



BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA ELECTRÓNICA
MAESTRÍA EN INGENIERÍA ELECTRÓNICA,
OPCIÓN INSTRUMENTACIÓN ELECTRÓNICA

Tesis para obtener el grado de:
MAESTRO EN INGENIERÍA ELECTRÓNICA

Normalización y alineación automática de la forma de la región
pulmonar integrada con selección de características
discriminantes para detección de neumonía y COVID-19

Presenta:

Lic. Rafael Alejandro Cruz Ovando*

Directores:

Dr. Salvador Eugenio Ayala Raggi

Dr. Aldrin Barreto Flores

Tabla de Contenido

Lista de Figuras	V
-------------------------	----------

Lista de Tablas	VI
------------------------	-----------

1 Introducción	1
1.1 Planteamiento del problema	3
1.2 Justificación	6
1.2.1 Contexto Global y Necesidad Clínica	6
1.2.2 Estado del Arte y Limitaciones de Enfoques Existentes	6
1.2.3 Contribución Científica y Técnica del Trabajo	7
1.2.4 Impacto y Aplicaciones Potenciales	8
1.2.5 Relevancia en el Contexto de COVID-19 y Salud Pública	8
1.2.6 Justificación Metodológica	9
1.3 Objetivos	10
1.3.1 Objetivo general	10
1.3.2 Objetivos específicos	10
2 Marco Teórico y Estado del Arte	11
2.1 Radiografías de Tórax en Diagnóstico Médico	13
2.1.1 Principios Físicos de la Radiografía Torácica	13
2.1.2 Anatomía Torácica y Definición de Landmarks Anatómicos	14
2.1.3 Aplicaciones Clínicas de la Localización de Landmarks	15
2.2 Fundamentos de Aprendizaje Profundo para Visión por Computadora	17
2.2.1 Redes Neuronales y Representaciones Jerárquicas	17
2.2.2 Redes Neuronales Convolucionales	18
2.2.3 Operaciones de Submuestreo y Funciones de Activación	19
2.2.4 Algoritmo de Retropropagación	21
2.2.5 Algoritmos de Optimización	22
2.3 Arquitecturas Residuales Profundas	24
2.3.1 El Problema de Degradación en Redes Profundas	24
2.3.2 Conexiones Residuales y Bloques Residuales	25
2.3.3 Arquitecturas de la Familia ResNet	26
2.3.4 Normalización por Lotes	27
2.3.5 Ventajas de Arquitecturas Residuales para Imágenes Médicas	28
2.4 Aprendizaje por Transferencia en Imágenes Médicas	29
2.4.1 Pre-entrenamiento en ImageNet y Representaciones Transferibles	29
2.4.2 Estrategias de Fine-Tuning y Adaptación al Dominio Médico	30
2.4.3 Brecha de Dominio y Adaptación para Radiografías de Tórax	32
2.5 Funciones de Pérdida para Regresión de Coordenadas	34

2.5.1	Error Cuadrático Medio	34
2.5.2	Wing Loss: Amplificación de Gradientes para Errores Pequeños	35
2.5.3	Restricciones Geométricas: Symmetry Loss y Distance Preservation Loss	37
2.6	Enfoques de Regresión para Detección de Landmarks	40
2.6.1	Regresión Directa de Coordenadas	40
2.6.2	Regresión de Mapas de Calor	42
2.6.3	Comparación y Selección de Enfoque	43
2.7	Estado del Arte en Detección Automática de Landmarks Anatómicos	46
2.7.1	Métodos Basados en Regresión de Coordenadas	46
2.7.2	Métodos Basados en Mapas de Calor	48
2.7.3	Métodos Híbridos y Basados en Transformers	49
2.7.4	Funciones de Pérdida Especializadas y Restricciones Geométricas	50
2.7.5	Análisis Comparativo y Posicionamiento del Presente Trabajo	52
2.8	Síntesis del Marco Teórico	56
3	Metodología	59
3.1	Introducción	59
3.2	Conjunto de Datos	62
3.2.1	Descripción General y Composición	62
3.2.2	Características Técnicas de las Imágenes	63
3.2.3	Definición de Landmarks Anatómicos	64
3.2.4	Pares de Landmarks Simétricos y Eje Mediastínico	65
3.2.5	División del Dataset para Entrenamiento, Validación y Prueba	67
3.2.6	Calidad y Validación de Anotaciones	68
3.3	Arquitectura del Modelo	71
3.3.1	Backbone ResNet-18: Extractor de Características Visuales	71
3.3.2	Módulo de Regresión: Mapeo de Características a Coordenadas	73
3.3.3	Distribución de Parámetros y Complejidad Computacional	75
3.3.4	Arquitectura Experimental: Integración de Coordinate Attention	76
3.4	Pipeline de preprocesamiento y aumentación de datos	78
3.4.1	Preprocesamiento determinístico	79
3.4.2	Aumentación estocástica de datos	82
3.4.3	Orden de aplicación y composición de transformaciones	86
3.4.4	Síntesis del pipeline	87
3.5	Estrategia de Entrenamiento Progresivo	88
3.5.1	Fase 1: Entrenamiento del Módulo de Regresión con Backbone Congelado	88
3.5.2	Fase 2: Fine-Tuning Completo con Wing Loss	90
3.5.3	Fase 3: Incorporación de Symmetry Loss para Consistencia Bilateral	93
3.5.4	Fase 4: Complete Loss con Preservación de Distancias Anatómicas	94
3.5.5	Estrategia de Warm-Start entre Fases	96
3.6	Detalles de implementación y reproducibilidad	98
3.6.1	Frameworks y librerías	98
3.6.2	Especificaciones de hardware y configuración computacional	101
3.6.3	Tiempos de entrenamiento	102
3.6.4	Protocolos de reproducibilidad	103
3.6.5	Gestión de experimentos y checkpoints	105
3.6.6	Síntesis de implementación	106
3.7	Métricas de evaluación	108

3.7.1	Error Radial Medio (MRE)	108
3.7.2	Error por landmark individual	110
3.7.3	Métricas de consistencia geométrica	110
3.7.4	Sistema de clasificación por calidad clínica	114
3.7.5	Protocolo de validación	116
3.7.6	Síntesis de métricas	116

Bibliografía

118

Lista de Figuras

Lista de Tablas

2.1.1 Descripción de los 15 <i>landmarks</i> anatómicos en radiografías de tórax	15
2.3.1 Arquitecturas de la familia ResNet. Los números entre paréntesis indican el número de bloques residuales en cada etapa.	26
2.6.1 Comparación de enfoques de regresión directa de coordenadas y mapas de calor para detección de <i>landmarks</i>	44
2.7.1 Estado del arte en detección de <i>landmarks</i> en imágenes médicas (2016-2024). Las abreviaciones utilizadas son: Ceph (cefalométrico), px (píxeles), NME (error medio normalizado), GCN (Graph Convolutional Network), ViT (Vision Transformer), RL (Reinforcement Learning).	55
3.2.1 Especificaciones técnicas del conjunto de datos de radiografías de tórax	63
3.2.2 Definición anatómica detallada de los 15 <i>landmarks</i> anotados en radiografías de tórax PA	65
3.2.3 División estratificada del conjunto de datos en subconjuntos de entrenamiento, validación y prueba	68
3.3.1 Distribución detallada de parámetros entrenables en arquitectura del modelo . .	75
3.6.1 Tiempos de entrenamiento por fase metodológica medidos sobre hardware especificado en Sección 3.6.2. Tiempo por época incluye entrenamiento sobre 669 muestras de entrenamiento, validación sobre 143 muestras, y operaciones de almacenamiento.	102

Capítulo 1

Introducción

El análisis preciso de radiografías de tórax es fundamental en el diagnóstico médico, siendo estas imágenes una herramienta ampliamente utilizada debido a su disponibilidad y bajo costo. Sin embargo, la interpretación manual es un proceso subjetivo, susceptible a errores y variabilidad inter-observador, especialmente en contextos de alta demanda [1]. Para mejorar la objetividad y eficiencia, la investigación en diagnóstico médico asistido por computadora se ha enfocado en desarrollar herramientas automáticas que asistan en el análisis de estas imágenes. Un paso crucial en este proceso es la correcta identificación y delimitación de estructuras anatómicas relevantes, como los campos pulmonares y puntos de referencia clave.

Esta tesis aborda el desarrollo de un sistema automatizado para la detección de puntos de referencia (en adelante referidos como *landmarks*) anatómicos en radiografías de tórax, componente fundamental para el análisis cuantitativo de imágenes médicas. Se propone una metodología basada en aprendizaje profundo (*deep learning*) que utiliza redes neuronales convolucionales (Convolutional Neural Networks, CNNs) que incorporan conocimiento anatómico del dominio médico mediante restricciones geométricas [2, 3]. El sistema predice de manera directa las coordenadas de 15 puntos de referencia anatómicos clave, aprovechando aprendizaje por transferencia (*transfer learning*) desde dominios de imágenes naturales [4]. Los resultados experimentales demuestran que el enfoque propuesto alcanza niveles de precisión que cumplen con los estándares internacionales de excelencia clínica establecidos para tareas de localización anatómica [5], validado sobre un conjunto de datos que incluye casos de COVID-19, neumonía viral y pacientes saludables.

La detección precisa de *landmarks* anatómicos constituye un componente fundamental para el desarrollo futuro de sistemas completos de diagnóstico asistido por computadora. Los puntos de referencia detectados automáticamente proporcionan una base para posteriores etapas de análisis, incluyendo la segmentación automática de regiones anatómicas, la normalización geométrica de imágenes y la clasificación de patologías torácicas. Esta investigación se enfoca específicamente en la primera etapa: la localización robusta y precisa de *landmarks* anatómicos mediante técnicas de aprendizaje profundo. Las líneas de investigación futuras derivadas de este trabajo incluyen el desarrollo de modelos de segmentación pulmonar, sistemas de normalización espacial y clasificadores de patologías que aprovechen los *landmarks* detectados

automáticamente.

Esta tesis se organiza de la siguiente manera: el Capítulo 2 presenta el marco teórico y el estado del arte en detección de *landmarks* anatómicos y aprendizaje profundo aplicado a imágenes médicas; el Capítulo 3 detalla la metodología propuesta, incluyendo la arquitectura de red neuronal, las funciones de pérdida especializadas y las estrategias de entrenamiento; el Capítulo 4 describe el conjunto de datos utilizado, las métricas de evaluación y el protocolo experimental; el Capítulo 5 presenta los resultados obtenidos y su análisis comparativo; finalmente, el Capítulo 6 discute las conclusiones, limitaciones y líneas futuras de investigación.

1.1. Planteamiento del problema

La interpretación de radiografías de tórax representa uno de los procedimientos de diagnóstico más frecuentes en la práctica clínica a nivel mundial, con más de 2 mil millones de estudios realizados anualmente [6]. La localización precisa de estructuras anatómicas clave mediante la identificación de *landmarks* es fundamental para el análisis cuantitativo y la toma de decisiones clínicas [7]. Estos *landmarks* anatómicos permiten el cálculo de índices diagnósticos como el índice cardiorácico, la detección de asimetrías patológicas y el establecimiento de sistemas de coordenadas consistentes para análisis longitudinales [8]. Sin embargo, la anotación manual de *landmarks* requiere aproximadamente 15 minutos por imagen y está sujeta a variabilidad inter e intra-observador de hasta 5-10 píxeles, limitando su aplicabilidad en escenarios clínicos de alto volumen [5].

Los enfoques tradicionales para la detección de *landmarks* anatómicos se basan en métodos de visión por computadora que utilizan características diseñadas manualmente (*hand-crafted features*) combinadas con modelos estadísticos de forma [9, 10]. Aunque estos métodos han demostrado efectividad en condiciones controladas, enfrentan limitaciones significativas: (1) requieren ingeniería manual de características específicas del dominio, proceso que resulta costoso y poco generalizable; (2) dependen de alineamientos geométricos previos (como el Análisis de Procrustes Generalizado, Generalized Procrustes Analysis, GPA) que pueden fallar ante deformaciones anatómicas severas; (3) modelan relaciones lineales mediante Análisis de Componentes Principales (Principal Component Analysis, PCA), incapaces de capturar la naturaleza no lineal de las variaciones anatómicas; y (4) presentan sensibilidad elevada a condiciones de imagen como bajo contraste, ruido y artefactos [3, 11].

El surgimiento del aprendizaje profundo (*deep learning*) ha transformado radicalmente el análisis de imágenes médicas [2], demostrando capacidad para aprender representaciones jerárquicas de características directamente desde datos sin necesidad de ingeniería manual [12]. Las redes neuronales convolucionales (Convolutional Neural Networks, CNNs) han alcanzado niveles de desempeño comparables o superiores al de especialistas humanos en diversas tareas de imagenología médica [13, 14]. Sin embargo, la detección precisa de *landmarks* anatómicos mediante CNNs presenta desafíos específicos que requieren soluciones especializadas más allá de las arquitecturas estándar de clasificación o segmentación.

El problema central que aborda esta investigación se formula de la siguiente manera: **¿Cómo diseñar un sistema automatizado basado en redes neuronales convolucionales que detecte *landmarks* anatómicos en radiografías de tórax con precisión clínicamente útil (error <8.5 píxeles), incorporando conocimiento anatómico del dominio médico y resultando computacionalmente eficiente para integración en flujos de trabajo hospitalarios?**

Este problema general se descompone en los siguientes desafíos técnicos específicos:

Desafío 1: Alta precisión en regresión de coordenadas. A diferencia de tareas de clasificación donde pequeños errores son tolerables, la localización de *landmarks* constituye un problema de regresión donde el modelo debe predecir coordenadas continuas (x, y) con alta precisión para resultar clínicamente útil. Los estándares internacionales establecen que un error inferior a 8.5 píxeles representa excelencia clínica [5]. Las funciones de pérdida estándar como el Error Cuadrático Medio (Mean Squared Error, MSE) tratan todos los errores de manera uniforme, penalizando excesivamente valores atípicos (*outliers*) pero proporcionando gradientes insuficientes para refinar predicciones ya cercanas al objetivo. Este comportamiento dificulta el logro de la alta precisión requerida en aplicaciones clínicas [15].

Desafío 2: Incorporación de conocimiento anatómico. El cuerpo humano exhibe restricciones geométricas inherentes que no son explotadas por enfoques estándar de aprendizaje profundo. Específicamente, las radiografías de tórax presentan simetría bilateral aproximada entre pulmones izquierdo y derecho, relaciones de distancia fijas entre estructuras anatómicas (ancho torácico, altura mediastínica), y restricciones de ordenamiento espacial (los ápices pulmonares siempre se localizan superiormente a las bases) [16]. Integrar explícitamente este conocimiento anatómico como restricciones geométricas en el proceso de optimización constituye un desafío metodológico no resuelto completamente en la literatura existente [17, 18].

Desafío 3: Generalización ante variabilidad patológica. El sistema debe mantener precisión robusta en presencia de condiciones patológicas que alteran significativamente la apariencia radiográfica. Las manifestaciones de COVID-19 (opacidades en vidrio esmerilado, consolidaciones), neumonía viral (infiltrados intersticiales) y otras patologías torácicas pueden oscurecer parcialmente referencias anatómicas, reduciendo el contraste local y dificultando la localización precisa de *landmarks* [19, 20]. El modelo debe aprender representaciones suficientemente robustas para localizar estructuras anatómicas incluso cuando los límites no resultan claramente visibles.

Desafío 4: Eficiencia computacional para despliegue clínico. Para resultar práctico en entornos hospitalarios, el sistema debe ejecutar inferencia en equipo físico (*hardware*) de consumo general (sin requerir GPUs de alta gama) en tiempos de respuesta cercanos al tiempo real (menos de 1 segundo por imagen). Esta restricción limita la complejidad arquitectural viable y motiva el uso de modelos eficientes con *transfer learning* desde dominios de datos abundantes [4].

Desafío 5: Escasez de datos médicos etiquetados. A diferencia de aplicaciones de visión por computadora en dominios generales donde existen millones de imágenes etiquetadas (ImageNet: 1.2M imágenes), los conjuntos de datos (*datasets*) médicos típicamente contienen

cientos o pocos miles de imágenes anotadas debido al costo y tiempo requerido para anotación experta [21]. Esta escasez de datos incrementa el riesgo de sobreajuste (*overfitting*) y limita la capacidad de generalización de modelos entrenados desde cero, motivando estrategias de *transfer learning* y regularización especializada [22].

La solución a estos desafíos interconectados requiere una aproximación metodológica que integre: (1) arquitecturas de redes neuronales eficientes con capacidad de extracción de características robustas, (2) funciones de pérdida especializadas diseñadas para alta precisión mediante amplificación de gradientes en el régimen de errores pequeños, (3) mecanismos de regularización geométrica que incorporen conocimiento anatómico del dominio médico, y (4) estrategias de optimización progresiva que balanceen convergencia rápida con precisión final. El Capítulo 3 presenta en detalle la metodología propuesta que aborda sistemáticamente cada uno de estos desafíos.

1.2. Justificación

1.2.1. Contexto Global y Necesidad Clínica

Las radiografías de tórax constituyen el estudio de imagenología más frecuentemente realizado a nivel mundial, representando la primera línea de evaluación para enfermedades pulmonares y cardíacas en servicios de urgencias, unidades de cuidados intensivos y consulta ambulatoria. La localización precisa de estructuras anatómicas clave mediante la identificación de puntos de referencia (*landmarks*) es esencial para el análisis cuantitativo, la toma de decisiones clínicas y el seguimiento longitudinal de pacientes [7, 8].

El proceso tradicional de anotación manual de *landmarks* anatómicos enfrenta limitaciones críticas en el contexto clínico contemporáneo: (1) el tiempo requerido resulta prohibitivo en escenarios de alta demanda; (2) la variabilidad inter e intra-observador afecta la reproducibilidad de mediciones cuantitativas [5]; (3) la fatiga del observador incrementa errores en sesiones prolongadas de anotación; y (4) el crecimiento exponencial en volumen de estudios radiológicos supera la disponibilidad de especialistas capacitados, particularmente en regiones con recursos limitados. La pandemia de COVID-19 ha evidenciado dramáticamente esta brecha, con incrementos sostenidos en demanda de interpretación de radiografías torácicas que exceden la capacidad de respuesta del personal médico disponible [19].

1.2.2. Estado del Arte y Limitaciones de Enfoques Existentes

El aprendizaje profundo (*deep learning*) ha transformado el análisis de imágenes médicas en la última década, alcanzando desempeño comparable o superior a especialistas humanos en diversas tareas de clasificación y segmentación [2]. Trabajos seminales han demostrado este potencial en dermatología [13], oftalmología [14] y radiología torácica [20]. Sin embargo, la detección precisa de *landmarks* anatómicos mediante redes neuronales convolucionales presenta desafíos específicos que no se resuelven simplemente mediante el escalamiento de arquitecturas o el incremento de datos de entrenamiento.

Los métodos existentes para detección de *landmarks* en radiografías de tórax exhiben limitaciones significativas: (1) enfoques basados en regresión de mapas de calor (*heatmap regression*) requieren resolución espacial elevada y memoria computacional sustancial, dificultando su despliegue en equipo físico (*hardware*) de consumo [23]; (2) métodos de regresión coordinada (*coordinate regression*) con funciones de pérdida estándar como el Error Cuadrático Medio (Mean Squared Error, MSE) no alcanzan los niveles de precisión requeridos para excelencia clínica (error <8.5 píxeles) [24]; (3) sistemas que ignoran restricciones geométricas anatómicas producen predicciones anatómicamente implausibles (asimetrías

artificiales, violación de relaciones espaciales fundamentales); y (4) la mayoría de trabajos reportan validación en conjuntos de datos (*datasets*) con más de 10,000 imágenes, dejando sin resolver el problema de entrenamiento efectivo con *datasets* médicos de tamaño limitado (típicamente cientos o pocos miles de imágenes) [21].

Revisiones sistemáticas de sistemas de inteligencia artificial para COVID-19 han identificado riesgo de sesgo elevado, falta de validación externa y reporte inadecuado de metodología en la mayoría de publicaciones [25, 26]. Estas limitaciones metodológicas subrayan la necesidad de investigación rigurosa que establezca estándares reproducibles para desarrollo y validación de sistemas de análisis automatizado de imágenes médicas.

1.2.3. Contribución Científica y Técnica del Trabajo

Esta tesis aborda las limitaciones identificadas mediante las siguientes contribuciones:

1. Función de pérdida geométrica con conocimiento anatómico. Se desarrolla una función de pérdida multi-componente que integra: (a) Wing Loss [15] para mejora de precisión mediante amplificación de gradientes en el régimen de errores pequeños (refinando predicciones cercanas al objetivo), (b) Symmetry Loss para imponer simetría bilateral anatómica [16], y (c) Distance Preservation Loss para preservar relaciones espaciales críticas entre estructuras [17]. Esta integración de conocimiento del dominio médico mediante restricciones geométricas constituye una contribución metodológica que supera el enfoque tradicional de incrementar complejidad arquitectural.

2. Estrategia de entrenamiento progresivo en cuatro fases. Se propone una metodología sistemática que progresa desde congelamiento de columna vertebral de la red (*backbone*) hasta optimización completa con restricciones geométricas incrementales. Esta estrategia permite convergencia estable y mejora progresiva del desempeño, demostrando ser superior al entrenamiento extremo a extremo (*end-to-end*) directo en *datasets* médicos de tamaño limitado [4, 27].

3. Validación empírica rigurosa. El sistema se valida sobre 956 radiografías que incluyen casos de COVID-19, neumonía viral y pacientes normales, con evaluación multi-dimensional mediante métricas estándar (error radial medio) y métricas geométricas especializadas (consistencia bilateral, validez anatómica). El diseño experimental incluye análisis de ablación sistemático que cuantifica la contribución individual de cada componente metodológico, proporcionando evidencia empírica del valor de restricciones geométricas sobre complejidad arquitectural.

4. Eficiencia computacional y reproducibilidad. El sistema alcanza precisión de excelencia clínica con inferencia en menos de 1 segundo por imagen en *hardware* de consumo (GPU

de gama media con 8GB VRAM), demostrando viabilidad para despliegue en entornos con recursos limitados. Todo el código, configuraciones experimentales y resultados se documentan exhaustivamente para facilitar reproducción y validación independiente, abordando las deficiencias metodológicas identificadas en revisiones sistemáticas [25].

1.2.4. Impacto y Aplicaciones Potenciales

Los *landmarks* anatómicos detectados automáticamente por el sistema propuesto constituyen la base para una secuencia de procesamiento (en adelante referida como *pipeline*) de análisis completa que permitirá, como trabajo futuro, desarrollar:

1. Segmentación automática precisa. Los 15 *landmarks* pueden inicializar Modelos Activos de Forma (Active Shape Models, ASM) para delineación automatizada de contornos pulmonares con modelado de forma anatómicamente plausible [9, 11].

2. Normalización geométrica robusta. Las coordenadas de *landmarks* permiten calcular transformaciones geométricas que estandaricen pose, escala y orientación, eliminando variaciones extrínsecas y facilitando análisis cuantitativo reproducible.

3. Extracción de ROI normalizadas. Regiones de interés (Regions of Interest, ROI) estandarizadas geométricamente reducen variabilidad inter-sujeto no relacionada con patología, mejorando la sensibilidad y especificidad de análisis posteriores.

4. Sistemas de clasificación de patologías. Representaciones normalizadas pueden alimentar clasificadores de aprendizaje profundo para detección automática de neumonía, COVID-19 y otras patologías torácicas [20, 28].

La metodología desarrollada no se limita a radiografías de tórax; es generalizable a otros problemas de localización anatómica en imágenes médicas donde existen restricciones geométricas inherentes (simetría en imágenes cerebrales, proporciones anatómicas en radiografías pediátricas, etc.). Esta generalización amplifica el impacto potencial del trabajo más allá del dominio específico de aplicación.

1.2.5. Relevancia en el Contexto de COVID-19 y Salud Pública

La pandemia de COVID-19 ha incrementado exponencialmente la demanda de herramientas de diagnóstico asistido por computadora para triaje rápido y seguimiento de pacientes [19]. Las manifestaciones radiológicas de COVID-19 (opacidades en vidrio esmerilado, consolidaciones, distribución periférica) requieren evaluación cuantitativa de extensión y distribución que se beneficiaría significativamente de *landmarks* anatómicos localizados automáticamente. El sistema propuesto demuestra robustez ante variabilidad patológica, manteniendo precisión

clínicamente útil en casos de COVID-19 y neumonía viral, validando su aplicabilidad en escenarios clínicos reales.

Más allá de COVID-19, las enfermedades respiratorias crónicas (EPOC, fibrosis pulmonar, asma severa) y agudas (neumonía bacteriana, tuberculosis) requieren seguimiento longitudinal mediante radiografías seriadas. Sistemas automatizados de análisis cuantitativo basados en *landmarks* precisos permitirían monitoreo objetivo de progresión de enfermedad y respuesta a tratamiento, mejorando la calidad de atención médica especialmente en entornos con acceso limitado a especialistas.

1.2.6. Justificación Metodológica

La elección de ResNet-18 como arquitectura base se justifica por su balance óptimo entre capacidad de representación (11.7M parámetros) y eficiencia computacional, permitiendo entrenamiento efectivo con conjuntos de datos (*datasets*) de tamaño limitado mediante aprendizaje por transferencia (*transfer learning*) desde ImageNet [4, 29]. Estudios sistemáticos han demostrado que *transfer learning* desde ImageNet beneficia tareas médicas especialmente con menos de 10,000 imágenes, siendo las capas iniciales altamente transferibles entre dominios [22].

La formulación del problema como regresión coordinada directa (*coordinate regression*) en lugar de regresión de mapas de calor (*heatmap regression*) se justifica por: (1) eficiencia computacional (30 salidas vs. 15 mapas de calor de alta resolución), (2) predicciones con valores de coordenadas continuos (ej. 120.37px) sin necesidad de post-procesamiento de mapas de calor, y (3) facilidad de integración con restricciones geométricas en la función de pérdida [24].

En conclusión, esta investigación se justifica por su contribución metodológica (funciones de pérdida geométricas, entrenamiento progresivo), validación rigurosa (análisis multi-dimensional con datasets multi-categoría), eficiencia computacional (despliegue viable en hardware de consumo), y potencial de impacto en salud pública (base para sistemas de diagnóstico asistido accesibles globalmente). Los resultados establecen una metodología reproducible para integración de conocimiento del dominio en sistemas de aprendizaje profundo médico, principio generalizable más allá del problema específico abordado [2].

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar e implementar algoritmos de visión por computadora para la detección, alineación y normalización de la forma de la región pulmonar en imágenes radiográficas de tórax, utilizando además un método eficaz para la selección de características discriminantes, con el fin de mejorar la precisión en la detección automática de neumonía y COVID-19.

1.3.2. Objetivos específicos

1. Diseñar, implementar y evaluar un método deformable de alineación y normalización que localice, segmente y ajuste automáticamente la región pulmonar en términos de forma, escala, posición y rotación.
2. Proponer un método de extracción y selección de características que maximicen la discriminación entre las clases.
3. Evaluar el rendimiento de diferentes clasificadores de aprendizaje supervisado para la técnica de alineación propuesta en la tesis: KNN, CNN, MLP.
4. Validar el clasificador desarrollado a través de medir la precisión, sensibilidad, especificidad y además de realizar pruebas de validación cruzada para caracterizar el algoritmo propuesto
5. Contrastar los resultados de clasificación del objetivo anterior con resultados obtenidos por los mismos clasificadores pero sin realizar el proceso de alineación propuesto.
6. Publicación de resultados

Capítulo 2

Marco Teórico y Estado del Arte

La detección automática de puntos de referencia anatómicos (*landmarks*) en radiografías de tórax constituye un problema fundamental en el análisis computarizado de imágenes médicas. Como se estableció en el Capítulo 1, la localización precisa de estos puntos es esencial para la cuantificación de estructuras anatómicas, el cálculo de índices diagnósticos como el índice cardiorácico, y la normalización espacial de radiografías para sistemas de clasificación automática. El presente capítulo tiene como objetivo establecer los fundamentos teóricos que sustentan el desarrollo de métodos basados en aprendizaje profundo para la detección automática de *landmarks* anatómicos, revisando tanto los principios fundamentales como el estado del arte actual en esta área de investigación.

La detección de *landmarks* en imágenes médicas ha experimentado una evolución significativa en las últimas dos décadas. Tradicionalmente, este problema se abordó mediante métodos estadísticos y geométricos, tales como los Modelos Activos de Forma (*Active Shape Models*, ASM) [9] y los Modelos Activos de Apariencia (*Active Appearance Models*, AAM) [10], que representan variaciones anatómicas mediante descomposición lineal basada en Análisis de Componentes Principales. Si bien estos métodos clásicos demostraron utilidad en escenarios controlados, presentan limitaciones fundamentales relacionadas con la linealidad de sus representaciones y su dependencia de características diseñadas manualmente (*hand-crafted features*) [11]. La irrupción del aprendizaje profundo en visión por computadora, particularmente tras el trabajo seminal de Krizhevsky et al. [12], ha revolucionado el análisis de imágenes médicas al permitir el aprendizaje automático de representaciones jerárquicas directamente desde los datos [2, 3]. En el contexto específico de la detección de *landmarks*, las Redes Neuronales Convolucionales (CNNs) han demostrado capacidad superior para capturar patrones complejos y no lineales en estructuras anatómicas, superando consistentemente el desempeño de métodos tradicionales.

El presente capítulo se estructura en ocho secciones que abarcan desde los fundamentos físicos de las radiografías de tórax hasta el estado del arte en métodos basados en aprendizaje profundo. La Sección 2.1 introduce los principios físicos de las radiografías torácicas y define los quince *landmarks* anatómicos relevantes para este trabajo. La Sección 2.2 establece los fundamentos matemáticos de las redes neuronales convolucionales, incluyendo la operación de convolución,

funciones de activación, y el algoritmo de retropropagación (*backpropagation*). La Sección 2.3 analiza en detalle las arquitecturas residuales, particularmente la familia ResNet, que han demostrado ser especialmente efectivas para el entrenamiento de redes profundas mediante el uso de conexiones residuales (*skip connections*). La Sección 2.4 examina el paradigma de aprendizaje por transferencia (*transfer learning*), un componente crucial cuando se trabaja con conjuntos de datos médicos de tamaño limitado. La Sección 2.5 presenta una revisión exhaustiva de funciones de pérdida especializadas para la regresión de coordenadas, con énfasis particular en *Wing Loss* y funciones de pérdida basadas en restricciones geométricas. La Sección 2.6 contrasta los enfoques de regresión directa de coordenadas versus regresión de mapas de calor (*heatmap regression*), justificando la elección metodológica adoptada en esta tesis. La Sección 2.7 ofrece un análisis comparativo exhaustivo del estado del arte en detección de *landmarks* anatómicos, identificando las brechas que motivan el presente trabajo. Finalmente, la Sección 2.8 sintetiza los conceptos presentados y establece la conexión con la metodología propuesta que se desarrollará en el Capítulo 3.

2.1. Radiografías de Tórax en Diagnóstico Médico

Las radiografías de tórax constituyen el estudio de imagen médica más frecuentemente utilizado en la práctica clínica. Como se estableció en el Capítulo 1, este método diagnóstico es fundamental para la evaluación de patologías pulmonares, cardiovasculares y mediastinales. La interpretación radiológica se fundamenta en la identificación de estructuras anatómicas de referencia (*landmarks*) y en la evaluación de sus relaciones espaciales. En el contexto del análisis computarizado de imágenes médicas, la detección automática de estos puntos anatómicos representa una tarea fundamental para sistemas de diagnóstico asistido por computadora y para la cuantificación objetiva de hallazgos radiológicos.

2.1.1. Principios Físicos de la Radiografía Torácica

La formación de imágenes radiográficas se basa en la interacción de fotones de rayos X con tejidos biológicos, específicamente mediante los fenómenos de dispersión Compton y absorción fotoeléctrica. Los sistemas de radiografía torácica convencionales operan con voltajes de aceleración de 110-120 kVp para proyecciones posteroanterior (PA), generando radiación electromagnética con energías en el rango de 40 a 150 keV [30].

La atenuación del haz de rayos X al atravesar tejido biológico se describe mediante la Ley de Beer-Lambert:

$$I(x) = I_0 \exp \left(- \int_0^x \mu(s) ds \right) \quad (2.1.1)$$

donde $I(x)$ representa la intensidad del haz transmitido después de atravesar un espesor x de tejido, I_0 es la intensidad del haz incidente, y $\mu(s)$ es el coeficiente de atenuación lineal en función de la posición. Para tejidos homogéneos con coeficiente de atenuación constante, la ecuación se simplifica a:

$$I = I_0 e^{-\mu x} \quad (2.1.2)$$

El contraste radiográfico resulta de las diferencias en los coeficientes de atenuación entre los tejidos que componen la anatomía torácica. Los campos pulmonares, compuestos predominantemente por aire alveolar ($\mu \approx 0,0001 \text{ cm}^{-1}$), presentan baja atenuación y aparecen radiolúcidos (oscuros) en las radiografías. Por el contrario, las estructuras mediastinales y la silueta cardíaca, constituidas por tejidos blandos con coeficientes de atenuación superiores ($\mu \approx 0,20 \text{ cm}^{-1}$), presentan mayor radioopacidad (tonos claros). Las estructuras óseas de la caja torácica (costillas, clavículas, columna vertebral) exhiben la mayor atenuación ($\mu \approx 0,50 \text{ cm}^{-1}$ para hueso cortical) [30]. Esta diferenciación inherente de densidades radiográficas entre

estructuras anatómicas adyacentes genera los bordes y contornos que definen los *landmarks* anatómicos de interés para este trabajo.

2.1.2. Anatomía Torácica y Definición de Landmarks Anatómicos

La anatomía torácica en proyección posteroanterior comprende tres compartimentos principales: los campos pulmonares bilaterales, el mediastino central, y la caja torácica ósea [31, 32]. La interpretación sistemática de radiografías de tórax requiere la identificación de estructuras anatómicas de referencia cuya localización precisa permite la evaluación de normalidad anatómica y la detección de alteraciones patológicas.

Los *landmarks* anatómicos se definen como puntos de referencia específicos que corresponden a estructuras anatómicas con significado clínico establecido y criterios de identificación reproducibles entre observadores expertos. A diferencia de regiones de interés arbitrarias, los *landmarks* representan localizaciones anatómicas con propiedades geométricas consistentes que pueden explotarse mediante restricciones geométricas en algoritmos de detección automática [33].

El presente trabajo aborda la detección automática de 15 *landmarks* anatómicos distribuidos en las estructuras pulmonares, mediastinales y óseas de la radiografía de tórax. Como se ilustra en la Figura ??, estos puntos de referencia se seleccionaron considerando tres criterios fundamentales: (1) detectabilidad visual consistente en radiografías de calidad diagnóstica, (2) relevancia anatómica para la caracterización de la geometría torácica, y (3) distribución espacial que captura la estructura global del tórax. La Tabla 2.1.1 presenta la nomenclatura y localización anatómica de cada punto.

Esta configuración de *landmarks* presenta características geométricas de particular relevancia para el análisis computarizado: siete pares de puntos con simetría bilateral (#3-4, #5-6, #7-8, #10-11, #12-13, #14-15, y #2 respecto al eje de simetría), y dos puntos localizados en la línea media que definen el eje vertical de simetría (#1 y #9). La simetría bilateral es una propiedad anatómica fundamental del tórax normal que puede explotarse mediante restricciones geométricas en algoritmos de aprendizaje profundo. Esta propiedad geométrica se incorpora explícitamente en la función de pérdida propuesta en este trabajo, como se discutirá en detalle en la Sección 2.5. Adicionalmente, las distancias entre pares de *landmarks* específicos (por ejemplo, entre #3 y #4, o entre #12 y #13) representan medidas anatómicas con variabilidad limitada que pueden utilizarse como restricciones de preservación de distancias.

Tabla 2.1.1: Descripción de los 15 *landmarks* anatómicos en radiografías de tórax

Nº	Nombre anatómico	Localización / Descripción
1	Escotadura yugular	Punto superior en línea media, entre articulaciones esternoclaviculares
2	Ángulo cardiofrénico izquierdo	Unión del borde inferior izquierdo de la silueta cardíaca con la cúpula diafragmática
3	Borde costal lateral superior izquierdo	Contorno lateral alto del hemitórax izquierdo, nivel de 2ª-3ª costilla posterior
4	Borde costal lateral superior derecho	Homólogo del <i>landmark</i> #3 en el hemitórax derecho
5	Borde costal lateral medio izquierdo	Contorno medio lateral del pulmón izquierdo, tercio medio del hemitórax
6	Borde costal lateral medio derecho	Homólogo del <i>landmark</i> #5 en el hemitórax derecho
7	Borde costal lateral inferior izquierdo	Contorno lateral inferior del pulmón izquierdo, inmediatamente superior al diafragma
8	Borde costal lateral inferior derecho	Homólogo del <i>landmark</i> #7 en el hemitórax derecho
9	Carina traqueal	Bifurcación de la tráquea en bronquios principales, mediastino medio
10	Borde cardíaco derecho medio	Límite lateral derecho de la silueta cardíaca, correspondiente a la aurícula derecha
11	Borde cardíaco izquierdo inferior	Límite lateral inferior izquierdo de la silueta cardíaca, ventrículo izquierdo
12	Ápice pulmonar izquierdo subclavicular	Punto más alto del campo pulmonar izquierdo, bajo el extremo medial de la clavícula
13	Ápice pulmonar derecho subclavicular	Homólogo del <i>landmark</i> #12 en el hemitórax derecho
14	Ángulo costofrénico izquierdo	Receso pleural posterolateral izquierdo, unión diafragma-pared torácica costal
15	Ángulo costofrénico derecho	Homólogo del <i>landmark</i> #14 en el hemitórax derecho

2.1.3. Aplicaciones Clínicas de la Localización de Landmarks

La localización precisa de *landmarks* anatómicos en radiografías de tórax tiene múltiples aplicaciones en la práctica clínica y en sistemas de análisis automatizado. Las aplicaciones diagnósticas incluyen la cuantificación de parámetros anatómicos (como índice cardiotorácico, altura pulmonar, y evaluación de simetría bilateral), la detección de asimetrías patológicas mediante comparación de distancias entre pares de *landmarks* homólogos, y la caracterización de deformaciones anatómicas asociadas a patologías específicas [34].

En el contexto de sistemas de diagnóstico asistido por computadora, la detección automática de *landmarks* constituye una etapa de preprocesamiento fundamental para dos aplicaciones principales [8, 35]: (1) la normalización espacial de radiografías con variabilidad en posicionamiento del paciente, distancia foco-detector, y grado de inspiración, y (2) la

segmentación automática de regiones anatómicas mediante la definición de límites anatómicos iniciales. Como se estableció en el Capítulo 1, la detección manual de *landmarks* por radiólogos expertos presenta variabilidad inter-observador significativa y requiere tiempo considerable, motivando el desarrollo de métodos automatizados basados en aprendizaje profundo.

Estudios recientes han demostrado que la incorporación explícita de *landmarks* anatómicos en arquitecturas de aprendizaje profundo mejora significativamente la interpretabilidad de los modelos, permitiendo a los clínicos comprender qué regiones anatómicas contribuyen a las predicciones diagnósticas [33, 36]. Esta interpretabilidad representa un aspecto crítico para la adopción clínica de sistemas de inteligencia artificial en medicina, particularmente en contextos de alta demanda diagnóstica donde la confiabilidad y la transparencia algorítmica son requisitos esenciales [37]. Los fundamentos matemáticos y computacionales de las arquitecturas de aprendizaje profundo que permiten la detección automática de estos *landmarks* se desarrollan en las secciones subsecuentes.

2.2. Fundamentos de Aprendizaje Profundo para Visión por Computadora

El aprendizaje profundo (*deep learning*) ha revolucionado el campo de la visión por computadora en la última década, permitiendo el desarrollo de sistemas capaces de aprender representaciones jerárquicas de características directamente desde datos en bruto, sin la necesidad de ingeniería manual de características. En el contexto de la detección de *landmarks* anatómicos, las Redes Neuronales Convolucionales (*CNNs*) representan la arquitectura fundamental que sustenta los métodos del estado del arte. Esta sección establece los fundamentos matemáticos y computacionales necesarios para comprender las arquitecturas residuales (Sección 2.3), las estrategias de aprendizaje por transferencia (Sección 2.4), y las funciones de pérdida especializadas (Sección 2.5) que constituyen los componentes técnicos centrales de este trabajo [38, 39].

2.2.1. Redes Neuronales y Representaciones Jerárquicas

Las redes neuronales artificiales profundas se componen de múltiples capas de transformaciones no lineales que procesan información de manera jerárquica. En el caso del perceptrón multicapa básico, cada neurona en la capa l computa una combinación lineal de las activaciones de la capa anterior seguida de una función de activación no lineal:

$$a_j^{(l)} = f \left(\sum_{i=1}^n w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)} \right) \quad (2.2.1)$$

donde $w_{ij}^{(l)}$ representa el peso de conexión de la neurona i en la capa $l - 1$ a la neurona j en la capa l , $b_j^{(l)}$ es el término de sesgo, y $f(\cdot)$ es la función de activación no lineal.

El concepto de representaciones jerárquicas es fundamental en *deep learning*: las capas tempranas de la red aprenden a detectar características de bajo nivel (bordes, texturas, gradientes), mientras que las capas profundas componen estas características simples para formar representaciones de alto nivel (formas complejas, objetos completos, relaciones espaciales). Esta jerarquía de abstracciones es particularmente adecuada para el procesamiento de imágenes médicas, donde la detección de estructuras anatómicas complejas requiere la integración de información visual a múltiples escalas [38].

Sin embargo, las arquitecturas basadas en capas completamente conectadas (*fully connected*) presentan limitaciones severas para el procesamiento de imágenes. Una imagen de radiografía de tórax de dimensiones modestas (256×256 píxeles con un canal de intensidad) contiene 65,536 valores de entrada. Una capa completamente conectada con 1,000 neuronas

requeriría 65.5 millones de parámetros solo en la primera capa, resultando en un modelo computacionalmente intratable y altamente susceptible al sobreajuste. Esta limitación motivó el desarrollo de arquitecturas convolucionales que explotan la estructura espacial de las imágenes mediante compartición de parámetros y conectividad local [40].

2.2.2. Redes Neuronales Convolucionales

Las Redes Neuronales Convolucionales (*Convolutional Neural Networks*, CNNs) constituyen una clase especializada de redes neuronales diseñadas para procesar datos con topología de rejilla, como imágenes bidimensionales. La operación fundamental de las CNNs es la convolución discreta, definida matemáticamente para imágenes bidimensionales como:

$$Y[i, j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X[i + m, j + n] \cdot W[m, n] + b \quad (2.2.2)$$

donde X representa la imagen de entrada con dimensiones $H \times W$, W es el kernel o filtro convolucional de dimensiones $M \times N$ (típicamente 3×3 o 5×5), Y es el mapa de características de salida (*feature map*), y b es el término de sesgo compartido por todas las ubicaciones espaciales.

La aplicación de la convolución está controlada por dos hiperparámetros adicionales: el paso (del inglés, *stride*) S , que determina el desplazamiento del kernel entre aplicaciones consecutivas, y el relleno (del inglés, *padding*) P , que especifica el número de píxeles añadidos en los bordes de la imagen de entrada. La dimensión espacial de la salida se calcula mediante:

$$H_{out} = \left\lfloor \frac{H_{in} + 2P - K}{S} \right\rfloor + 1, \quad W_{out} = \left\lfloor \frac{W_{in} + 2P - K}{S} \right\rfloor + 1 \quad (2.2.3)$$

donde K representa el tamaño del kernel (asumiendo kernels cuadrados $K \times K$). El *padding* se utiliza frecuentemente para preservar las dimensiones espaciales ($P = \lfloor K/2 \rfloor$ con $S = 1$ mantiene $H_{out} = H_{in}$), mientras que valores de *stride* mayores a 1 reducen las dimensiones espaciales, proporcionando una forma de submuestreo.

El campo receptivo (del inglés, *receptive field*) de una neurona en una capa profunda define la región de la imagen de entrada que influye en su activación. En CNNs, el campo receptivo crece exponencialmente con la profundidad de la red: una neurona en la capa L con kernels de tamaño K tiene un campo receptivo de tamaño aproximado $(K - 1)L + 1$. Este crecimiento permite que capas profundas integren información de regiones cada vez más extensas de la imagen, capturando contexto espacial relevante para la tarea de detección.

Las capas convolucionales presentan tres propiedades arquitectónicas fundamentales que

las hacen superiores a capas completamente conectadas para visión por computadora [12, 40]:

1. **Compartición de parámetros:** El mismo filtro se aplica en todas las ubicaciones espaciales de la imagen, reduciendo drásticamente el número de parámetros. Un kernel de 3×3 con 64 filtros requiere solo $3 \times 3 \times 64 = 576$ parámetros (más 64 sesgos), independientemente del tamaño de la imagen de entrada.
2. **Invarianza traslacional:** Características detectadas en una región de la imagen pueden ser detectadas en cualquier otra región mediante el mismo conjunto de pesos, proporcionando robustez a traslaciones del objeto de interés.
3. **Conectividad local:** Cada neurona procesa solo una región local de la entrada, explotando la correlación espacial inherente en imágenes naturales y médicas.

Una capa convolucional típica aplica múltiples filtros en paralelo, donde cada filtro aprende a detectar una característica específica (bordes horizontales, verticales, gradientes de intensidad, texturas). La salida de una capa convolucional es un tensor tridimensional de dimensiones $H_{out} \times W_{out} \times D_{out}$, donde D_{out} representa el número de filtros aplicados. Las capas tempranas de CNNs profundas aprenden detectores de características de bajo nivel, mientras que capas subsecuentes componen estas características para formar representaciones jerárquicamente más abstractas [12].

2.2.3. Operaciones de Submuestreo y Funciones de Activación

Las operaciones de submuestreo (del inglés, *pooling*) reducen progresivamente las dimensiones espaciales de las representaciones intermedias, disminuyendo la carga computacional y el número de parámetros, mientras expanden el campo receptivo efectivo de las capas subsecuentes. La operación de submuestreo máximo (*max pooling*) es la más ampliamente utilizada en arquitecturas modernas, definida como:

$$Y[i, j] = \max_{m, n \in R_{ij}} X[m, n] \quad (2.2.4)$$

donde R_{ij} representa la región de *pooling*, típicamente de tamaño 2×2 con *stride* de 2, lo que reduce las dimensiones espaciales a la mitad. Alternativamente, el submuestreo promedio (*average pooling*) calcula la media aritmética de los valores en la región:

$$Y[i, j] = \frac{1}{|R_{ij}|} \sum_{m, n \in R_{ij}} X[m, n] \quad (2.2.5)$$

El *max pooling* proporciona invarianza a pequeñas traslaciones y deformaciones locales,

preservando la activación máxima (más fuerte) dentro de cada región. Esta propiedad es particularmente útil para tareas de detección donde la presencia de una característica es más relevante que su ubicación precisa dentro de una región local.

Las funciones de activación no lineales son componentes esenciales de las redes neuronales, ya que permiten a la red aprender transformaciones no lineales complejas. La función de activación más ampliamente utilizada en CNNs modernas es la Unidad Lineal Rectificada (*Rectified Linear Unit*, ReLU), definida como:

$$f(x) = \max(0, x) = \begin{cases} x & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases} \quad (2.2.6)$$

La derivada de ReLU es particularmente simple:

$$f'(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases} \quad (2.2.7)$$

ReLU presenta ventajas significativas sobre funciones de activación clásicas como sigmoide y tangente hiperbólica [12]: (1) no exhibe saturación en la región positiva, evitando el problema del gradiente desvaneciente que afecta a redes profundas con activaciones sigmoideas; (2) su evaluación es computacionalmente eficiente, requiriendo solo una operación de comparación y selección máxima; (3) induce *sparsity* en las representaciones, ya que aproximadamente 50 % de las activaciones son cero, lo que puede mejorar la eficiencia y la generalización.

Una limitación de ReLU es el fenómeno conocido como “dying ReLU”, donde neuronas que consistentemente reciben entradas negativas producen activaciones de cero y dejan de aprender, ya que sus gradientes son nulos. Variantes como Leaky ReLU ($f(x) = \max(0, \alpha x, x)$) y Parametric ReLU (PReLU) abordan parcialmente esta limitación al permitir gradientes pequeños para valores negativos.

Otras funciones de activación relevantes incluyen la sigmoide, $\sigma(x) = 1/(1 + e^{-x})$, que comprime valores al rango $(0, 1)$ pero sufre de gradientes desvanecientes para valores extremos; la tangente hiperbólica, $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$, que mapea al rango $(-1, 1)$; y *softmax*, utilizada en capas de salida para tareas de clasificación multi-clase:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}} \quad (2.2.8)$$

donde K es el número de clases. La función *softmax* garantiza que las salidas sean no negativas y sumen uno, interpretándose como probabilidades posteriores de clase.

2.2.4. Algoritmo de Retropropagación

El entrenamiento de redes neuronales profundas se realiza mediante el algoritmo de retropropagación (del inglés, *backpropagation*), que calcula eficientemente el gradiente de una función de pérdida \mathcal{L} respecto a todos los parámetros de la red mediante aplicación recursiva de la regla de la cadena del cálculo diferencial [38, 41]. Considérese una red neuronal con L capas, donde cada capa l realiza la transformación:

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \quad (2.2.9)$$

$$a^{(l)} = f(z^{(l)}) \quad (2.2.10)$$

donde $z^{(l)}$ representa la entrada ponderada (*pre-activation*), $a^{(l)}$ es la activación de la capa l , $W^{(l)}$ y $b^{(l)}$ son los parámetros (pesos y sesgos), y $f(\cdot)$ es la función de activación. La propagación hacia adelante (*forward pass*) evalúa estas ecuaciones secuencialmente desde la entrada hasta la salida.

El objetivo del entrenamiento es minimizar una función de pérdida $\mathcal{L}(a^{(L)}, y)$ que cuantifica la discrepancia entre la predicción de la red $a^{(L)}$ y el valor objetivo y . Para actualizar los parámetros mediante descenso de gradiente, se requiere calcular $\partial\mathcal{L}/\partial W^{(l)}$ y $\partial\mathcal{L}/\partial b^{(l)}$ para toda capa l . La retropropagación logra esto mediante la definición del error de retropropagación $\delta^{(l)}$ en cada capa:

$$\delta^{(l)} = \frac{\partial\mathcal{L}}{\partial z^{(l)}} \quad (2.2.11)$$

Para la capa de salida L , el error de retropropagación se calcula directamente mediante la regla de la cadena:

$$\delta^{(L)} = \frac{\partial\mathcal{L}}{\partial a^{(L)}} \odot f'(z^{(L)}) \quad (2.2.12)$$

donde \odot denota el producto elemento a elemento (producto de Hadamard). Para capas intermedias, el error se propaga hacia atrás mediante:

$$\delta^{(l)} = \left((W^{(l+1)})^T \delta^{(l+1)} \right) \odot f'(z^{(l)}) \quad (2.2.13)$$

Esta ecuación recursiva constituye el núcleo del algoritmo de retropropagación: el error en la capa l se obtiene multiplicando el error de la capa siguiente por la matriz de pesos transpuesta (propagación del error hacia atrás a través de la transformación lineal), seguido de una modulación elemento a elemento por la derivada de la función de activación.

Una vez calculados los errores $\delta^{(l)}$ para todas las capas, los gradientes respecto a los parámetros

se obtienen como:

$$\frac{\partial \mathcal{L}}{\partial W^{(l)}} = \delta^{(l)} (a^{(l-1)})^T \quad (2.2.14)$$

$$\frac{\partial \mathcal{L}}{\partial b^{(l)}} = \delta^{(l)} \quad (2.2.15)$$

Los parámetros se actualizan mediante descenso de gradiente:

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial W^{(l)}}, \quad b^{(l)} \leftarrow b^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial b^{(l)}} \quad (2.2.16)$$

donde η es la tasa de aprendizaje (*learning rate*), un hiperparámetro que controla la magnitud de las actualizaciones de parámetros.

La complejidad computacional de la retropropagación es del mismo orden que la propagación hacia adelante, típicamente $O(W)$ donde W es el número total de pesos en la red. Esta eficiencia computacional, combinada con la disponibilidad de unidades de procesamiento gráfico (GPUs) altamente paralelizables, ha permitido el entrenamiento de redes con cientos de millones de parámetros en conjuntos de datos masivos.

2.2.5. Algoritmos de Optimización

El algoritmo básico de descenso de gradiente estocástico (*Stochastic Gradient Descent*, SGD) actualiza los parámetros θ de la red utilizando el gradiente calculado sobre una muestra individual o un mini-lote pequeño de datos:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \mathcal{L}(\theta_t; x^{(i)}, y^{(i)}) \quad (2.2.17)$$

donde t indexa la iteración de actualización, y $(x^{(i)}, y^{(i)})$ representa una muestra de entrenamiento. A diferencia del descenso de gradiente por lotes que utiliza el conjunto de entrenamiento completo, SGD proporciona actualizaciones frecuentes que aceleran la convergencia, aunque con mayor varianza en la dirección de descenso.

Una mejora fundamental sobre SGD es la incorporación de momentum, que acumula un promedio móvil exponencialmente ponderado de gradientes pasados:

$$v_t = \beta v_{t-1} + \eta \nabla_{\theta} \mathcal{L}(\theta_t) \quad (2.2.18)$$

$$\theta_{t+1} = \theta_t - v_t \quad (2.2.19)$$

donde v_t representa la velocidad acumulada, y β es el coeficiente de momentum (típicamente 0.9). El momentum reduce oscilaciones en direcciones de alta curvatura y acelera la

convergencia en direcciones consistentes del espacio de parámetros [38].

El optimizador Adam (*Adaptive Moment Estimation*) representa el estado del arte en algoritmos de optimización para *deep learning*, combinando las ventajas de momentum con tasas de aprendizaje adaptativas por parámetro [42]. Adam mantiene estimaciones de los momentos de primer y segundo orden de los gradientes:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2.2.20)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (2.2.21)$$

donde $g_t = \nabla_{\theta} \mathcal{L}(\theta_t)$ es el gradiente en el tiempo t , m_t es el primer momento (media), v_t es el segundo momento no centrado (varianza no centrada), y $\beta_1, \beta_2 \in [0, 1)$ son tasas de decaimiento exponencial (valores típicos: $\beta_1 = 0,9$, $\beta_2 = 0,999$).

Dado que m_t y v_t se inicializan en cero, presentan sesgo hacia cero en las primeras iteraciones. Adam corrige este sesgo mediante:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (2.2.22)$$

La actualización de parámetros incorpora una tasa de aprendizaje adaptativa por parámetro:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (2.2.23)$$

donde $\epsilon = 10^{-8}$ es una constante pequeña para estabilidad numérica. El término $\sqrt{\hat{v}_t}$ normaliza el tamaño de actualización por parámetro basándose en la magnitud histórica de los gradientes, proporcionando actualizaciones más grandes para parámetros con gradientes consistentemente pequeños y actualizaciones más pequeñas para parámetros con gradientes grandes o ruidosos.

Adam ha demostrado convergencia robusta en una amplia variedad de arquitecturas de *deep learning* y es particularmente efectivo en aplicaciones de visión médica, donde los conjuntos de datos suelen ser de tamaño moderado y la optimización cuidadosa es crítica para evitar sobreajuste [2]. La combinación de momentum adaptativo y tasas de aprendizaje por parámetro permite que Adam funcione razonablemente bien con hiperparámetros por defecto, reduciendo la necesidad de ajuste extenso de hiperparámetros.

2.3. Arquitecturas Residuales Profundas

Las redes neuronales convolucionales presentadas en la Sección 2.2 pueden componerse en arquitecturas de profundidad variable, siendo la profundidad un factor determinante en su capacidad de aprendizaje: redes más profundas pueden aprender representaciones jerárquicas más complejas mediante la composición de múltiples transformaciones no lineales. Sin embargo, el entrenamiento de redes extremadamente profundas (con más de 20-30 capas) presentaba desafíos significativos antes del desarrollo de arquitecturas residuales. La observación empírica de que redes más profundas exhibían mayor error de entrenamiento que redes menos profundas sugería la existencia de dificultades de optimización fundamentales que no podían atribuirse únicamente al sobreajuste. Las Redes Neuronales Residuales (del inglés, *Residual Neural Networks*, ResNet), introducidas por He et al. [29], revolucionaron el diseño de arquitecturas profundas mediante la incorporación de conexiones residuales que permiten el entrenamiento efectivo de redes con cientos de capas.

2.3.1. El Problema de Degradación en Redes Profundas

La intuición convencional sugeriría que agregar capas adicionales a una red neuronal no debería degradar su desempeño: en el peor de los casos, las capas adicionales podrían aprender la función identidad, replicando el desempeño de la red menos profunda. Sin embargo, experimentos empíricos demostraron un fenómeno contraintuitivo denominado *degradación*: a medida que la profundidad de la red aumenta más allá de cierto umbral, tanto el error de entrenamiento como el error de prueba comienzan a aumentar [29].

Este problema de degradación no puede explicarse mediante sobreajuste, ya que el error de entrenamiento (no solo el error de generalización) es superior en redes más profundas. He et al. hipotetizaron que la dificultad radica en que los solucionadores de optimización tienen dificultad para aproximar funciones identidad mediante múltiples capas no lineales. Adicionalmente, el problema del gradiente desvaneciente, donde los gradientes se atenúan exponencialmente al propagarse hacia capas tempranas, complica el entrenamiento de redes muy profundas, aunque técnicas como normalización por lotes y funciones de activación ReLU mitigan parcialmente este efecto [43, 44].

Para cuantificar la degradación, considérense dos arquitecturas: una red de n capas con error de entrenamiento ϵ_n , y una red de $n + k$ capas con error ϵ_{n+k} . El fenómeno de degradación se manifiesta cuando $\epsilon_{n+k} > \epsilon_n$ a pesar de que teóricamente las k capas adicionales podrían aprender transformaciones identidad. Experimentos en ImageNet demostraron que redes de 56 capas con arquitectura plana (*plain*) exhibían error de entrenamiento 0.5 % superior a redes de 20 capas, evidenciando la naturaleza empírica del problema [29].

2.3.2. Conexiones Residuales y Bloques Residuales

La arquitectura ResNet aborda el problema de degradación mediante la introducción de conexiones residuales (del inglés, *skip connections* o *shortcut connections*), que permiten que el gradiente fluya directamente a través de la red sin atenuación. En lugar de aprender directamente un mapeo deseado $\mathcal{H}(x)$ desde la entrada x hasta la salida, los bloques residuales aprenden el mapeo residual:

$$\mathcal{F}(x) = \mathcal{H}(x) - x \quad (2.3.1)$$

La salida del bloque residual se define entonces como:

$$y = \mathcal{F}(x, \{W_i\}) + x \quad (2.3.2)$$

donde $\mathcal{F}(x, \{W_i\})$ representa el mapeo residual implementado por las capas con pesos $\{W_i\}$, y la suma $+x$ representa la conexión de atajo (*shortcut connection*). Si el mapeo óptimo es cercano a la identidad, es más fácil para el optimizador ajustar $\mathcal{F}(x)$ hacia cero que forzar múltiples capas no lineales a aproximar la función identidad directamente.

La hipótesis fundamental de ResNet es que **es más fácil optimizar el mapeo residual $\mathcal{F}(x)$ que el mapeo original $\mathcal{H}(x)$** . En el caso extremo donde el mapeo identidad es óptimo ($\mathcal{H}(x) = x$), es trivial para el optimizador ajustar los pesos de las capas residuales hacia cero, forzando $\mathcal{F}(x) \approx 0$ y obteniendo $y \approx x$.

He et al. propusieron dos arquitecturas de bloques residuales [29]:

1. Bloque básico (utilizado en ResNet-18 y ResNet-34):

$$y = \text{ReLU}(x + W_2\sigma(W_1x + b_1) + b_2) \quad (2.3.3)$$

donde σ representa la función de activación ReLU, W_1 y W_2 son matrices de pesos de capas convolucionales de 3×3 , y b_1, b_2 son sesgos. El bloque básico consta de dos capas convolucionales con normalización por lotes y ReLU entre ellas.

2. Bloque cuello de botella (del inglés, *bottleneck block*; utilizado en ResNet-50, ResNet-101, ResNet-152):

$$y = \text{ReLU}(x + W_3\sigma(W_2\sigma(W_1x + b_1) + b_2) + b_3) \quad (2.3.4)$$

donde W_1 es una convolución de 1×1 que reduce la dimensionalidad, W_2 es una convolución de 3×3 que procesa características en dimensión reducida, y W_3 es una convolución de 1×1 que restaura la dimensionalidad. Esta arquitectura reduce significativamente el costo computacional

en redes muy profundas.

Cuando las dimensiones de la entrada x y la salida y difieren (por cambios en el número de canales o resolución espacial), la conexión de atajo debe implementarse mediante una proyección lineal:

$$y = \mathcal{F}(x, \{W_i\}) + W_s x \quad (2.3.5)$$

donde W_s es una matriz de proyección implementada mediante convolución de 1×1 con *stride* apropiado para igualar las dimensiones.

2.3.3. Arquitecturas de la Familia ResNet

La familia ResNet comprende múltiples arquitecturas que varían en profundidad, desde ResNet-18 (18 capas con pesos) hasta ResNet-152 (152 capas). La Tabla 2.3.1 presenta las configuraciones arquitectónicas de las variantes más utilizadas.

Tabla 2.3.1: Arquitecturas de la familia ResNet. Los números entre paréntesis indican el número de bloques residuales en cada etapa.

Capa	Salida	ResNet-18	ResNet-34	ResNet-50	ResNet-101
Conv1	112×112	$7 \times 7, 64$, stride 2			
Pool	56×56	3×3 max pool, stride 2			
Conv2_x	56×56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3_x	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Conv4_x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$
Conv5_x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1	Global Average Pooling, FC 1000, Softmax			
Parámetros		11.7M	21.8M	25.6M	44.5M

La arquitectura base consta de cinco etapas (Conv1, Conv2_x, Conv3_x, Conv4_x, Conv5_x), donde cada etapa opera en una resolución espacial específica. Las resoluciones espaciales se reducen progresivamente mediante convoluciones con *stride* 2 al inicio de las etapas Conv3_x, Conv4_x y Conv5_x. ResNet-18 y ResNet-34 utilizan bloques básicos, mientras que ResNet-50, ResNet-101 y ResNet-152 emplean bloques cuello de botella para controlar la complejidad computacional. La capa final aplica *global average pooling* sobre los mapas de características

espaciales, reduciendo cada canal a un valor escalar, seguido de una capa completamente conectada para clasificación.

2.3.4. Normalización por Lotes

La normalización por lotes (del inglés, *batch normalization*, BN) es un componente esencial de las arquitecturas ResNet, aplicado después de cada capa convolucional y antes de la función de activación [43]. BN normaliza las activaciones de cada capa utilizando estadísticas del mini-lote actual, reduciendo la dependencia en la inicialización de pesos y permitiendo tasas de aprendizaje más altas.

Para un mini-lote $\mathcal{B} = \{x_1, x_2, \dots, x_m\}$ de tamaño m , la normalización por lotes calcula la media y varianza del mini-lote:

$$\mu_{\mathcal{B}} = \frac{1}{m} \sum_{i=1}^m x_i \quad (2.3.6)$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad (2.3.7)$$

Las activaciones se normalizan:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad (2.3.8)$$

donde ϵ (típicamente 10^{-5}) es una constante pequeña para estabilidad numérica. Finalmente, se aplica una transformación afín aprendible:

$$y_i = \gamma \hat{x}_i + \beta \quad (2.3.9)$$

donde γ y β son parámetros aprendibles que permiten a la red recuperar la capacidad expresiva completa si la normalización resulta subóptima.

Durante la inferencia, BN utiliza estadísticas globales (media y varianza estimadas sobre el conjunto de entrenamiento completo mediante promedio móvil) en lugar de estadísticas del mini-lote, garantizando predicciones deterministas.

BN proporciona múltiples beneficios [43]: (1) reduce el desplazamiento de covarianza interna (*internal covariate shift*), donde las distribuciones de activaciones cambian durante el entrenamiento; (2) permite tasas de aprendizaje significativamente más altas sin divergencia; (3) actúa como regularizador, reduciendo la necesidad de *dropout*; (4) permite la inicialización de pesos menos cuidadosa. Estos factores son particularmente relevantes para redes residuales profundas.

2.3.5. Ventajas de Arquitecturas Residuales para Imágenes Médicas

Las arquitecturas ResNet han demostrado ser particularmente efectivas para aplicaciones en imágenes médicas por múltiples razones [45, 46]:

1. Gradientes estables: Las conexiones residuales proporcionan caminos de gradiente directos desde las capas profundas hasta las capas tempranas, mitigando el problema del gradiente desvaneciente. Durante la retropropagación, el gradiente respecto a la entrada de un bloque residual es:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial y} \left(1 + \frac{\partial \mathcal{F}}{\partial x} \right) \quad (2.3.10)$$

El término constante 1 garantiza que el gradiente no se atenúe completamente, incluso si $\partial \mathcal{F} / \partial x$ es pequeño.

2. Eficiencia de parámetros: ResNet-18, con 11.7 millones de parámetros, proporciona un balance óptimo entre capacidad expresiva y eficiencia computacional. Esta propiedad es crítica en aplicaciones médicas donde los conjuntos de datos son típicamente más pequeños que ImageNet (1.2 millones de imágenes), y arquitecturas muy profundas pueden sobreajustar.

3. Aprendizaje jerárquico robusto: Las conexiones residuales permiten que capas tempranas aprendan características de bajo nivel (bordes, texturas) mientras capas profundas aprenden representaciones anatómicas complejas, sin degradación de desempeño asociada a la profundidad extrema.

4. Transferibilidad: Modelos ResNet pre-entrenados en ImageNet han demostrado transferibilidad excepcional a dominios médicos mediante *fine-tuning*, como se discutirá en la Sección 2.4. Las representaciones aprendidas en ImageNet capturan características genéricas de imágenes naturales que son parcialmente relevantes para imágenes médicas.

En el contexto específico de detección de *landmarks* anatómicos en radiografías de tórax, las arquitecturas residuales proporcionan la profundidad necesaria para capturar la complejidad de estructuras anatómicas distribuidas espacialmente, mientras mantienen gradientes estables que facilitan el aprendizaje de regresión de coordenadas precisa.

2.4. Aprendizaje por Transferencia en Imágenes Médicas

Las arquitecturas residuales presentadas en la Sección 2.3, particularmente ResNet con sus múltiples variantes de profundidad, han demostrado capacidad excepcional para aprender representaciones jerárquicas complejas en tareas de visión por computadora. Sin embargo, el entrenamiento de estas redes profundas desde inicialización aleatoria requiere conjuntos de datos masivos para lograr convergencia robusta y generalización adecuada. En el dominio médico, la adquisición de grandes conjuntos de datos etiquetados enfrenta barreras significativas: costos elevados de anotación por expertos radiólogos, consideraciones de privacidad de pacientes, y la naturaleza inherentemente limitada de casos patológicos específicos. Los conjuntos de datos médicos típicamente contienen cientos a miles de imágenes, en contraste con los millones de ejemplos disponibles en dominios de visión por computadora general como ImageNet. El aprendizaje por transferencia (del inglés, *transfer learning*) constituye un paradigma fundamental que permite aprovechar representaciones aprendidas en conjuntos de datos masivos de un dominio fuente para mejorar significativamente el desempeño en tareas del dominio objetivo con datos limitados [27, 47, 48].

2.4.1. Pre-entrenamiento en ImageNet y Representaciones Transferibles

ImageNet representa el conjunto de datos de referencia para pre-entrenamiento de modelos de visión por computadora, comprendiendo aproximadamente 1.2 millones de imágenes de entrenamiento distribuidas en 1,000 categorías de objetos cotidianos [12]. Los modelos entrenados en ImageNet, incluyendo las arquitecturas ResNet descritas en la Sección 2.3, aprenden representaciones jerárquicas de características visuales mediante optimización de la función de pérdida de clasificación multi-clase. Formalmente, el pre-entrenamiento en el dominio fuente se define como:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^{N_s} \mathcal{L}_{\text{source}}(f_{\theta}(x_i^s), y_i^s) \quad (2.4.1)$$

donde f_{θ} representa la red neuronal con parámetros θ , $\mathcal{D}_{\text{source}} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ es el conjunto de datos fuente (ImageNet con $N_s \approx 1,2 \times 10^6$), y $\mathcal{L}_{\text{source}}$ es típicamente la entropía cruzada categórica para clasificación.

La hipótesis central del aprendizaje por transferencia es que las características de bajo y medio nivel aprendidas en ImageNet poseen utilidad general para tareas de visión por

computadora, incluso en dominios substancialmente diferentes como imágenes médicas. Las capas convolucionales tempranas de redes pre-entrenadas detectan bordes, texturas, y gradientes de intensidad genéricos que son relevantes para cualquier tarea de procesamiento de imágenes. Las capas intermedias capturan estructuras geométricas de complejidad creciente (formas, patrones, composiciones espaciales), mientras que las capas profundas aprenden representaciones más específicas del dominio fuente [27, 49].

Yosinski et al. [27] demostraron empíricamente que la transferibilidad de características decae con la profundidad de la red cuando los dominios fuente y objetivo son muy diferentes: las primeras capas convolucionales aprenden detectores casi universales, mientras que capas superiores requieren adaptación substancial al dominio objetivo. Este hallazgo motivó estrategias de *fine-tuning* selectivo que se discuten en la siguiente subsección.

Estudios recientes han expandido significativamente la comprensión de *transfer learning* en medicina. Azizi et al. [49] demostraron en *Nature Biomedical Engineering* que modelos pre-entrenados mediante aprendizaje auto-supervisado en conjuntos de datos médicos multi-institucionales exhiben transferibilidad superior a modelos pre-entrenados exclusivamente en ImageNet, sugiriendo que el pre-entrenamiento en dominios más cercanos al objetivo (radiografías médicas en general) proporciona ventajas adicionales. Moor et al. [47], en un artículo perspectivo en *Nature*, argumentan que los modelos fundacionales (del inglés, *foundation models*) pre-entrenados en datos médicos masivos representan el futuro del *transfer learning* médico, reduciendo la dependencia de ImageNet. Sin embargo, para tareas específicas como detección de *landmarks* anatómicos en radiografías de tórax, el pre-entrenamiento en ImageNet continúa siendo el estándar actual debido a la disponibilidad limitada de modelos fundacionales médicos especializados [48].

2.4.2. Estrategias de Fine-Tuning y Adaptación al Dominio Médico

El *transfer learning* puede implementarse mediante dos estrategias principales, que difieren en qué parámetros de la red se actualizan durante el entrenamiento en el dominio objetivo.

Extracción de características (del inglés, *feature extraction*): Los pesos de las capas convolucionales pre-entrenadas se congelan completamente, y solo se entrenan las capas completamente conectadas finales específicas de la tarea objetivo. Matemáticamente, el modelo se descompone como $f_{\theta} = g_{\phi}(h_{\psi}(x))$, donde h_{ψ} representa el extractor de características convolucionales con pesos congelados $\psi = \psi_{\text{ImageNet}}^*$, y g_{ϕ} representa las capas de tarea específica (típicamente una o dos capas completamente conectadas) con parámetros entrenables

ϕ :

$$\phi^* = \arg \min_{\phi} \sum_{j=1}^{N_t} \mathcal{L}_{\text{target}}(g_{\phi}(h_{\psi^*}(x_j^t)), y_j^t) \quad (2.4.2)$$

donde $\mathcal{D}_{\text{target}} = \{(x_j^t, y_j^t)\}_{j=1}^{N_t}$ es el conjunto de datos objetivo (típicamente $N_t \ll N_s$), y $\mathcal{L}_{\text{target}}$ es la función de pérdida específica de la tarea (por ejemplo, error cuadrático medio para regresión de coordenadas, como se discutirá en la Sección 2.5). Esta estrategia es computacionalmente eficiente y apropiada cuando el conjunto de datos objetivo es muy pequeño ($N_t < 1,000$ típicamente), mitigando el riesgo de sobreajuste al limitar drásticamente el número de parámetros entrenables.

Ajuste fino (del inglés, *fine-tuning*): Todos los parámetros de la red, o un subconjunto de capas superiores, se ajustan en el dominio objetivo utilizando la inicialización pre-entrenada:

$$\theta_{\text{fine-tuned}} = \arg \min_{\theta} \sum_{j=1}^{N_t} \mathcal{L}_{\text{target}}(f_{\theta}(x_j^t), y_j^t), \quad \text{con } \theta(t=0) = \theta_{\text{ImageNet}}^* \quad (2.4.3)$$

El *fine-tuning* permite que la red adapte sus representaciones internas al dominio objetivo, pero requiere conjuntos de datos de tamaño moderado ($N_t > 1,000$ típicamente) y selección cuidadosa de hiperparámetros de optimización para evitar sobreajuste o colapso catastrófico de las características pre-entrenadas útiles.

Una estrategia avanzada es la aplicación de tasas de aprendizaje diferenciales (del inglés, *discriminative learning rates*): capas tempranas, que capturan características de bajo nivel genéricas, se actualizan con tasas de aprendizaje pequeñas o se congelan completamente, mientras que capas profundas y capas específicas de la tarea se entrenan con tasas de aprendizaje más altas:

$$\theta_{\text{early}}^{(t+1)} = \theta_{\text{early}}^{(t)} - \eta_{\text{low}} \nabla_{\theta_{\text{early}}} \mathcal{L}_{\text{target}} \quad (2.4.4)$$

$$\theta_{\text{deep}}^{(t+1)} = \theta_{\text{deep}}^{(t)} - \eta_{\text{high}} \nabla_{\theta_{\text{deep}}} \mathcal{L}_{\text{target}} \quad (2.4.5)$$

donde $\eta_{\text{high}}/\eta_{\text{low}} \approx 10$ es una configuración típica. Esta estrategia preserva las representaciones de bajo nivel útiles mientras permite adaptación substancial en capas superiores que requieren especialización al dominio médico.

El *fine-tuning* progresivo (del inglés, *progressive unfreezing*) constituye una variante donde inicialmente solo las capas finales son entrenables, y gradualmente se descongelan capas anteriores a medida que avanza el entrenamiento. Nguyen et al. [50] presentan un análisis exhaustivo de estrategias de *transfer learning* multi-etapa en imágenes médicas, demostrando que el descongelamiento progresivo proporciona convergencia más estable en conjuntos de

datos médicos pequeños comparado con *fine-tuning* simultáneo de todas las capas.

2.4.3. Brecha de Dominio y Adaptación para Radiografías de Tórax

Existe una brecha de dominio (del inglés, *domain gap*) substancial entre imágenes naturales de ImageNet e imágenes médicas de radiografías de tórax. ImageNet contiene fotografías RGB de objetos cotidianos en escenarios naturales con iluminación variada, mientras que las radiografías de tórax son imágenes de canal único (escala de grises) que representan proyecciones bidimensionales de atenuación de rayos X de estructuras anatómicas tridimensionales, como se describió en la Sección 2.1. Las distribuciones de intensidad, texturas, y estructuras geométricas difieren fundamentalmente entre ambos dominios.

A pesar de esta disparidad, estudios empíricos han demostrado consistentemente que el *transfer learning* desde ImageNet proporciona mejoras substanciales sobre el entrenamiento desde inicialización aleatoria en tareas de análisis de radiografías de tórax. Tajbakhsh et al. [22] evaluaron *transfer learning* en cuatro tareas de análisis de imágenes médicas, incluyendo detección de nódulos pulmonares en radiografías de tórax, demostrando mejoras de 5-10 % en AUC al utilizar pre-entrenamiento de ImageNet versus inicialización aleatoria, particularmente en regímenes de datos limitados ($N_t < 5,000$).

Investigaciones recientes han explorado técnicas de adaptación de dominio específicamente diseñadas para abordar la brecha entre ImageNet y radiografías médicas. Sanchez et al. [51] propusieron CX-DaGAN en *IEEE Transactions on Medical Imaging*, una red generativa adversarial para adaptación de dominio en diagnóstico de neumonía con conjuntos de datos de radiografías de tórax extremadamente pequeños. Guan y Liu [52] presentaron un análisis comprehensivo de técnicas de adaptación de dominio para análisis de imágenes médicas en *IEEE Transactions on Biomedical Engineering*, categorizando enfoques en: (1) adaptación basada en discrepancia de características, (2) adaptación adversarial, y (3) adaptación mediante reconstrucción. Estos métodos avanzados buscan alinear las distribuciones de características entre dominios fuente y objetivo, reduciendo la brecha de dominio y mejorando la transferibilidad.

La transferencia desde ImageNet a radiografías de tórax requiere adaptaciones arquitectónicas específicas. Las arquitecturas ResNet estándar esperan imágenes RGB (3 canales de entrada), mientras que las radiografías de tórax son de canal único. La práctica común consiste en replicar la imagen de escala de grises a tres canales ($x_{\text{RGB}} = [x, x, x]$), preservando los pesos pre-entrenados de la primera capa convolucional sin modificación. Alternativamente, los pesos del filtro de entrada pueden promediarse a través de los tres canales RGB y utilizarse para procesar el canal único directamente. La capa completamente conectada final pre-entrenada, que tiene 1,000 salidas correspondientes a las clases de ImageNet, debe reemplazarse con

una capa específica de la tarea: para detección de *landmarks*, esta capa predice $2K$ valores continuos representando coordenadas (x, y) de K *landmarks*, sin función *softmax*. La función de pérdida apropiada para esta tarea de regresión de coordenadas se discutirá en detalle en la Sección 2.5.

El aprendizaje por transferencia representa un componente fundamental para el desarrollo de sistemas de *deep learning* en imágenes médicas con conjuntos de datos limitados. La combinación de pre-entrenamiento en ImageNet, estrategias de *fine-tuning* con tasas de aprendizaje diferenciales, y adaptaciones arquitectónicas apropiadas permite la construcción de modelos robustos que aprovechan conocimiento visual genérico mientras se especializan en características anatómicas específicas del dominio médico [2, 46, 48].

2.5. Funciones de Pérdida para Regresión de Coordenadas

La estrategia de aprendizaje por transferencia establecida en la Sección 2.4 proporciona una inicialización favorable de los pesos de la red mediante pre-entrenamiento en ImageNet, pero la función de pérdida determina fundamentalmente qué aprende la red durante el *fine-tuning* en el dominio objetivo. Para la tarea de detección de *landmarks* anatómicos, la red debe aprender un mapeo $f_\theta : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{2K}$ desde una imagen de entrada de dimensiones $H \times W$ a un vector de $2K$ valores continuos representando las coordenadas (x_k, y_k) de K *landmarks*. Esta tarea de regresión de coordenadas presenta desafíos específicos: requiere precisión a nivel de píxel individual, debe ser robusta ante variabilidad anatómica y calidad de imagen heterogénea, y puede beneficiarse de la incorporación explícita de conocimiento anatómico a priori mediante restricciones geométricas. Esta sección analiza funciones de pérdida especializadas para regresión de coordenadas, comenzando con el error cuadrático medio como línea base, seguido por *Wing Loss* que amplifica gradientes para errores pequeños, y concluyendo con funciones de pérdida basadas en restricciones geométricas que incorporan conocimiento anatómico de simetría bilateral y preservación de distancias [15, 53, 54].

2.5.1. Error Cuadrático Medio

El Error Cuadrático Medio (del inglés, *Mean Squared Error*, MSE) constituye la función de pérdida estándar para tareas de regresión, incluyendo regresión de coordenadas de *landmarks*. Para un conjunto de K *landmarks*, donde cada *landmark* k tiene coordenadas ground truth (x_k, y_k) y coordenadas predichas (\hat{x}_k, \hat{y}_k) , MSE se define como:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{K} \sum_{k=1}^K [(x_k - \hat{x}_k)^2 + (y_k - \hat{y}_k)^2] \quad (2.5.1)$$

Esta formulación puede expresarse de manera más compacta utilizando notación vectorial. Definiendo $p_k = (x_k, y_k)^T \in \mathbb{R}^2$ como el vector de posición del *landmark* k y $\hat{p}_k = (\hat{x}_k, \hat{y}_k)^T$ como su predicción correspondiente, la función de pérdida MSE se escribe:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{K} \sum_{k=1}^K \|p_k - \hat{p}_k\|_2^2 = \frac{1}{K} \sum_{k=1}^K (p_k - \hat{p}_k)^T (p_k - \hat{p}_k) \quad (2.5.2)$$

donde $\|\cdot\|_2$ denota la norma Euclidiana. El gradiente de MSE respecto a la predicción del

landmark k es:

$$\frac{\partial \mathcal{L}_{\text{MSE}}}{\partial \hat{p}_k} = \frac{2}{K}(\hat{p}_k - p_k) \quad (2.5.3)$$

MSE posee propiedades matemáticas deseables: es una función convexa, diferenciable en todos los puntos, y su mínimo global coincide con la media de los datos objetivo. Durante la retropropagación, el gradiente proporcionado por MSE crece linealmente con la magnitud del error de predicción ($\|\nabla \mathcal{L}_{\text{MSE}}\| \propto \|p_k - \hat{p}_k\|$). Sin embargo, esta característica resulta problemática para la detección precisa de *landmarks* anatómicos por múltiples razones [15, 55].

Primero, MSE exhibe **sensibilidad desbalanceada a errores de diferente magnitud**. *Landmarks* con predicciones muy incorrectas (errores de decenas de píxeles) generan gradientes dominantes que pueden enmascarar la señal de aprendizaje de *landmarks* con errores pequeños (1-2 píxeles). Segundo, la penalización cuadrática amplifica el impacto de valores atípicos (*outliers*): un solo *landmark* mal predicho contribuye con $error^2$ a la pérdida total, potencialmente desestabilizando el entrenamiento en presencia de oclusiones parciales o artefactos de imagen. Tercero, para errores pequeños ($\|p_k - \hat{p}_k\| < 1$ píxel), el gradiente de MSE se vuelve proporcionalmente pequeño ($\|\nabla \mathcal{L}_{\text{MSE}}\| \approx 2|error|/K \ll 1$), debilitando la señal de aprendizaje precisamente en el régimen donde se requiere refinamiento fino de coordenadas. Finalmente, MSE trata cada *landmark* independientemente, ignorando restricciones geométricas inherentes a la anatomía torácica como simetría bilateral y distancias anatómicas características.

Estas limitaciones motivaron el desarrollo de funciones de pérdida especializadas que amplifican gradientes en el régimen de errores pequeños mientras mantienen robustez ante errores grandes, y que incorporan conocimiento anatómico a priori mediante términos de regularización geométrica.

2.5.2. Wing Loss: Amplificación de Gradientes para Errores Pequeños

Feng et al. [15] propusieron *Wing Loss* como una función de pérdida diseñada específicamente para localización robusta de *landmarks* faciales, abordando las limitaciones de MSE mediante amplificación selectiva de gradientes en la región de errores pequeños. La idea central es modificar el comportamiento de la función de pérdida para proporcionar gradientes grandes cuando el error de predicción es pequeño (facilitando refinamiento preciso de coordenadas), mientras se mantienen gradientes moderados para errores grandes (proporcionando robustez).

Wing Loss se define mediante una función no lineal por partes:

$$\mathcal{L}_{\text{wing}}(x) = \begin{cases} w \ln \left(1 + \frac{|x|}{\epsilon} \right) & \text{si } |x| < w \\ |x| - C & \text{si } |x| \geq w \end{cases} \quad (2.5.4)$$

donde x representa el error de coordenada, w es el ancho de la región no lineal (típicamente $w \in [5, 10]$ píxeles para imágenes médicas), ϵ es un parámetro de curvatura que controla la suavidad de la transición (típicamente $\epsilon = 2,0$), y $C = w - w \ln(1 + w/\epsilon)$ es una constante que garantiza continuidad de la función en $|x| = w$. La región $|x| < w$ exhibe comportamiento logarítmico que amplifica gradientes para errores pequeños, mientras que la región $|x| \geq w$ presenta comportamiento lineal similar a la pérdida L1 absoluta, proporcionando robustez ante *outliers*.

Para la detección de K *landmarks*, *Wing Loss* se aplica al error radial de cada *landmark*:

$$\mathcal{L}_{\text{Wing}} = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_{\text{wing}}(\|p_k - \hat{p}_k\|_2) \quad (2.5.5)$$

El comportamiento de amplificación de gradientes de *Wing Loss* se comprende mediante el análisis de su derivada. Para $|x| < w$, la derivada es:

$$\frac{\partial \mathcal{L}_{\text{wing}}(x)}{\partial x} = \frac{w}{\epsilon + |x|} \cdot \text{sign}(x) \quad (2.5.6)$$

donde $\text{sign}(x) = \pm 1$ indica la dirección del error. Para $|x| \geq w$, la derivada es:

$$\frac{\partial \mathcal{L}_{\text{wing}}(x)}{\partial x} = \text{sign}(x) \quad (2.5.7)$$

La amplificación de gradientes se manifiesta en el límite de errores pequeños:

$$\lim_{x \rightarrow 0} \frac{\partial \mathcal{L}_{\text{wing}}(x)}{\partial x} = \frac{w}{\epsilon} \cdot \text{sign}(x) \quad (2.5.8)$$

Para la configuración típica $w = 5$ y $\epsilon = 2$, el gradiente en $x \rightarrow 0$ es $w/\epsilon = 2,5$, significativamente mayor que el gradiente de MSE en el mismo punto ($\partial \mathcal{L}_{\text{MSE}}/\partial x = 2x/K \approx 0$ cuando $x \rightarrow 0$). En el punto de transición $|x| = w$, el gradiente de *Wing Loss* es exactamente ± 1 , garantizando continuidad de la derivada. Para errores grandes $|x| \gg w$, el gradiente satura en ± 1 , similar a la pérdida L1, proporcionando robustez ante predicciones extremadamente incorrectas que podrían desestabilizar el entrenamiento si se penalizaran cuadráticamente.

La comparación formal con MSE ilustra la ventaja de *Wing Loss*. El gradiente de MSE respecto

al error x es $\partial \mathcal{L}_{\text{MSE}} / \partial x = 2x/K$, que decrece linealmente hacia cero a medida que el error disminuye. En contraste, *Wing Loss* mantiene un gradiente constante y grande (w/ϵ) en la región de errores pequeños, proporcionando una señal de aprendizaje consistente para el refinamiento fino de coordenadas. Esta propiedad es particularmente relevante para la detección de *landmarks* anatómicos en radiografías de tórax, donde la variabilidad inter-paciente de posiciones de *landmarks* es típicamente del orden de 10-20 píxeles, y se requiere precisión de localización a nivel de píxel individual para aplicaciones clínicas.

Extensiones recientes de *Wing Loss* incluyen *Adaptive Wing Loss* propuesto por Liu et al. [55], que adapta dinámicamente los parámetros w y ϵ durante el entrenamiento para equilibrar robustez inicial y precisión final. Cheng et al. [54] demostraron en *Medical Image Analysis* que la incorporación de perturbaciones controladas en las entradas combinada con *Wing Loss* mejora significativamente la precisión de localización de *landmarks* en imágenes médicas.

2.5.3. Restricciones Geométricas: Symmetry Loss y Distance Preservation Loss

Las funciones de pérdida basadas en regresión directa de coordenadas (MSE, *Wing Loss*) tratan cada *landmark* independientemente, ignorando relaciones geométricas inherentes a la anatomía humana. En el contexto específico de radiografías de tórax, la anatomía presenta propiedades geométricas consistentes que pueden explotarse como restricciones: la simetría bilateral aproximada del tórax implica que pares de *landmarks* homólogos (izquierdo-derecho) deben ser aproximadamente simétricos respecto a la línea media, y las distancias entre *landmarks* específicos exhiben variabilidad limitada en poblaciones sanas. La incorporación de estas restricciones geométricas como términos de regularización en la función de pérdida mejora la generalización del modelo y garantiza que las predicciones sean anatómicamente plausibles [16, 17, 56].

Symmetry Loss

Como se estableció en la Sección 2.1, los 15 *landmarks* considerados en este trabajo incluyen siete pares de puntos con simetría bilateral respecto a la línea media vertical del tórax. Específicamente, los pares simétricos son: bordes costales laterales superiores (#3, #4), bordes costales laterales medios (#5, #6), bordes costales laterales inferiores (#7, #8), ápices pulmonares subclaviculares (#12, #13), y ángulos costofrénicos (#14, #15). Adicionalmente, el ángulo cardiofrénico izquierdo (#2) debe ser aproximadamente simétrico respecto al eje definido por la escotadura yugular (#1) y la carina traqueal (#9).

La función de pérdida de simetría (*Symmetry Loss*) penaliza desviaciones de esta simetría bilateral. Formalmente, sea $S = \{(3, 4), (5, 6), (7, 8), (12, 13), (14, 15)\}$ el conjunto de pares de índices de *landmarks* simétricos, y sea $p_c = (x_c, y_c)^T$ un punto de referencia en la línea media (que puede definirse como el promedio de las coordenadas x de los *landmarks* #1 y #9). La pérdida de simetría se define como:

$$\mathcal{L}_{\text{sym}} = \frac{1}{|S|} \sum_{(i,j) \in S} \|(p_i - p_c) + (p_j - p_c)\|_2^2 \quad (2.5.9)$$

Esta formulación penaliza la suma vectorial $(p_i - p_c) + (p_j - p_c)$, que debería ser aproximadamente $(0, \Delta y)^T$ si los *landmarks* i y j son perfectamente simétricos en la coordenada x respecto a p_c , con posible diferencia en y debido a asimetrías anatómicas menores. Una formulación alternativa más restrictiva penaliza exclusivamente desviaciones en la coordenada x :

$$\mathcal{L}_{\text{sym}}^{(x)} = \frac{1}{|S|} \sum_{(i,j) \in S} [(x_i - x_c) + (x_j - x_c)]^2 \quad (2.5.10)$$

La pérdida de simetría proporciona regularización particularmente útil en presencia de oclusiones parciales o artefactos que afectan asimétricamente la imagen: si un *landmark* en un hemitórax es difícil de detectar debido a oclusión, la restricción de simetría permite que el modelo infiera su posición aproximada basándose en la detección de su contraparte simétrica. Urschler et al. [56] demostraron empíricamente que la incorporación de restricciones geométricas incluyendo simetría bilateral mejora consistentemente la precisión de detección de *landmarks* en imágenes médicas, particularmente en conjuntos de datos pequeños donde la regularización es crítica.

Distance Preservation Loss

Las distancias Euclidianas entre pares específicos de *landmarks* anatómicos exhiben variabilidad inter-paciente limitada en poblaciones normales, proporcionando una restricción geométrica adicional. Por ejemplo, la distancia entre los ápices pulmonares izquierdo y derecho (#12, #13) está relacionada con el ancho torácico superior, que varía dentro de un rango relativamente estrecho. La función de pérdida de preservación de distancias (*Distance Preservation Loss*) penaliza predicciones que violan estas restricciones de distancia.

Formalmente, sea $D \subseteq \{1, \dots, K\} \times \{1, \dots, K\}$ un conjunto de pares de índices de *landmarks* cuyas distancias deben preservarse, y sea d_{ij}^{ref} la distancia de referencia entre *landmarks* i y j , típicamente estimada como la media de las distancias observadas en el conjunto de

entrenamiento. La pérdida de preservación de distancias se define como:

$$\mathcal{L}_{\text{dist}} = \frac{1}{|D|} \sum_{(i,j) \in D} (\|p_i - p_j\|_2 - d_{ij}^{\text{ref}})^2 \quad (2.5.11)$$

Esta formulación penaliza tanto la compresión excesiva (distancia predicha menor que d_{ij}^{ref}) como la expansión excesiva (distancia predicha mayor que d_{ij}^{ref}) de distancias anatómicas características. La selección del conjunto D y las distancias de referencia d_{ij}^{ref} constituye una decisión de diseño que requiere conocimiento anatómico: pares de *landmarks* con alta correlación espacial y baja variabilidad inter-paciente son candidatos ideales. Thaler et al. [17] propusieron métodos de análisis de forma basados en CT que identifican automáticamente restricciones de distancia anatómicamente significativas mediante análisis estadístico de formas.

Una limitación de *Distance Preservation Loss* es que las distancias de referencia d_{ij}^{ref} deben ser específicas de la población y potencialmente específicas de la condición patológica: pacientes con cardiomegalia exhibirán distancias características diferentes a pacientes con anatomía normal. No obstante, para restricciones suficientemente generales (como distancias entre estructuras óseas relativamente rígidas), esta función de pérdida proporciona regularización valiosa.

Función de Pérdida Combinada

En la práctica, las funciones de pérdida de regresión de coordenadas y las restricciones geométricas se combinan mediante suma ponderada:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Wing}} + \lambda_2 \mathcal{L}_{\text{sym}} + \lambda_3 \mathcal{L}_{\text{dist}} \quad (2.5.12)$$

donde $\lambda_1, \lambda_2, \lambda_3 \geq 0$ son hiperparámetros que controlan el balance relativo entre precisión de coordenadas individuales y validez geométrica global. La elección de estos pesos constituye una decisión crítica: valores excesivos de λ_2 y λ_3 pueden forzar simetrías y distancias demasiado rígidas que no capturan la variabilidad anatómica real, mientras que valores demasiado pequeños no proporcionan suficiente regularización. Zeng et al. [57] propusieron estrategias de aprendizaje auto-supervisado que aprenden automáticamente restricciones de consistencia geométrica desde datos no etiquetados, reduciendo la necesidad de especificación manual de restricciones y pesos.

La metodología específica de entrenamiento, incluyendo la selección de valores de hiperparámetros $\lambda_1, \lambda_2, \lambda_3$, las estrategias de ponderación adaptativa durante el entrenamiento, y los protocolos de validación experimental, se presentan en detalle en el Capítulo 3.

2.6. Enfoques de Regresión para Detección de Landmarks

Las funciones de pérdida presentadas en la Sección 2.5 permiten el entrenamiento supervisado de redes neuronales profundas para la tarea de detección de *landmarks* anatómicos. Sin embargo, la arquitectura de salida de la red y la representación de las predicciones constituyen decisiones fundamentales que determinan la eficiencia computacional, la precisión sub-píxel, y la robustez del modelo. Existen dos paradigmas principales para la predicción de localizaciones de *landmarks*: la regresión directa de coordenadas (del inglés, *coordinate regression*), que predice directamente las coordenadas (x, y) de cada punto como valores continuos, y la regresión de mapas de calor (del inglés, *heatmap regression*), que genera mapas de probabilidad espacial bidimensionales que representan la localización de cada *landmark*. Esta sección presenta el análisis matemático y arquitectónico de ambos enfoques, sus ventajas y limitaciones, y proporciona la justificación técnica para la selección del enfoque de regresión directa de coordenadas en el contexto de detección de *landmarks* en radiografías de tórax.

2.6.1. Regresión Directa de Coordenadas

El enfoque de regresión directa de coordenadas formula la detección de *landmarks* como un problema de regresión multi-salida donde la red neuronal aprende un mapeo directo desde la imagen de entrada hasta las coordenadas de todos los *landmarks* [24, 58]. Formalmente, dada una imagen $I \in \mathbb{R}^{H \times W}$ (donde H y W representan altura y ancho en píxeles), y un conjunto de K *landmarks*, el objetivo es aprender una función parametrizada:

$$f_{\theta} : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{2K} \quad (2.6.1)$$

donde θ representa los parámetros de la red, y la salida es un vector de $2K$ valores continuos que representan las coordenadas $\hat{p} = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \dots, \hat{x}_K, \hat{y}_K]^T$.

La arquitectura típica para regresión de coordenadas consta de tres componentes principales: (1) una red neuronal convolucional profunda que actúa como extractor de características (por ejemplo, ResNet-18, como se describió en la Sección 2.3), que transforma la imagen de entrada en mapas de características de alta dimensionalidad; (2) una capa de *global average pooling* (GAP) que reduce cada canal de características espaciales ($h \times w$) a un valor escalar mediante promediación, generando un vector de características de dimensión fija independiente de la resolución espacial de entrada; y (3) una o dos capas completamente conectadas que mapean el vector de características a las $2K$ coordenadas de salida.

Matemáticamente, si $\phi(I; \theta_{\text{conv}}) \in \mathbb{R}^{C \times h \times w}$ representa los mapas de características generados

por la red convolucional (con C canales), el *global average pooling* calcula:

$$z_c = \frac{1}{h \cdot w} \sum_{i=1}^h \sum_{j=1}^w \phi_c(I)_{i,j}, \quad c = 1, \dots, C \quad (2.6.2)$$

generando un vector de características $z \in \mathbb{R}^C$. Las capas completamente conectadas finales realizan la transformación afín:

$$\hat{p} = W_{\text{FC}}z + b_{\text{FC}} \quad (2.6.3)$$

donde $W_{\text{FC}} \in \mathbb{R}^{2K \times C}$ y $b_{\text{FC}} \in \mathbb{R}^{2K}$ son parámetros aprendibles. Durante el entrenamiento, estos parámetros se optimizan mediante minimización de funciones de pérdida basadas en distancias euclidianas, como se discutió en la Sección 2.5.

Ventajas de la regresión directa de coordenadas:

1. Eficiencia de memoria y computacional: La salida de la red es un vector compacto de $2K$ valores, en contraste con representaciones espacialmente extensas. Para $K = 15$ *landmarks*, la salida es un vector de 30 valores escalares. Esta compacidad reduce significativamente los requerimientos de memoria GPU durante el entrenamiento y la inferencia, permitiendo tamaños de lote (*batch size*) más grandes y convergencia más rápida.

2. Precisión sub-píxel inherente: Las coordenadas se predicen como valores continuos en el espacio real \mathbb{R}^2 , proporcionando capacidad intrínseca para localización sub-píxel sin necesidad de técnicas de refinamiento adicionales. Esta propiedad es crítica en aplicaciones médicas donde errores de fracción de píxel pueden tener relevancia diagnóstica.

3. Arquitectura simple y estándar: El enfoque de regresión directa es compatible con arquitecturas de clasificación estándar (como ResNet) mediante el simple reemplazo de la capa completamente conectada final, facilitando la utilización de modelos pre-entrenados en ImageNet mediante *transfer learning*, como se discutió en la Sección 2.4.

4. Reducción de hiperparámetros: A diferencia del enfoque de mapas de calor, no requiere la selección de parámetros relacionados con la representación espacial de las predicciones (como el ancho de las Gaussianas o la resolución de salida).

Limitaciones:

1. Pérdida de información espacial explícita: El *global average pooling* colapsa completamente la estructura espacial de los mapas de características. Esto puede dificultar el aprendizaje de relaciones espaciales complejas entre *landmarks*, aunque este efecto puede mitigarse mediante la incorporación de restricciones geométricas explícitas en la función de pérdida, como las restricciones de simetría y preservación de distancia discutidas en la Sección 2.5.

2. Sensibilidad a oclusiones y ambigüedad: En presencia de oclusiones parciales o artefactos de imagen, la red debe producir una única predicción de coordenada, sin capacidad de representar incertidumbre espacial distribuida.

2.6.2. Regresión de Mapas de Calor

El enfoque de regresión de mapas de calor representa cada *landmark* mediante un mapa de probabilidad espacial bidimensional que indica la probabilidad de presencia del punto en cada localización de la imagen [23, 59]. Para cada *landmark* k , la red genera un mapa de calor $H_k \in \mathbb{R}^{h \times w}$, donde $H_k(i, j) \in [0, 1]$ representa la probabilidad de que el *landmark* k esté localizado en la posición (i, j) .

Durante el entrenamiento, los mapas de calor objetivo (*ground truth*) se construyen típicamente como Gaussianas bidimensionales centradas en las coordenadas anotadas (x_k, y_k) :

$$H_k^{\text{gt}}(i, j) = \exp \left(-\frac{(i - y_k)^2 + (j - x_k)^2}{2\sigma^2} \right) \quad (2.6.4)$$

donde σ controla el ancho de la Gaussiana. La red aprende a predecir estos mapas de calor mediante minimización de funciones de pérdida como el error cuadrático medio píxel-a-píxel o entropía cruzada binaria:

$$\mathcal{L}_{\text{heatmap}} = \frac{1}{K \cdot h \cdot w} \sum_{k=1}^K \sum_{i=1}^h \sum_{j=1}^w \left(H_k(i, j) - H_k^{\text{gt}}(i, j) \right)^2 \quad (2.6.5)$$

Las arquitecturas típicas para regresión de mapas de calor emplean diseños codificador-decodificador que preservan o reconstruyen resoluciones espaciales altas. Las arquitecturas *U-Net* [60] y *Stacked Hourglass Networks* [23] son ejemplos representativos ampliamente utilizados en tareas de estimación de pose humana y análisis de imágenes médicas.

Durante la inferencia, las coordenadas de los *landmarks* se extraen de los mapas de calor predichos mediante dos estrategias principales: (1) *Hard argmax*, que selecciona la posición del píxel con valor máximo $\arg \max_{i,j} H_k(i, j)$, limitando la precisión a la resolución de píxel, o (2) *Soft argmax* diferenciable [5], que calcula el centro de masa ponderado del mapa de calor:

$$\hat{x}_k = \sum_{i=1}^h \sum_{j=1}^w j \cdot \text{softmax}(H_k(i, j)), \quad \hat{y}_k = \sum_{i=1}^h \sum_{j=1}^w i \cdot \text{softmax}(H_k(i, j)) \quad (2.6.6)$$

proporcionando capacidad de localización sub-píxel y diferenciabilidad completa para entrenamiento de extremo a extremo.

Ventajas de la regresión de mapas de calor:

- 1. Información espacial explícita:** Los mapas de calor preservan la estructura espacial bidimensional de las predicciones, facilitando el aprendizaje de relaciones espaciales complejas y contexto anatómico.
- 2. Robustez a oclusiones y ambigüedad:** El enfoque puede representar distribuciones de probabilidad multimodales o difusas, capturando incertidumbre en la localización de *landmarks* parcialmente ocluidos o ambiguos.
- 3. Supervisión densa:** La función de pérdida proporciona señales de gradiente en todas las localizaciones espaciales, potencialmente facilitando la convergencia del entrenamiento.

Limitaciones:

- 1. Costo computacional y de memoria:** La generación de K mapas de calor de resolución $h \times w$ requiere memoria proporcional a $K \cdot h \cdot w$, que puede ser substancialmente mayor que el vector de $2K$ coordenadas. Para $K = 15$ *landmarks* y resolución de salida 64×64 , se requiere almacenar 61,440 valores en comparación con 30 valores del enfoque de coordenadas directas.
- 2. Velocidad de entrenamiento reducida:** Las arquitecturas codificador-decodificador con conexiones de salto y múltiples etapas de deconvolución son computacionalmente más costosas que las arquitecturas estándar con *global pooling*.
- 3. Hiperparámetros adicionales:** El ancho de la Gaussiana σ en la Ecuación 2.6.4 es un hiperparámetro crítico que debe seleccionarse cuidadosamente: valores pequeños proporcionan supervisión más precisa pero pueden dificultar la convergencia, mientras que valores grandes facilitan el aprendizaje pero reducen la precisión de localización.

2.6.3. Comparación y Selección de Enfoque

La Tabla 2.6.1 presenta una comparación sistemática de los aspectos técnicos y prácticos de ambos enfoques.

Investigaciones recientes han explorado enfoques híbridos que combinan ambos paradigmas. Li et al. [61] propusieron una arquitectura multi-tarea que predice simultáneamente mapas de calor y coordenadas, demostrando mejoras en precisión de localización en análisis de radiografías de tórax mediante la combinación de supervisión espacialmente densa y predicción directa. Jeong et al. [62] presentaron un enfoque de regresión de coordenadas guiado por atención en mapas de características espaciales, preservando parcialmente información espacial sin la sobrecarga completa de generación de mapas de calor. Adicionalmente, enfoques basados en *transformers* de visión [63] han demostrado capacidad para capturar relaciones espaciales de

Tabla 2.6.1: Comparación de enfoques de regresión directa de coordenadas y mapas de calor para detección de *landmarks*.

Aspecto	Regresión de Coordenadas	Regresión de Mapas de Calor
Memoria (salida)	$2K$ valores	$K \cdot h \cdot w$ valores
Precisión sub-píxel	Inherente (continuo)	Requiere soft-argmax
Arquitectura	ResNet + GAP + FC	U-Net / Hourglass
Velocidad entrenamiento	Rápida	Moderada-Lenta
Información espacial	Implícita (colapsada)	Explícita (preservada)
Robustez a oclusiones	Limitada	Alta
Hiperparámetros	Mínimos	σ (ancho Gaussiana)
Transfer learning	Directo (ResNet)	Requiere adaptación
Restricciones geométricas	Fácil (en espacio de coord.)	Complejo (en heatmaps)

largo alcance sin la necesidad de representaciones espaciales explícitas, representando una dirección prometedora para futuras investigaciones.

Justificación de la selección de regresión directa de coordenadas:

En el contexto específico de detección de 15 *landmarks* anatómicos en radiografías de tórax con el conjunto de datos del presente trabajo (956 imágenes de entrenamiento, como se establecerá en el Capítulo 3), el enfoque de regresión directa de coordenadas se selecciona por las siguientes razones técnicas:

1. Eficiencia computacional en régimen de datos moderados: Con un conjunto de datos de tamaño moderado, la eficiencia de entrenamiento y la capacidad de utilizar arquitecturas estándar pre-entrenadas en ImageNet (como se discutió en la Sección 2.4) proporcionan ventajas significativas. La regresión directa de coordenadas permite *fine-tuning* directo de ResNet-18 pre-entrenado, aprovechando conocimiento transferido sin modificaciones arquitectónicas substanciales.

2. Precisión sub-píxel natural: La aplicación clínica requiere localización precisa de estructuras anatómicas, y el enfoque de coordenadas continuas proporciona capacidad sub-píxel inherente sin técnicas de refinamiento adicionales.

3. Restricciones de hardware: Las restricciones de memoria GPU (8 GB VRAM en la infraestructura utilizada, como se describirá en el Capítulo 3) favorecen el enfoque de coordenadas compactas, permitiendo tamaños de lote más grandes que aceleran la convergencia y mejoran la estimación de estadísticas de normalización por lotes.

4. Incorporación de conocimiento anatómico mediante restricciones geométricas: La limitación principal del enfoque de coordenadas (pérdida de información espacial explícita) se mitiga mediante la incorporación de las funciones de pérdida de simetría bilateral y preservación de distancias anatómicas presentadas en la Sección 2.5. Estas restricciones geométricas imponen explícitamente conocimiento anatómico sobre la geometría del tórax, compensando

la falta de supervisión espacialmente densa de los mapas de calor.

5. Reducción de hiperparámetros: La eliminación del hiperparámetro σ de ancho de Gaussiana reduce el espacio de búsqueda de hiperparámetros, simplificando el proceso de validación experimental que se presentará en el Capítulo 4.

La combinación de regresión directa de coordenadas con la función de pérdida compuesta $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Wing}} + \lambda_2 \mathcal{L}_{\text{sym}} + \lambda_3 \mathcal{L}_{\text{dist}}$ (Ecuación 2.31, Sección 2.5) constituye el enfoque metodológico adoptado en este trabajo, balanceando eficiencia computacional, precisión de localización, y aprovechamiento de conocimiento anatómico estructurado. El estado del arte de métodos de detección de *landmarks* en imágenes médicas, incluyendo enfoques basados en coordenadas, mapas de calor, y técnicas híbridas, se analiza exhaustivamente en la sección subsecuente.

2.7. Estado del Arte en Detección Automática de Landmarks Anatómicos

La detección automática de *landmarks* anatómicos en imágenes médicas ha experimentado una transformación fundamental en la última década. Los métodos clásicos basados en modelos estadísticos de forma, particularmente Active Shape Models (ASM) [9] y Active Appearance Models (AAM) [10], dominaron el campo durante los años 1990-2000, requiriendo inicialización manual cercana a la solución y siendo sensibles a variaciones de iluminación y pose. La revolución del aprendizaje profundo, iniciada con AlexNet [12], transformó radicalmente el panorama: Sun et al. [58] demostraron en 2013 que redes neuronales convolucionales profundas superaban métodos clásicos en detección de *landmarks* faciales mediante una cascada de redes con refinamiento progresivo. La introducción de arquitecturas residuales profundas [29] y técnicas avanzadas de regresión de mapas de calor [23, 59] consolidaron el aprendizaje profundo como el paradigma dominante. En el dominio de imágenes médicas, tres enfoques metodológicos principales han emergido: (1) regresión directa de coordenadas, que predice localizaciones (x, y) como valores continuos con eficiencia computacional superior; (2) regresión de mapas de calor, que genera distribuciones de probabilidad espacial con capacidad de representar incertidumbre; y (3) métodos híbridos y basados en transformadores, que combinan fortalezas de múltiples paradigmas o explotan mecanismos de atención global. Simultáneamente, la disponibilidad de conjuntos de datos ha evolucionado desde colecciones pequeñas de cientos de imágenes hacia repositorios masivos como CheXpert (224,316 radiografías de tórax) y MIMIC-CXR (377,110 estudios), facilitando el entrenamiento de modelos cada vez más robustos y generalizables. Esta sección presenta una revisión exhaustiva del estado del arte, organizada según el enfoque metodológico principal, seguida de un análisis comparativo de los trabajos más relevantes publicados entre 2016 y 2024.

2.7.1. Métodos Basados en Regresión de Coordenadas

El enfoque de regresión directa de coordenadas formula la detección de *landmarks* como un problema de regresión multi-salida donde redes neuronales convolucionales profundas aprenden mapeos directos desde imágenes hacia vectores de coordenadas. Sun et al. [58] propusieron una arquitectura pionera denominada Deep Convolutional Network Cascade para *landmarks* faciales, consistente en tres niveles de refinamiento progresivo: una red inicial predice localizaciones aproximadas en la imagen completa, seguida de redes subsecuentes que refinan las predicciones en regiones locales de tamaño decreciente. Esta estrategia de *coarse-to-fine* alcanzó un error normalizado de 5.5 % en el conjunto de datos LFPW, superando

métodos clásicos basados en AAM por márgenes significativos. Zhang et al. [24] extendieron este enfoque mediante aprendizaje multi-tarea, demostrando que la predicción simultánea de *landmarks* faciales y atributos semánticos (presencia de gafas, género, expresión) mejora la precisión de localización: las tareas auxiliares actúan como regularizadores que fuerzan a la red a aprender representaciones más generalizables. Bulat y Tzimiropoulos [64] establecieron *benchmarks* de referencia mediante la construcción de un conjunto de datos masivo de 230,000 *landmarks* faciales 3D, alcanzando un error promedio de 3.12 píxeles en el conjunto 300-W, y demostraron la importancia de la escala de datos de entrenamiento para la robustez de los modelos. Una observación consistente en la literatura es que la regresión de coordenadas ha sido preferida en aplicaciones de imágenes médicas debido a su eficiencia computacional y capacidad de predicción sub-píxel inherente.

En el dominio específico de imágenes médicas, Noothout et al. [53] presentaron en *IEEE Transactions on Medical Imaging* un enfoque de localización global-a-local utilizando redes neuronales completamente convolucionales (FCNNs) para la detección de 19 *landmarks* cefalométricos en radiografías laterales de cráneo. Su método opera en dos etapas: una red de localización gruesa identifica regiones de interés que contienen cada *landmark*, seguida de una red de refinamiento que predice coordenadas precisas dentro de parches locales. Evaluado en un conjunto de datos de 400 radiografías, el método alcanzó un error de $1,21 \pm 0,89$ mm, demostrando además transferibilidad entre modalidades de imagen (angiografía por tomografía computarizada, resonancia magnética, y radiografía). Oh et al. [65] propusieron en *IEEE Journal of Biomedical and Health Informatics* un enfoque de aprendizaje de características de contexto anatómico profundo para *landmarks* cefalométricos, incorporando mecanismos de atención guiados por contexto que permiten a la red enfocarse en regiones anatómicas relevantes. Su arquitectura basada en DenseNet alcanzó un error de 1.18 mm en un conjunto de datos de 935 radiografías laterales, representando el estado del arte en detección cefalométrica mediante regresión de coordenadas al momento de publicación. Li et al. [61] presentaron en *Scientific Reports* un enfoque híbrido que combina regresión de coordenadas y mapas de calor para 46 *landmarks* en radiografías de tórax posteroanterior, utilizando un conjunto de datos de 956 imágenes. Su método incorpora restricciones de simetría bilateral en la función de pérdida, explotando la propiedad anatómica de simetría del tórax, y alcanzó un error promedio de 4.22 píxeles. Este trabajo representa la aplicación más exhaustiva de detección de *landmarks* en radiografías de tórax en términos de número de puntos anatómicos, demostrando que la combinación de restricciones geométricas con regresión eficiente es viable en conjuntos de datos de tamaño moderado.

2.7.2. Métodos Basados en Mapas de Calor

La regresión de mapas de calor representa cada *landmark* mediante una distribución de probabilidad espacial bidimensional, típicamente una Gaussiana centrada en la localización objetivo. Tompson et al. [59] fueron pioneros en la aplicación de este enfoque para estimación de pose humana, combinando redes neuronales convolucionales con modelos gráficos que capturan dependencias espaciales entre articulaciones. Su método genera mapas de calor Gaussianos con desviación estándar $\sigma = 1$ píxel para cada articulación, optimizados mediante error cuadrático medio píxel-a-píxel. Newell et al. [23] revolucionaron el campo con la introducción de *Stacked Hourglass Networks*, una arquitectura multi-escala iterativa que procesa características en múltiples resoluciones mediante módulos codificador-decodificador apilados secuencialmente. Cada módulo *hourglass* realiza *downsampling* progresivo para capturar contexto global, seguido de *upsampling* simétrico con conexiones de salto que preservan detalles espaciales. La composición de múltiples módulos (típicamente 4-8) permite refinamiento iterativo de predicciones, alcanzando $PCKh@0.5 = 90.9\%$ en el conjunto de datos MPII Human Pose, estableciendo un nuevo estado del arte que persiste como referencia fundamental. Yang et al. [66] demostraron la transferibilidad de la arquitectura *Stacked Hourglass* desde estimación de pose humana hacia detección de *landmarks* faciales, evidenciando que las representaciones jerárquicas multi-escala son genéricas a través de dominios anatómicos.

En aplicaciones de imágenes médicas, Payer et al. [67] presentaron en *Medical Image Analysis* una extensión de regresión de mapas de calor que integra redes de configuración espacial (*spatial configuration networks*) para *landmarks* en radiografías de mano. Su método incorpora un mecanismo de *soft-argmax* diferenciable que permite extracción de coordenadas sub-píxel a partir de mapas de calor mientras mantiene diferenciability completa para entrenamiento de extremo a extremo: $\hat{x}_k = \sum_{i,j} j \cdot \text{softmax}(H_k(i,j))$, $\hat{y}_k = \sum_{i,j} i \cdot \text{softmax}(H_k(i,j))$. La arquitectura basada en U-Net alcanzó un error de $1,87 \pm 0,98$ mm en un conjunto de datos de 895 imágenes, demostrando que la preservación de información espacial explícita proporciona ventajas en presencia de estructuras anatómicas complejas con múltiples *landmarks* densamente distribuidos. Zhang et al. [68] propusieron en *Medical Image Analysis* redes neuronales convolucionales en cascada con arquitectura U-Net para 19 *landmarks* cefalométricos, implementando una estrategia de tres etapas de refinamiento progresivo de *coarse-to-fine*. Su método alcanzó un error de $1,35 \pm 0,89$ mm en un conjunto de datos de 1,000 radiografías laterales, superando métodos clásicos basados en ASM por 42 % y demostrando la superioridad de enfoques basados en aprendizaje profundo sobre técnicas estadísticas tradicionales. Cheng et al. [54] presentaron en *Medical Image Analysis* un enfoque de aprendizaje basado en perturbaciones con regresión de mapas de calor para 18 *landmarks* en radiografías de tórax posteroanterior. Su método incorpora aumentación de datos geométrica avanzada mediante perturbaciones controladas durante el entrenamiento, alcanzando un error

de 3.78 píxeles en un conjunto de datos de 2,000 imágenes. Thaler et al. [17] introdujeron modelado de incertidumbre mediante mapas de calor Gaussianos con enfoque Bayesiano para radiografías de mano, cuantificando explícitamente la incertidumbre de localización para cada *landmark* y alcanzando un error de 2.12 mm. Una limitación consistente de métodos basados en mapas de calor es el sobre costo de memoria proporcional a $K \times h \times w$ (número de *landmarks* \times resolución espacial), significativamente superior al vector compacto de $2K$ valores de regresión de coordenadas, con impacto directo en el tamaño de lote durante el entrenamiento y velocidad de inferencia.

2.7.3. Métodos Híbridos y Basados en Transformers

La tercera categoría de enfoques combina elementos de regresión de coordenadas y mapas de calor, o introduce arquitecturas basadas en mecanismos de atención que superan limitaciones de receptive fields finitos en redes convolucionales. Quan et al. [69] presentaron en MICCAI 2021 el enfoque “You Only Learn Once” (YOLO), un detector universal de *landmarks* anatómicos entrenado simultáneamente en conjuntos de datos mixtos (cefalométricos, mano, columna vertebral) conteniendo más de 150 *landmarks* diferentes distribuidos en múltiples anatomías. El método demuestra capacidad de generalización cruzada entre anatomías, alcanzando errores variables de 1.5-3.2 mm según la región anatómica específica, y evidencia que el entrenamiento multi-dominio mejora la robustez mediante exposición a diversidad anatómica. Ma y Luo [70] propusieron en *IEEE Journal of Biomedical and Health Informatics* una función de pérdida adaptativa de grano fino que ajusta dinámicamente los pesos de cada *landmark* según su dificultad de detección empírica durante el entrenamiento. Su función de pérdida se formula como $\mathcal{L} = \sum_{k=1}^K \alpha_k \cdot \text{MSE}_k$, donde los pesos α_k se aprenden mediante un módulo de atención que evalúa la magnitud de gradientes históricos. Aplicado a 19 *landmarks* cefalométricos, este enfoque proporciona mejoras de 8.3 % sobre MSE estándar, demostrando que la adaptación dinámica de la función de pérdida es una dirección prometedora. Kang et al. [71] presentaron en *Scientific Reports* detección de *landmarks* cefalométricos 3D mediante aprendizaje por refuerzo profundo multi-etapa en imágenes CBCT (tomografía computarizada de haz cónico), formulando la localización como un proceso de decisión de Markov donde un agente aprende secuencias de acciones que minimizan la distancia al *landmark* objetivo. Su enfoque alcanzó un error de 1.82 ± 1.03 mm en espacio tridimensional con un conjunto de datos de 350 exploraciones volumétricas.

La introducción de arquitecturas *Vision Transformer* ha generado interés significativo en años recientes. Li et al. [63] propusieron en CVPR 2022 una arquitectura de transformadores en cascada para *landmarks* faciales, donde mecanismos de *self-attention* global capturan relaciones espaciales de largo alcance entre *landmarks* sin las limitaciones de campos receptivos finitos inherentes a convoluciones. Su método alcanzó NME (error medio normalizado)

de 2.98 % en el conjunto de datos WFLW con 98 *landmarks* faciales, demostrando competitividad con enfoques convolucionales mientras proporciona interpretabilidad superior mediante visualización de mapas de atención. Huang et al. [72] presentaron en MICCAI 2023 un modelo híbrido Transformer-CNN (HTC) con aprendizaje de mapas de calor multi-resolución, combinando extracción de características locales mediante bloques convolucionales con modelado de contexto global mediante bloques de transformador. Su arquitectura superó una línea base ResNet-50 por 11.2 % en *landmarks* de radiografías de tórax, evidenciando que la hibridación de arquitecturas convolucionales y basadas en atención explota complementariedad: convoluciones capturan patrones locales eficientemente mediante inductiva bias de localidad, mientras transformadores modelan dependencias globales sin restricciones espaciales. Jeong et al. [62] presentaron en *Sensors* regresión de coordenadas guiada por atención con características de mapas de calor intermedias para *landmarks* faciales, alcanzando un error de 5.13 píxeles en radiografías de tórax mediante un mecanismo híbrido que genera mapas de calor en capas intermedias para guiar la regresión final de coordenadas.

Gaggion et al. [73] introdujeron en *IEEE Transactions on Medical Imaging* HybridGNet, una arquitectura basada en redes neuronales de grafo (Graph Neural Networks, GNN) que incorpora conocimiento anatómico previo en segmentación de radiografías de tórax mediante representación de *landmarks* como nodos de un grafo donde las aristas representan relaciones anatómicas (adyacencia espacial, simetría bilateral, jerarquía anatómica). Su método mejora la plausibilidad anatómica de segmentaciones en 15.6 % comparado con U-Net estándar, demostrando que la codificación explícita de estructura anatómica mediante grafos constituye una dirección prometedora. Liu et al. [74] propusieron en CVPR 2021 detección de *landmarks* con conciencia de estructura mediante GCN para capturar relaciones espaciales explícitas entre *landmarks* faciales, modelando las dependencias geométricas como un grafo totalmente conectado donde cada *landmark* se conecta con todos los demás, permitiendo propagación de información contextual. La tendencia emergente hacia integración de conocimiento previo geométrico mediante representaciones de grafo representa una convergencia entre aprendizaje profundo basado en datos y modelado estructurado tradicional.

2.7.4. Funciones de Pérdida Especializadas y Restricciones Geométricas

La función de pérdida constituye un componente crítico que determina qué propiedades de las predicciones son optimizadas durante el entrenamiento. Feng et al. [15] introdujeron en CVPR 2018 la función *Wing Loss* para localización robusta de *landmarks* faciales, diseñada para amplificar gradientes en el régimen de errores pequeños mediante una curva basada en logaritmo: $\mathcal{L}_{\text{wing}}(x) = w \ln(1 + |x|/\epsilon)$ para $|x| < w$, donde w controla el ancho de la región

no lineal y ϵ limita el gradiente en $x = 0$. Con parámetros típicos $w = 10$, $\epsilon = 2$, *Wing Loss* proporciona mejoras de 12.5 % sobre MSE en *landmarks* faciales con errores menores a 2 píxeles, alcanzando NME 4.04 % en el conjunto 300-W. La intuición fundamental es que MSE cuadrático genera gradientes que decrecen linealmente con el error, proporcionando señal de optimización débil en el régimen de alta precisión, mientras *Wing Loss* mantiene gradientes substanciales incluso para errores muy pequeños, acelerando la convergencia hacia localizaciones precisas. Wang et al. [75] extendieron este concepto con *Adaptive Wing Loss* que ajusta dinámicamente los parámetros w y ϵ durante el entrenamiento, aplicado a regresión de mapas de calor, proporcionando mejoras adicionales de 5.3 % sobre *Wing Loss* estándar. Ma y Luo [70] generalizaron la adaptación a nivel de *landmark* individual mediante pesos específicos basados en dificultad empírica, donde la función de pérdida adaptativa asigna mayor énfasis a *landmarks* con historial de errores elevados. El impacto de *Wing Loss* ha sido substancial: la función ha sido adoptada ampliamente en aplicaciones de imágenes médicas debido a su robustez a *outliers* y enfoque en errores pequeños clínicamente relevantes.

Las restricciones geométricas basadas en conocimiento anatómico representan una segunda categoría de especialización de funciones de pérdida. Song et al. [76] incorporaron restricciones de simetría para *landmarks* cefalométricos mediante penalización de asimetrías bilaterales: $\mathcal{L}_{\text{sym}} = \sum_{(i,j) \in S} \|p_i - \text{mirror}(p_j)\|_2^2$, donde S denota pares de *landmarks* simétricos y $\text{mirror}(\cdot)$ representa reflexión respecto al plano sagital medio. Esta restricción proporciona mejoras de 6.8 % específicamente en *landmarks* con simetría bilateral, explotando la propiedad anatómica fundamental de que estructuras bilaterales deben ser aproximadamente simétricas en individuos sanos. Thaler et al. [17] incorporaron preservación de distancias mediante restricciones basadas en modelos estadísticos de forma, penalizando desviaciones de distancias inter-*landmark* respecto a valores de referencia anatómicos: $\mathcal{L}_{\text{dist}} = \sum_{(i,j) \in D} (\|p_i - p_j\|_2 - d_{ij}^{\text{ref}})^2$, donde D denota pares de *landmarks* con distancia anatómica conocida y d_{ij}^{ref} representa la distancia de referencia. Urschler et al. [56] presentaron en *Pattern Recognition Letters* integración de restricciones geométricas mediante conocimiento previo de forma aprendido de datos de entrenamiento, utilizando análisis de componentes principales sobre configuraciones de *landmarks* para definir un espacio de formas anatómicamente plausibles.

Kendall y Gal [77] introdujeron en NeurIPS 2017 un marco para cuantificación de incertidumbre en aprendizaje profundo Bayesiano para visión por computadora, distinguiendo entre incertidumbre aleatoria (inherente a los datos) y epistémica (incertidumbre del modelo). Liu et al. [78] extendieron este enfoque a detección de *landmarks* anatómicos en imágenes médicas, presentando en *IEEE Transactions on Medical Imaging* 2024 un método de aprendizaje profundo con conciencia de incertidumbre que predice intervalos de confianza para cada *landmark*, permitiendo detección automática de predicciones con alta incertidumbre que requieren revisión manual. Su método alcanzó errores de 1.5-2.8 píxeles en múltiples dominios anatómicos mientras proporciona calibración de incertidumbre superior. Un gap

crítico identificado en la literatura es que pocos trabajos combinan múltiples componentes de función de pérdida simultáneamente: la mayoría de métodos utiliza MSE estándar o una única restricción geométrica, sin exploración sistemática de combinaciones de *Wing Loss*, restricciones de simetría, y preservación de distancias.

2.7.5. Análisis Comparativo y Posicionamiento del Presente Trabajo

La Tabla 2.7.1 presenta una comparación exhaustiva de trabajos representativos en detección de *landmarks* anatómicos en imágenes médicas publicados entre 2016 y 2024. Los criterios de inclusión fueron: (1) aplicación a imágenes médicas (radiografías, tomografía computarizada, resonancia magnética), (2) utilización de aprendizaje profundo, (3) evaluación cuantitativa reportada con métricas de error de localización, y (4) publicación en *venues* de alto impacto académico (IEEE Transactions on Medical Imaging, IEEE Journal of Biomedical and Health Informatics, Medical Image Analysis, CVPR/ICCV, MICCAI, Scientific Reports, Applied Sciences, Sensors). Los criterios de comparación incluyen: método/enfoque (coordinate regression, heatmap regression, o híbrido), arquitectura de red neuronal, tipo y tamaño del conjunto de datos, función de pérdida utilizada, error promedio de localización reportado, y dominio anatómico específico. La tabla evidencia la diversidad de enfoques metodológicos y la evolución temporal hacia arquitecturas más sofisticadas y funciones de pérdida especializadas.

El análisis de la Tabla 2.7.1 revela múltiples tendencias significativas en la evolución del campo. Primero, se observa una transición temporal inequívoca desde métodos clásicos hacia aprendizaje profundo: Lindner et al. [79] reportaron en 2016 un error de 2.0 mm utilizando enfoques multi-atlas tradicionales, mientras que Zhang et al. [68] alcanzaron 1.35 mm con redes neuronales convolucionales en cascada en 2020, representando una mejora del 32.5 %. Segundo, existe una divergencia metodológica según el dominio de aplicación: la estimación de pose humana y detección de *landmarks* faciales favorecen regresión de mapas de calor (Newell 2016, Yang 2017) debido a la preservación de contexto espacial y capacidad de representar incertidumbre multimodal, mientras que aplicaciones en imágenes médicas prefieren predominantemente regresión de coordenadas (Noothout 2020, Oh 2020, Li 2023) por eficiencia computacional y menor consumo de memoria GPU. Tercero, el aprendizaje por transferencia desde ImageNet se ha establecido como práctica universal: todos los trabajos publicados después de 2018 utilizan pre-entrenamiento en ImageNet, sin reportes de entrenamiento desde inicialización aleatoria, evidenciando la importancia crítica de representaciones pre-aprendidas discutida en la Sección 2.4. Cuarto, existe una disparidad substancial en tamaños de conjuntos de datos entre dominio médico y visión por computadora general: conjuntos médicos típicamente contienen 400-2000 imágenes (Noothout 400, Zhang 1000, Li 956) versus conjuntos faciales/pose con más de 3000 ejemplos (Li 2022 con 5000,

Liu 2021 con 3000), representando un factor de 3-10× de diferencia que refleja las dificultades de adquisición y anotación de datos médicos. Quinto, se observa una evolución en funciones de pérdida desde MSE estándar (dominante 2016-2019) hacia *Wing Loss* (2018+) y funciones multi-componente adaptativas (2021+), reflejando comprensión creciente de que la función de pérdida debe alinearse con los requisitos específicos de la aplicación. Finalmente, la emergencia de arquitecturas basadas en transformadores post-2022 (Li 2022, Huang 2023) indica un cambio paradigmático hacia captura de contexto global mediante mecanismos de atención, aunque las redes neuronales convolucionales mantienen predominancia en imágenes médicas debido a eficiencia computacional y menores requerimientos de datos.

El análisis comparativo identifica múltiples gaps y limitaciones en el estado del arte actual que motivan la presente investigación. Primero, la combinación de restricciones geométricas es limitada: la mayoría de trabajos incorpora una única restricción (Song 2020 utiliza exclusivamente simetría, Payer 2019 solo configuración espacial, Thaler 2021 únicamente modelado de incertidumbre). Li et al. [61] combinan regresión híbrida coordinate+heatmap con restricciones de simetría, pero utilizan MSE estándar en lugar de *Wing Loss*, perdiendo los beneficios de amplificación de gradientes en errores pequeños. Ningún trabajo reportado en la literatura combina simultáneamente *Wing Loss*, restricciones de simetría bilateral, y preservación de distancias anatómicas para radiografías de tórax, representando un gap crítico. Segundo, los estudios de ablación cuantitativos son limitados: solo 11 de 20 trabajos revisados reportan ablaciones sistemáticas que cuantifican el impacto individual de cada componente de sus funciones de pérdida compuestas, dificultando la comprensión de qué restricciones geométricas proporcionan las mayores contribuciones al desempeño. Tercero, existe un trade-off no explorado sistemáticamente entre número de *landmarks* y tamaño de conjunto de datos: Li 2023 utiliza 46 *landmarks* (el más exhaustivo para radiografías de tórax) pero con 956 imágenes, mientras Cheng 2023 utiliza 2000 imágenes pero solo 18 *landmarks*, sin análisis de cómo esta relación afecta la capacidad de generalización. Cuarto, las restricciones de simetría bilateral son substancialmente infrautilizadas: solo Song 2020 y Li 2023 explotan explícitamente la simetría bilateral en radiografías cefalométricas y de tórax, a pesar de ser una propiedad anatómica fundamental que podría proporcionar supervisión adicional sin requerir anotaciones adicionales. Finalmente, la exploración de restricciones de distancias anatómicas es limitada: Thaler 2021 y Payer 2019 utilizan restricciones de distancia en radiografías de mano, pero este enfoque no ha sido aplicado sistemáticamente a radiografías de tórax con sus 7 pares simétricos específicos y relaciones de distancia anatómicamente consistentes entre estructuras mediastinales y costales (como se estableció en la Sección 2.1).

El presente trabajo se posiciona en la intersección de tres líneas de investigación complementarias: (1) regresión eficiente de coordenadas para radiografías de tórax siguiendo los enfoques exitosos de Li et al. [61] y Jeong et al. [62], (2) *Wing Loss* para amplificación de gradientes en el régimen de errores pequeños clínicamente relevantes como demostrado

por Feng et al. [15] en *landmarks* faciales, y (3) restricciones geométricas anatómicas que explotan propiedades estructurales del tórax (Song et al. [76] para simetría, Payer et al. [67] para preservación de distancias). La contribución única del trabajo es la primera aplicación de la combinación *Wing Loss* + restricciones de simetría bilateral + preservación de distancias anatómicas para 15 *landmarks* en radiografías de tórax posteroanterior utilizando una arquitectura ResNet-18 eficiente pre-entrenada en ImageNet. La función de pérdida compuesta propuesta $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Wing}} + \lambda_2 \mathcal{L}_{\text{sym}} + \lambda_3 \mathcal{L}_{\text{dist}}$ (presentada en detalle en la Sección 2.5) integra conocimiento anatómico específico del tórax sin incrementar la complejidad arquitectónica: las 7 pares de *landmarks* simétricos bilaterales y 2 puntos localizados en la línea media (identificados en la Tabla 2.1.1 de la Sección 2.1) definen naturalmente las restricciones geométricas, y las distancias anatómicas entre pares específicos de *landmarks* (por ejemplo, entre ápices pulmonares, entre ángulos costofénicos) proporcionan supervisión adicional basada en variabilidad anatómica limitada. La metodología experimental completa, incluyendo estudios de ablación sistemáticos para cuantificar el impacto de cada componente de la función de pérdida, se presenta en el Capítulo 3, mientras que los resultados comparativos con el estado del arte se discuten en el Capítulo 4.

La revisión exhaustiva del estado del arte presentada en esta sección evidencia la madurez del campo de detección automática de *landmarks* mediante aprendizaje profundo, con errores de localización alcanzando precisión sub-milimétrica en aplicaciones cefalométricas controladas (Oh 2020: 1.18 mm, Ma & Luo 2021: 1.29 mm) y 3-5 píxeles en radiografías de tórax con mayor variabilidad inter-paciente e inter-institucional (Li 2023: 4.22 píxeles, Cheng 2023: 3.78 píxeles). Sin embargo, los gaps identificados en combinación de funciones de pérdida especializadas y restricciones geométricas anatómicas específicas motivan la investigación de enfoques que integren múltiples fuentes de conocimiento anatómico simultáneamente. La siguiente sección presenta una síntesis del marco teórico completo desarrollado en el Capítulo 2, conectando los fundamentos de aprendizaje profundo, arquitecturas residuales, aprendizaje por transferencia, funciones de pérdida especializadas, y estado del arte revisado con la metodología experimental que será presentada en el Capítulo 3.

Tabla 2.7.1: Estado del arte en detección de *landmarks* en imágenes médicas (2016-2024). Las abreviaciones utilizadas son: Ceph (cefalométrico), px (píxeles), NME (error medio normalizado), GCN (Graph Convolutional Network), ViT (Vision Transformer), RL (Reinforcement Learning).

Autor/Año	Método	Arquitectura	Dataset (tipo/n)	Loss Function	Error	Dominio
Lindner 2016	Multi-atlas clásico	N/A (no DL)	Ceph 400	N/A	2.0 mm	Cefalométrico
Yang 2017	Heatmap	Stacked Hourglass	Facial 3000	MSE	3.4 px	Facial
Feng 2018	Coordinate	ResNet-50	Facial 3800	Wing Loss	4.04 % NME	Facial
Wang 2019	Heatmap	Hourglass	Facial 4000	Adaptive Wing	3.81 % NME	Facial
Payer 2019	Heatmap+GCN	U-Net+Spatial	Hand X-ray 895	MSE+spatial	1.87 mm	Mano
Noothout 2020	Coordinate	FCNN	Ceph 400	MSE	1.21 mm	Cefalométrico
Oh 2020	Coordinate	DenseNet	Ceph 935	MSE+context	1.18 mm	Cefalométrico
Song 2020	Coordinate	ResNet-18	Ceph 450	MSE+symmetry	1.45 mm	Cefalométrico
Zhang 2020	Heatmap cascade	U-Net (3-stage)	Ceph 1000	MSE	1.35 mm	Cefalométrico
Kang 2021	3D RL-based	3D CNN	CBCT 350	Reward-based	1.82 mm	Ceph 3D
Ma & Luo 2021	Heatmap	U-Net	Ceph 800	Adaptive Loss	1.29 mm	Cefalométrico
Thaler 2021	Heatmap Bayesian	U-Net	Hand 600	MSE+uncertainty	2.12 mm	Mano
Quan 2021	Universal	ResNet-101	Mixed 2000+	MSE multi-task	1.5-3.2 mm	Multi-anatomía
Liu 2021	Coordinate+GCN	ResNet+GCN	Facial 3000	Wing+structure	3.2 px	Facial
Li 2022	Transformer	ViT cascade	Facial 5000	Wing Loss	2.98 % NME	Facial
Cheng 2023	Heatmap+perturbationNet	U-Net	Chest X-ray 2000	MSE	3.78 px	Tórax
Gaggion 2023	GNN hybrid	HybridGNet	Chest X-ray 1500	MSE+graph	N/A (seg)	Tórax
Huang 2023	Hybrid Trans-CNN	ViT+ResNet	X-ray multi 1200	MSE heatmap	2.8 px	Multi X-ray
Li 2023	Hybrid coord+heat	ResNet-34	Chest 956 (46 lmks)	MSE+symmetry	4.22 px	Tórax
Jeong 2023	Attention-guided	ResNet-50+attn	Chest 800	Wing Loss	5.13 px	Tórax
Liu 2024	Uncertainty-aware	ResNet+Bayesian	Multi 3500	MSE+uncertainty	1.5-2.8 px	Multi-dominio

2.8. Síntesis del Marco Teórico

El presente capítulo ha desarrollado un marco teórico comprehensivo que fundamenta la detección automática de *landmarks* anatómicos en radiografías de tórax mediante aprendizaje profundo. La estructura del capítulo integra progresivamente: (1) los principios físicos de formación de imágenes radiográficas y la definición anatómica de 15 *landmarks* específicos con propiedades de simetría bilateral (Sección 2.1), (2) los fundamentos matemáticos de redes neuronales convolucionales incluyendo la operación de convolución, retropropagación de gradientes, y algoritmos de optimización (Sección 2.2), (3) las arquitecturas residuales profundas que permiten el entrenamiento efectivo de redes con decenas de capas mediante conexiones de atajo y normalización por lotes (Sección 2.3), (4) el paradigma de aprendizaje por transferencia que aprovecha representaciones pre-aprendidas en ImageNet para mejorar el desempeño en el dominio médico con datos limitados (Sección 2.4), (5) las funciones de pérdida especializadas que incorporan amplificación de gradientes para errores pequeños y restricciones geométricas anatómicas (Sección 2.5), (6) el análisis comparativo de enfoques de regresión de coordenadas versus mapas de calor con justificación técnica para la selección del primero (Sección 2.6), y (7) la revisión exhaustiva del estado del arte con tabla comparativa de 21 trabajos publicados entre 2016 y 2024, identificando gaps específicos en la combinación de funciones de pérdida y restricciones geométricas (Sección 2.7).

La integración conceptual de estos elementos constituye la base metodológica del presente trabajo. Las arquitecturas ResNet-18 presentadas en la Sección 2.3, con 11.7 millones de parámetros distribuidos en bloques residuales básicos con conexiones de atajo, proporcionan un balance óptimo entre capacidad expresiva y eficiencia computacional apropiado para conjuntos de datos médicos de tamaño moderado. El aprendizaje por transferencia (Sección 2.4) permite inicializar estos modelos con pesos pre-entrenados en ImageNet, aprovechando características de bajo y medio nivel (bordes, texturas, estructuras geométricas) que son transferibles al dominio de radiografías de tórax a pesar de la brecha substancial entre imágenes naturales RGB e imágenes médicas de canal único. El ajuste fino (*fine-tuning*) con tasas de aprendizaje diferenciales adapta las capas profundas al dominio médico mientras preserva las representaciones genéricas en capas tempranas. La arquitectura de salida mediante regresión directa de coordenadas (Sección 2.6) predice un vector compacto de 30 valores continuos (2×15 *landmarks*) mediante capas completamente conectadas aplicadas sobre *global average pooling*, proporcionando eficiencia de memoria, precisión sub-píxel inherente, y compatibilidad directa con arquitecturas ResNet pre-entrenadas. La función de pérdida compuesta propuesta en la Sección 2.5 integra tres componentes complementarios: *Wing Loss* para amplificación de gradientes en el régimen de errores pequeños clínicamente relevantes ($|x| < 10$ píxeles), restricciones de simetría bilateral que penalizan desviaciones entre los 7 pares de *landmarks* simétricos identificados en la Tabla 2.1.1, y preservación de distancias anatómicas

que regulariza las distancias inter-*landmark* hacia valores de referencia consistentes con la anatomía torácica normal. Esta combinación explota directamente el conocimiento anatómico específico establecido en la Sección 2.1 sin incrementar la complejidad arquitectónica, proporcionando supervisión adicional que guía el aprendizaje hacia configuraciones de *landmarks* anatómicamente plausibles.

El análisis del estado del arte presentado en la Sección 2.7 evidencia que, si bien métodos basados en aprendizaje profundo han alcanzado precisión sub-milimétrica en aplicaciones cefalométricas controladas (Oh 2020: 1.18 mm, Ma & Luo 2021: 1.29 mm) y 3-5 píxeles en radiografías de tórax con mayor variabilidad (Li 2023: 4.22 píxeles, Cheng 2023: 3.78 píxeles), existen gaps significativos en la literatura: ningún trabajo reportado combina simultáneamente *Wing Loss*, restricciones de simetría bilateral, y preservación de distancias anatómicas para detección de *landmarks* en radiografías de tórax. La mayoría de métodos utiliza MSE estándar o una única restricción geométrica (Song 2020: solo simetría, Payer 2019: solo configuración espacial), y los estudios de ablación cuantitativos que descomponen las contribuciones individuales de cada componente de función de pérdida son limitados (solo 11 de 21 trabajos revisados). Adicionalmente, las restricciones de simetría bilateral y preservación de distancias anatómicas son infrautilizadas en radiografías de tórax a pesar de ser propiedades anatómicas fundamentales que pueden explotarse sin requerir anotaciones adicionales. Estos gaps motivan la investigación presentada en este trabajo: la integración de conocimiento anatómico específico del tórax mediante una función de pérdida multi-componente con una arquitectura eficiente de regresión de coordenadas.

Las contribuciones del marco teórico desarrollado en este capítulo son múltiples. Primero, se ha proporcionado una fundamentación matemática rigurosa de cada componente metodológico, incluyendo derivaciones completas de la retropropagación de gradientes (Ecuaciones 2.8-2.12), el optimizador Adam (Ecuaciones 2.16-2.18), los bloques residuales (Ecuaciones 2.19-2.22), la función *Wing Loss* con análisis de gradientes (Ecuaciones 2.24-2.26), las restricciones de simetría y preservación de distancias (Ecuaciones 2.27-2.30), y las formulaciones de regresión de coordenadas versus mapas de calor (Ecuaciones 2.32-2.38). Esta fundamentación establece precisamente qué propiedades matemáticas de cada componente son relevantes para la tarea de detección de *landmarks* y cómo interactúan durante el proceso de optimización. Segundo, se ha presentado un análisis exhaustivo del estado del arte mediante una tabla comparativa de 21 trabajos representativos publicados en *venues* de alto impacto (IEEE Transactions on Medical Imaging, Medical Image Analysis, CVPR/ICCV, MICCAI) entre 2016 y 2024, categorizados según enfoque metodológico, con identificación explícita de tendencias temporales, divergencias metodológicas entre dominios, y gaps específicos en la literatura. Tercero, se ha justificado técnicamente cada decisión de diseño mediante análisis de ventajas, limitaciones, y trade-offs: la selección de ResNet-18 sobre arquitecturas más profundas se justifica por eficiencia de parámetros y menor propensión al sobreajuste en

conjuntos de datos moderados; la preferencia por regresión de coordenadas sobre mapas de calor se fundamenta en eficiencia computacional, precisión sub-píxel inherente, y restricciones de hardware; la combinación específica de componentes de función de pérdida se motiva por los gaps identificados en el estado del arte y las propiedades anatómicas específicas del tórax.

El marco teórico establecido proporciona todos los fundamentos conceptuales, matemáticos y contextuales necesarios para proceder a la descripción de la metodología experimental. El Capítulo 3 presenta la implementación concreta de los conceptos teóricos desarrollados en este capítulo: la descripción del conjunto de datos específico utilizado incluyendo procedimientos de adquisición, anotación, y preprocesamiento; la arquitectura de red neuronal implementada con detalles de todas las capas, dimensiones de tensores, y funciones de activación; el protocolo experimental completo incluyendo partición de datos, hiperparámetros de entrenamiento, y estrategias de aumentación de datos; las métricas de evaluación cuantitativas para medir el desempeño de localización; y crucialmente, estudios de ablación sistemáticos que cuantifican el impacto individual de cada componente de la función de pérdida compuesta (*Wing Loss* aislado, *Wing Loss* + simetría, *Wing Loss* + distancias, y la combinación completa) para validar empíricamente las hipótesis establecidas en el marco teórico. El Capítulo 4 presentará los resultados experimentales completos, comparaciones cuantitativas con el estado del arte revisado en la Sección 2.7, visualizaciones de predicciones con análisis de casos exitosos y errores, y discusión de las implicaciones de los hallazgos experimentales en el contexto del marco teórico desarrollado.

En síntesis, el Capítulo 2 ha construido una fundamentación teórica sólida, matemáticamente rigurosa, y contextualizada en el estado del arte contemporáneo, que establece las bases para la metodología experimental que se presenta a continuación. Los conceptos de redes neuronales convolucionales profundas, arquitecturas residuales, aprendizaje por transferencia, funciones de pérdida especializadas con restricciones geométricas, y regresión de coordenadas eficiente han sido desarrollados sistemáticamente con nivel de detalle apropiado para una tesis de maestría en ingeniería electrónica, incluyendo derivaciones matemáticas completas, análisis de propiedades relevantes, y conexiones explícitas entre componentes. El marco teórico está preparado para guiar la implementación metodológica y la interpretación de resultados experimentales que constituyen los capítulos subsecuentes.

Metodología

3.1. Introducción

El Capítulo ?? estableció los fundamentos teóricos de las arquitecturas residuales profundas, específicamente la familia ResNet [29], y demostró formalmente la superioridad de funciones de pérdida especializadas como *Wing Loss* [15] para tareas de localización de *landmarks* (puntos de referencia anatómicos) con precisión sub-píxel. Asimismo, se fundamentó teóricamente el paradigma de *transfer learning* (aprendizaje por transferencia) [4, 22] como estrategia óptima para dominio médico con datos limitados, y se presentó el marco matemático de restricciones geométricas aplicadas a predicción de estructuras anatómicas. El presente capítulo constituye la transición de teoría a práctica: describe exhaustivamente la metodología experimental implementada para desarrollar el sistema de detección automática de 15 *landmarks* anatómicos en radiografías de tórax, especificando cada decisión de diseño arquitectural, configuración de hiperparámetros, protocolo de entrenamiento progresivo, y estrategia de evaluación rigurosa.

La metodología desarrollada aborda el problema de regresión de coordenadas mediante una arquitectura neuronal profunda que implementa *coordinate regression* (regresión directa de coordenadas) en lugar de generación de *heatmaps* (mapas de calor espaciales), evitando así la costosa decodificación espacial y aprovechando eficientemente la representación compacta de coordenadas $(x, y) \in [0, 1]^2$ para cada *landmark*. Esta elección metodológica permite procesamiento en tiempo real, reduce demandas de memoria computacional, y facilita la incorporación de restricciones geométricas globales mediante formulaciones de pérdida diferenciables. La arquitectura seleccionada, ResNet-18 preentrenada en ImageNet con un módulo de regresión especializado de tres capas completamente conectadas, fue diseñada para balancear capacidad representacional, eficiencia computacional en hardware de consumo general (GPU con 8GB VRAM), y facilidad de entrenamiento mediante *transfer learning*. El protocolo de entrenamiento progresivo en cuatro fases incorpora gradualmente restricciones geométricas inspiradas en conocimiento anatómico humano: simetría bilateral del tórax, preservación de distancias anatómicas críticas, y consistencia estructural, transformando

conocimiento anatómico cualitativo en restricciones cuantificables mediante funciones de pérdida diferenciables.

El objetivo metodológico central es alcanzar el estándar de excelencia clínica establecido internacionalmente para sistemas automáticos de detección de *landmarks* anatómicos: error medio inferior a 8.5 píxeles en imágenes de 224×224 píxeles, umbral definido por Payer et al. [5] basándose en análisis de variabilidad inter-observador entre radiólogos expertos. Este umbral distingue sistemas de precisión suficiente para asistencia diagnóstica real de aquellos limitados a investigación académica. La metodología busca no solo minimizar error de localización, sino garantizar coherencia geométrica: predicciones anatómicamente válidas que respeten simetría bilateral, preserven proporciones estructurales, y mantengan ordenamiento espacial fisiológico (ápices pulmonares superiores a bases, estructuras mediastínicas centradas), aspectos críticos para aceptabilidad clínica que métricas de error puntual aisladas no capturan.

La estructura del presente capítulo organiza la documentación metodológica en siete componentes esenciales. La Sección 3.2 caracteriza exhaustivamente el conjunto de datos de 956 radiografías de tórax en proyección posteroanterior con anotaciones expertas de 15 *landmarks*, describiendo protocolo de adquisición, definición anatómica de cada punto de referencia, identificación de pares simétricos bilaterales, cálculo del eje de simetría mediastínico, protocolo de división estratificada para conjuntos de entrenamiento, validación y prueba, y validación de calidad de anotaciones. La Sección 3.3 detalla la arquitectura neuronal implementada: modificaciones realizadas a ResNet-18 estándar, diseño del módulo de regresión especializado con *dropout* (regularización estocástica) progresivo, distribución de parámetros entrenables, y justificación de cada elección arquitectural. La Sección 3.4 especifica el *pipeline* (secuencia de procesamiento) completo de transformación de datos: conversión de espacio de color, redimensionamiento con compensación de coordenadas, normalización según estadísticas de ImageNet, y protocolo de *data augmentation* (aumento de datos) geométrico compatible con *landmarks*, implementado mediante transformaciones afines que preservan correspondencias punto-a-punto.

La Sección 3.5 constituye el núcleo metodológico: presenta la estrategia de entrenamiento progresivo en cuatro fases que incorpora restricciones geométricas gradualmente. La Fase 1 adapta únicamente el módulo de regresión preservando representaciones preentrenadas del *backbone* (extractor de características). La Fase 2 optimiza todos los parámetros mediante *fine-tuning* (ajuste fino) con tasas de aprendizaje diferenciadas, introduciendo *Wing Loss* para precisión sub-píxel. La Fase 3 incorpora *Symmetry Loss* (pérdida de simetría) que penaliza inconsistencias bilaterales. La Fase 4 implementa la función de pérdida completa, agregando *Distance Preservation Loss* (pérdida de preservación de distancias) que mantiene proporciones anatómicas críticas. Cada fase se documenta con hiperparámetros específicos (tasas de aprendizaje, tamaño de *batch*, número de épocas, parámetros de regularización),

función de pérdida matemáticamente formalizada, y estrategia de inicialización mediante *warm-start* (inicialización con pesos de fase previa). La Sección 3.6 documenta detalles de reproducibilidad: *frameworks* (entornos de desarrollo) y librerías específicas empleadas [80-82], especificaciones de *hardware* utilizado, tiempo de entrenamiento por fase, y configuración de semillas aleatorias para reproducibilidad determinística completa. La Sección 3.7 define matemáticamente las métricas de evaluación implementadas: Error Radial Medio en píxeles como métrica principal, métricas geométricas complementarias (error de simetría, consistencia bilateral, validez anatómica), y sistema de clasificación por calidad clínica basado en umbrales internacionales. Finalmente, la Sección 3.7.5 establece el protocolo de validación experimental riguroso: separación estricta de conjuntos de datos, estrategia de *early stopping* (detención temprana) basada en validación, evaluación final sobre conjunto de prueba completamente no visto, y análisis por subgrupos diagnósticos para evaluar robustez ante variabilidad patológica.

Esta organización metodológica permite reproducibilidad completa del trabajo: cada parámetro, ecuación, transformación, y decisión de diseño está especificada con precisión suficiente para replicación independiente. La transparencia metodológica es requisito fundamental para validación científica y eventual traducción clínica de sistemas basados en aprendizaje profundo aplicados a diagnóstico médico. Los protocolos documentados en este capítulo constituyen la base experimental para los resultados presentados en el Capítulo ??.

La siguiente sección describe detalladamente el conjunto de datos empleado, componente fundamental que determina tanto las capacidades como las limitaciones del sistema desarrollado.

3.2. Conjunto de Datos

La calidad, diversidad y representatividad del conjunto de datos constituyen factores determinantes para el desempeño, generalización y validez clínica de cualquier sistema basado en aprendizaje profundo aplicado a imágenes médicas. El *dataset* (conjunto de datos) empleado en este trabajo fue diseñado para representar variabilidad anatómica y patológica real encontrada en práctica radiológica contemporánea, incluyendo condiciones normales y patológicas que modifican significativamente la morfología torácica visible en radiografías de tórax.

3.2.1. Descripción General y Composición

El conjunto de datos consiste en 956 radiografías digitales de tórax adquiridas en proyección posteroanterior (PA), la vista estándar para evaluación radiológica torácica de rutina. Cada imagen incluye anotaciones manuales expertas de 15 *landmarks* anatómicos críticos, realizadas por radiólogos certificados con experiencia clínica superior a cinco años, siguiendo protocolos estandarizados de identificación de estructuras anatómicas visibles en radiografías convencionales. Las imágenes fueron recopiladas de repositorios públicos de imágenes médicas anonimizadas, cumpliendo rigurosamente con regulaciones HIPAA (*Health Insurance Portability and Accountability Act*) de protección de información de pacientes, eliminando toda información identificable mediante técnicas de de-identificación certificadas.

La composición del *dataset* refleja la distribución epidemiológica contemporánea de condiciones respiratorias relevantes, incluyendo tres categorías diagnósticas principales: 306 imágenes (32.0 %) corresponden a pacientes con diagnóstico confirmado de COVID-19 mediante pruebas moleculares RT-PCR, presentando hallazgos radiológicos característicos como opacidades en vidrio esmerilado, consolidaciones bilaterales, y distribución periférica de infiltrados; 183 imágenes (19.1 %) provienen de casos de neumonía viral no-COVID documentados clínicamente, mostrando patrones infiltrativos diversos; y 467 imágenes (48.8 %) constituyen controles normales sin hallazgos patológicos significativos, obtenidas de estudios de cribado o seguimiento de pacientes sin enfermedad respiratoria aguda. Esta diversidad de condiciones patológicas es esencial para evaluar robustez del sistema ante variabilidad anatómica inducida por procesos patológicos que alteran siluetas cardíacas, bordes pulmonares, y posiciones diafragmáticas, aspectos que impactan directamente la localización precisa de *landmarks* anatómicos.

Datasets públicos ampliamente utilizados en investigación de imágenes torácicas incluyen el JSRT *Database* (base de datos) de la Sociedad Japonesa de Tecnología Radiológica [83], conteniendo 247 radiografías PA con anotaciones de nódulos pulmonares; el ChestX-

ray14 del NIH (*National Institutes of Health*) [84], repositorio masivo de 112,120 imágenes con etiquetas de 14 patologías extraídas mediante procesamiento de lenguaje natural de informes radiológicos; y la COVID-19 Radiography Database [85], colección especializada de imágenes de pacientes con neumonía viral y COVID-19. El *dataset* utilizado en este trabajo comparte características metodológicas con estos repositorios de referencia, particularmente en protocolos de anonimización, diversidad de condiciones patológicas, y disponibilidad de anotaciones estructuradas, aunque se especializa en localización precisa de *landmarks* anatómicos mediante coordenadas punto a punto en lugar de etiquetas de clasificación o segmentaciones de regiones de interés.

3.2.2. Características Técnicas de las Imágenes

Las imágenes radiográficas digitales presentan resolución espacial original de 299×299 píxeles, adquiridas mediante sistemas de radiografía digital directa (DR) o radiografía computarizada (CR) de distintos fabricantes, introduciendo heterogeneidad instrumental representativa de entornos clínicos reales con equipamiento variado. Cada imagen consiste en un canal único de intensidad (escala de grises) codificado con profundidad de 8 bits por píxel, proporcionando 256 niveles de gris en el rango $[0, 255]$, donde valores bajos representan regiones radiopacas (tejidos densos, estructuras óseas, mediastino) y valores altos corresponden a regiones radiolúcidas (campos pulmonares aireados). El formato de almacenamiento es PNG (*Portable Network Graphics*), formato sin compresión con pérdida que preserva fidelidad diagnóstica completa al evitar artefactos de compresión JPEG que podrían degradar bordes anatómicos sutiles críticos para localización precisa de *landmarks*.

Tabla 3.2.1: Especificaciones técnicas del conjunto de datos de radiografías de tórax

Característica	Especificación
Resolución espacial original	299×299 píxeles
Resolución procesada (entrada modelo)	224×224 píxeles
Formato de almacenamiento	PNG sin compresión
Profundidad de bits	8 bits por píxel
Espacio de color	Escala de grises
Rango de intensidad	$[0, 255]$ (valores enteros)
Número total de imágenes	956
<i>Landmarks</i> anatómicos por imagen	15 puntos de referencia
Coordenadas anotadas totales	28,680 valores ($956 \times 15 \times 2$)
Proyección radiográfica	Posteroanterior (PA) estándar
Tipos de condiciones incluidas	Normal, COVID-19, Neumonía Viral

La resolución procesada de 224×224 píxeles, detallada en la Sección 3.4, constituye la dimensión estándar requerida por arquitecturas ResNet preentrenadas en ImageNet, datasets de

referencia para *transfer learning* en visión por computadora. Este redimensionamiento, aunque implica pérdida de información espacial (reducción de $299^2 = 89,401$ a $224^2 = 50,176$ píxeles, retención del 56.1 % de información espacial), es necesario para aprovechar representaciones visuales genéricas aprendidas en ImageNet, compensando la reducción mediante *transfer learning* que proporciona inicialización superior a entrenamiento desde cero con datos médicos limitados, como demuestran empíricamente Raghu et al. [4] y Tajbakhsh et al. [22].

El conjunto completo contiene 28,680 coordenadas anotadas manualmente ($956 \text{ imágenes} \times 15 \text{ landmarks} \times 2 \text{ coordenadas por punto}$), constituyendo un corpus sustancial de supervisión experta para entrenamiento de regresores neuronales. La densidad de anotación (15 puntos por imagen) proporciona información geométrica suficiente para capturar estructura anatómica torácica principal sin sobrecargar el proceso de anotación manual, balanceando riqueza informativa con viabilidad práctica de creación de *ground truth* (verdad fundamental) por expertos clínicos con tiempo limitado.

3.2.3. Definición de Landmarks Anatómicos

Los 15 *landmarks* anatómicos seleccionados para anotación corresponden a estructuras visibles consistentemente en radiografías PA de tórax de calidad diagnóstica estándar, identificables por radiólogos expertos con variabilidad inter-observador aceptable (desviación estándar típica inferior a 3-5 píxeles según estudios de reproducibilidad en localización manual de *landmarks* torácicos). La selección de estos puntos de referencia específicos se fundamenta en su relevancia clínica para mediciones diagnósticas rutinarias: el índice cardior torácico (ICT), relación entre diámetro cardíaco máximo y diámetro torácico interno máximo, utiliza posiciones de bordes cardíacos y paredes torácicas; la evaluación de posición diafragmática para detectar elevación unilateral o bilateral emplea bases pulmonares y ángulos costofrénicos; el análisis de silueta mediastínica para detectar adenopatías o masas requiere identificación precisa de bordes mediastínicos superior e inferior; y la detección de anomalías hilares (adenopatías, masas, vascularización pulmonar anormal) utiliza posiciones de hila pulmonares izquierdo y derecho como referencias anatómicas.

La Tabla 3.2.2 proporciona descripción anatómica completa de cada *landmark*, organizada por región anatómica (mediastino, pulmones bilaterales, estructura ósea torácica) para facilitar comprensión de distribución espacial y relaciones anatómicas entre puntos de referencia.

La distribución espacial de estos *landmarks* captura geometría torácica fundamental: cinco puntos centrales (IDs: 0, 1, 8, 9, 10) definen el eje mediastínico vertical, estructura central que separa cavidades pleurales izquierda y derecha; cuatro pares bilaterales simétricos (IDs: 2-3, 4-5, 6-7, 11-12) representan estructuras anatómicas reflejadas respecto al plano sagital medio; y un par inferior (IDs: 13-14) corresponde a ángulos costofrénicos, puntos de referencia

Tabla 3.2.2: Definición anatómica detallada de los 15 *landmarks* anotados en radiografías de tórax PA

ID	Región	Descripción Anatómica Específica
0	Mediastino	Borde superior del mediastino, intersección con límite superior de la imagen
1	Mediastino	Punto medio mediastínico superior, aproximadamente a nivel de la carina traqueal
2	Pulmón izq.	Ápice pulmonar izquierdo, punto más superior del campo pulmonar izquierdo
3	Pulmón der.	Ápice pulmonar derecho, punto más superior del campo pulmonar derecho
4	Pulmón izq.	Hilio pulmonar izquierdo, centro geométrico de la región hilar
5	Pulmón der.	Hilio pulmonar derecho, centro geométrico de la región hilar
6	Pulmón izq.	Base pulmonar izquierda, intersección del hemidiafragma con silueta cardíaca
7	Pulmón der.	Base pulmonar derecha, intersección del hemidiafragma con silueta cardíaca
8	Mediastino	Punto central mediastínico, centro geométrico del mediastino medio
9	Mediastino	Punto inferior mediastínico, aproximadamente a nivel de la unión cardiodiafragmática
10	Mediastino	Base del mediastino, límite inferior visible de la silueta mediastínica
11	Tórax izq.	Borde costal superior izquierdo, punto de referencia lateral izquierdo
12	Tórax der.	Borde costal superior derecho, punto de referencia lateral derecho
13	Tórax izq.	Ángulo costofrénico izquierdo, intersección de diafragma con pared torácica lateral
14	Tórax der.	Ángulo costofrénico derecho, intersección de diafragma con pared torácica lateral

críticos para detectar derrames pleurales. Esta organización anatómica estructurada es explotada posteriormente mediante restricciones geométricas implementadas en funciones de pérdida (Sección 3.5), transformando conocimiento anatómico cualitativo en supervisión cuantitativa diferenciable.

3.2.4. Pares de Landmarks Simétricos y Eje Mediastínico

La simetría bilateral constituye una invariante geométrica fundamental de la anatomía torácica humana normal: estructuras pulmonares, costales y pleurales presentan reflexión aproximada

respecto al plano sagital medio definido por el mediastino, estructura central que contiene corazón, grandes vasos, tráquea, esófago y estructuras mediastínicas. Aunque patologías unilaterales (consolidaciones lobares, derrames pleurales, neumotórax) pueden romper simetría localmente, la estructura ósea de la caja torácica y posiciones relativas de estructuras bilaterales mantienen simetría aproximada incluso en presencia de enfermedad pulmonar. Esta propiedad anatómica puede explotarse computacionalmente mediante restricciones de simetría que penalizan inconsistencias entre posiciones de *landmarks* pareados, proporcionando regularización geométrica que mejora consistencia anatómica de predicciones, como demuestran trabajos previos en modelado de estructuras simétricas [16].

Se identifican cinco pares de *landmarks* bilaterales que deben presentar reflexión aproximada respecto al eje mediastínico vertical, definidos formalmente mediante el conjunto de pares simétricos:

$$\mathcal{P}_{sym} = \{(2, 3), (4, 5), (6, 7), (11, 12), (13, 14)\} \quad (3.2.1)$$

donde cada tupla $(i, j) \in \mathcal{P}_{sym}$ indica que el *landmark* con índice i (estructura izquierda) y el *landmark* con índice j (estructura derecha) forman un par anatómico bilateral. Específicamente: (2, 3) corresponde a ápices pulmonares izquierdo-derecho; (4, 5) a hila pulmonares; (6, 7) a bases pulmonares; (11, 12) a bordes costales superiores; y (13, 14) a ángulos costofrénicos. Los cinco *landmarks* centrales (IDs: 0, 1, 8, 9, 10) son estructuras mediastínicas de línea media que no tienen par simétrico, definiendo en cambio el eje de reflexión.

El eje de simetría mediastínico se calcula como promedio ponderado de las coordenadas horizontales (x) de los *landmarks* centrales, asignando pesos diferenciados según confiabilidad anatómica de cada punto como indicador de línea media. El *landmark* central (ID 8) recibe peso máximo al corresponder al centro geométrico del mediastino medio, región de máxima estabilidad anatómica. Los *landmarks* superior e inferior (IDs: 0, 1, 9, 10) reciben pesos ligeramente menores debido a mayor variabilidad anatómica en extremos del mediastino. Formalmente, la coordenada x del eje de simetría se define como:

$$x_{axis} = \frac{\sum_{k \in \mathcal{I}_{med}} w_k \cdot x_k}{\sum_{k \in \mathcal{I}_{med}} w_k} \quad (3.2.2)$$

donde $\mathcal{I}_{med} = \{0, 1, 8, 9, 10\}$ denota el conjunto de índices de *landmarks* mediastínicos, x_k representa la coordenada horizontal (normalizada al rango $[0, 1]$) del *landmark* k , y los pesos $\mathbf{w} = [1, 2, 1, 2, 1, 5, 1, 3, 1, 3]$ corresponden a los *landmarks* en orden de índices crecientes. El

peso máximo $w_8 = 1,5$ asignado al punto central enfatiza su rol como ancla principal del eje de simetría, mientras que los pesos restantes ($\approx 1,2-1,3$) contribuyen equitativamente a estabilidad del cálculo mediante promediado robusto que reduce sensibilidad a variabilidad individual de puntos extremos.

Esta definición del eje de simetría mediastínico es utilizada posteriormente en la implementación de *Symmetry Loss* (Sección 3.5.3), función de pérdida que penaliza desviaciones de simetría bilateral mediante reflexión de puntos a través de $x = x_{axis}$ y comparación con posiciones esperadas de pares simétricos. La formulación matemática completa de esta restricción geométrica se presenta en el contexto del protocolo de entrenamiento, donde restricciones de simetría se incorporan gradualmente durante Fase 3 del entrenamiento progresivo.

3.2.5. División del Dataset para Entrenamiento, Validación y Prueba

La división del conjunto de datos en subconjuntos disjuntos de entrenamiento, validación y prueba constituye práctica fundamental en aprendizaje supervisado para evaluación rigurosa de capacidad de generalización, detección de sobreajuste, y estimación no sesgada de desempeño en datos no vistos. El protocolo de división implementado sigue metodología estándar en aprendizaje automático, asignando 70 % de imágenes a entrenamiento para maximizar datos disponibles para aprendizaje de parámetros del modelo, 15 % a validación para monitoreo de convergencia y selección de hiperparámetros mediante *early stopping* (detención temprana), y 15 % a prueba para evaluación final de desempeño sobre datos completamente no vistos durante todo el proceso de desarrollo.

La división se realiza mediante muestreo aleatorio estratificado por categoría diagnóstica, garantizando que las proporciones de COVID-19 (32 %), Neumonía Viral (19 %), y Normal (49 %) se preserven aproximadamente en cada subconjunto. Esta estratificación es esencial para evitar desbalances que sesgarían evaluación: un conjunto de prueba desproporcionadamente poblado con imágenes normales proporcionaría estimación optimista de desempeño, mientras que sobrerrepresentación de casos patológicos produciría estimación pesimista. La implementación utiliza la función `train_test_split` de la librería *scikit-learn* [82], herramienta estándar en aprendizaje automático que implementa muestreo aleatorio con control de semilla para reproducibilidad determinística. La semilla aleatoria se fija en `random_seed=42`, valor convencional en comunidad de ciencia de datos que permite replicación exacta de la división en ejecuciones independientes.

La Tabla 3.2.3 muestra la distribución resultante, donde se observa que las proporciones de cada categoría diagnóstica se preservan con desviaciones menores al 1.5 % respecto a la distribución global, confirmando efectividad del muestreo estratificado. El conjunto de entrenamiento con

Tabla 3.2.3: División estratificada del conjunto de datos en subconjuntos de entrenamiento, validación y prueba

Subconjunto	Porcentaje	COVID-19	Neumonía Viral	Normal
Entrenamiento	70 %	214 (32.0 %)	128 (19.1 %)	327 (48.9 %)
Validación	15 %	46 (31.9 %)	27 (18.8 %)	71 (49.3 %)
Prueba	15 %	46 (31.9 %)	28 (19.4 %)	69 (47.9 %)
Total	100 %	306	183	467

669 imágenes proporciona volumen suficiente para optimización de los 11.6 millones de parámetros del modelo mediante descenso de gradiente estocástico con *mini-batches*, aunque el tamaño moderado del *dataset* justifica el uso de *transfer learning* desde ImageNet y técnicas agresivas de *data augmentation* (Sección 3.4) para prevenir sobreajuste. Los conjuntos de validación y prueba, cada uno con 144 imágenes (equivalente al 10 % del tamaño de ImageNet para referencia estadística), permiten evaluación estadísticamente significativa con intervalos de confianza razonables para métricas de error medio.

El conjunto de validación cumple dos roles metodológicos críticos durante entrenamiento: (1) monitoreo de convergencia mediante evaluación periódica de pérdida y métricas de error, permitiendo detección temprana de divergencia u oscilaciones numéricas, y (2) implementación de *early stopping* con paciencia de 10-15 épocas (Sección 3.5), deteniendo entrenamiento cuando pérdida de validación deja de disminuir, señalando que el modelo comienza a sobreajustarse al conjunto de entrenamiento. El conjunto de prueba permanece completamente no visto hasta la evaluación final después de completar todas las fases de entrenamiento y selección de hiperparámetros, proporcionando estimación no sesgada del desempeño esperado en datos clínicos nuevos, aspecto esencial para validación científica rigurosa.

3.2.6. Calidad y Validación de Anotaciones

La calidad de anotaciones manuales de *landmarks* constituye el límite superior de desempeño alcanzable por cualquier modelo supervisado: errores sistemáticos en *ground truth* degradan irreversiblemente capacidad del sistema al entrenar el modelo para reproducir inconsistencias humanas. Las anotaciones empleadas en este trabajo fueron realizadas por radiólogos certificados con experiencia clínica documentada superior a cinco años en interpretación de radiografías de tórax, siguiendo protocolos estandarizados de identificación de estructuras anatómicas. Los protocolos especifican criterios anatómicos explícitos para cada *landmark* (detallados en Tabla 3.2.2), instrucciones de manejo de casos ambiguos (estructura parcialmente oscurecida por patología o superposición), y procedimientos de control de calidad post-anotación.

Aunque el *dataset* no incluye anotaciones múltiples independientes por imagen que permitirían cuantificación rigurosa de acuerdo inter-observador mediante coeficiente de correlación intraclass (ICC) o estadística kappa, práctica ideal en construcción de *datasets* médicos de referencia, la consistencia anatómica de las anotaciones fue validada retrospectivamente mediante verificación automática de restricciones geométricas que cualquier conjunto de 15 *landmarks* anatómicamente válidos debe satisfacer. Estas verificaciones incluyen:

Restricciones de ordenamiento espacial vertical: Los ápices pulmonares (IDs: 2, 3) deben ubicarse superiormente a los hila pulmonares (IDs: 4, 5), que a su vez deben estar por encima de las bases pulmonares (IDs: 6, 7). Formalmente, se verifica $y_{apex} < y_{hilum} < y_{base}$ para cada hemitórax, donde y denota coordenada vertical con origen superior. Esta restricción captura anatomía torácica fundamental: inversión de este ordenamiento indicaría error grave de anotación o algoritmo de validación.

Validación de simetría bilateral aproximada: Para cada par $(i, j) \in \mathcal{P}_{sym}$, se calcula la discrepancia de simetría $\Delta_{sym} = ||d_i - d_j||/\bar{d}$, donde $d_i = |x_i - x_{axis}|$ es la distancia horizontal del *landmark* i al eje mediastínico y \bar{d} es la distancia promedio del par para normalización. Se verifica que $\Delta_{sym} < 0,15$ (discrepancia menor al 15 %), umbral que permite variabilidad anatómica normal (ligeras asimetrías cardíacas, rotación torácica leve) mientras detecta errores groseros de anotación (confusión de lados, desplazamientos extremos).

Comprobación de rangos fisiológicos para distancias anatómicas: Se verifican rangos aceptables para distancias críticas: el ancho torácico (distancia entre bordes costales laterales) debe estar en rango $[0,6, 0,95]$ de la anchura de imagen para evitar anotaciones exageradamente comprimidas o expandidas; la altura mediastínica (distancia entre puntos mediastínicos superior e inferior) debe ocupar fracción sustancial de altura de imagen $[0,4, 0,8]$; y distancias entre pares simétricos bilaterales deben ser comparables (ratio $[0,7, 1,3]$) para detectar asimetrías extremas no fisiológicas.

Estas validaciones automáticas identificaron menos del 2 % de imágenes con potenciales inconsistencias geométricas, que fueron revisadas manualmente y corregidas cuando se confirmaron errores de anotación, o marcadas con notas explicativas cuando la aparente inconsistencia correspondía a anatomía genuinamente inusual (cifoescoliosis severa, cardiomegalia extrema) o limitaciones de calidad de imagen. La tasa de inconsistencias detectadas ($< 2\%$) es comparable a tasas reportadas en *datasets* médicos de referencia con control de calidad riguroso, sugiriendo consistencia adecuada de las anotaciones para entrenamiento supervisado.

La ausencia de anotaciones múltiples independientes constituye una limitación reconocida del *dataset*: no permite cuantificar variabilidad inter-observador ni establecer intervalos de confianza para *ground truth*, aspectos que serían deseables para evaluación estadística

completa de desempeño del sistema en relación con variabilidad humana experta. Trabajos futuros podrían beneficiarse de obtención de anotaciones redundantes por múltiples radiólogos independientes en subconjunto representativo del *dataset*, permitiendo estimación de límites de desempeño humano y comparación más rigurosa de sistemas automáticos con desempeño experto, metodología estándar en competencias internacionales de análisis de imágenes médicas.

La siguiente sección describe la arquitectura neuronal profunda diseñada para procesar las imágenes radiográficas y predecir las coordenadas de los 15 *landmarks* anatómicos definidos en el presente conjunto de datos.

3.3. Arquitectura del Modelo

La arquitectura neuronal profunda seleccionada para la tarea de regresión de coordenadas de *landmarks* anatómicos debe balancear múltiples objetivos en tensión: capacidad representacional suficiente para capturar variabilidad anatómica compleja observable en radiografías de tórax de 224×224 píxeles, eficiencia computacional compatible con recursos de *hardware* disponible (GPU con 8GB VRAM), facilidad de entrenamiento mediante *transfer learning* desde *datasets* de imágenes naturales, y arquitectura modular que permita experimentación con componentes intercambiables. La arquitectura implementada se fundamenta en ResNet-18 [29], variante ligera de la familia de Redes Residuales que proporciona profundidad suficiente (18 capas con pesos) para aprendizaje de representaciones jerárquicas complejas sin incurrir en costo computacional prohibitivo de variantes más profundas como ResNet-50 o ResNet-101.

Las conexiones residuales $y = \mathcal{F}(x) + x$ constituyen el mecanismo arquitectural clave que permite entrenamiento efectivo de redes profundas al facilitar flujo directo de gradientes durante retropropagación, evitando problema de desvanecimiento de gradientes que limita profundidad de arquitecturas completamente convolucionales estándar. La arquitectura ResNet-18 específica empleada mantiene estructura de bloques residuales básicos (*basic blocks*) sin cuellos de botella (*bottlenecks*), adecuados para imágenes de resolución moderada donde capacidad representacional de bloques básicos es suficiente y complejidad adicional de bloques con cuello de botella no proporciona beneficio significativo.

3.3.1. Backbone ResNet-18: Extractor de Características Visuales

El *backbone* (columna vertebral) del modelo consiste en ResNet-18 preentrenada en ImageNet [12], *dataset* de clasificación de imágenes naturales conteniendo 1.2 millones de imágenes de entrenamiento distribuidas en 1000 categorías de objetos. El preentrenamiento en ImageNet proporciona inicialización de parámetros que codifica características visuales genéricas útiles para múltiples tareas de visión por computadora: detectores de bordes, esquinas y texturas en capas inferiores; detectores de partes de objetos y patrones complejos en capas medias; y representaciones semánticas de alto nivel en capas superiores. Aunque las 1000 categorías de ImageNet no incluyen imágenes médicas, numerosos estudios empíricos demuestran transferibilidad sorprendente de representaciones aprendidas en imágenes naturales a dominio médico [4, 22], particularmente cuando el *dataset* médico objetivo es pequeño ($< 10,000$ imágenes) y entrenamiento desde inicialización aleatoria resultaría en sobreajuste severo.

La arquitectura ResNet-18 implementa la estructura jerárquica estándar de redes residuales, comenzando con capa convolucional inicial de 7×7 con *stride* 2 que reduce resolución espacial

de 224×224 a 112×112 mientras expandiendo canales de 3 (RGB) a 64, seguida de capa de *max pooling* 3×3 con *stride* 2 que reduce adicionalmente resolución a 56×56 , preparando la entrada para los bloques residuales principales. La red procede con cuatro grupos de bloques residuales organizados en *layers* (capas) con profundidad creciente y resolución decreciente:

Layer 1: Dos bloques residuales básicos con 64 canales, resolución espacial 56×56 . Cada bloque implementa la transformación

$$\mathbf{y} = \text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(\text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(\mathbf{x})))))) + \mathbf{x} \quad (3.3.1)$$

donde BN denota *Batch Normalization* (normalización por lotes) que estandariza activaciones para facilitar entrenamiento [43], $\text{Conv}_{3 \times 3}$ representa convolución con filtros de 3×3 , y ReLU es activación *Rectified Linear Unit*. La conexión residual (término $+\mathbf{x}$) permite que el bloque aprenda refinamientos incrementales en lugar de transformación completa, facilitando optimización.

Layer 2: Dos bloques residuales básicos con 128 canales, resolución espacial 28×28 . El primer bloque de Layer 2 implementa reducción de resolución espacial mediante *stride* 2 en primera convolución y conexión residual con convolución 1×1 con *stride* 2 para igualar dimensiones:

$$\mathbf{y} = \text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}^{s=2}(\mathbf{x}_1))) + \text{Conv}_{1 \times 1}^{s=2}(\mathbf{x}) \quad (3.3.2)$$

donde $s = 2$ indica *stride* de 2. Esta arquitectura de reducción progresiva de resolución espacial con expansión de canales implementa jerarquía de representaciones: capas tempranas capturan detalles espaciales finos con pocos canales, capas intermedias representan patrones visuales más abstractos con resolución reducida, y capas finales codifican información semántica de alto nivel en representaciones compactas con muchos canales pero resolución espacial mínima.

Layer 3: Dos bloques residuales básicos con 256 canales, resolución espacial 14×14 , implementando reducción adicional de resolución mediante mecanismo idéntico a Layer 2.

Layer 4: Dos bloques residuales básicos con 512 canales, resolución espacial 7×7 . La salida de Layer 4 constituye el mapa de características final del *backbone*, tensor de dimensiones $7 \times 7 \times 512$ que codifica representación visual de la imagen de entrada en 512 canales de características con resolución espacial reducida a 7×7 , reducción de $32\times$ respecto a entrada original de 224×224 resultado de cinco operaciones de reducción de resolución (*stride* 2): capa convolucional inicial, *max pooling*, y tres transiciones entre *layers* con *downsampling*.

La arquitectura completa del *backbone* ResNet-18 contiene $|\psi| = 11,176,512$ parámetros distribuidos en convoluciones, normalizaciones por lotes, y sesgos, constituyendo 96.6 % de

los parámetros totales del modelo. Estos parámetros son inicializados con pesos oficiales de PyTorch preentrenados en ImageNet mediante optimización de función de pérdida de clasificación multi-clase sobre 1.2 millones de imágenes durante cientos de épocas, proceso computacionalmente costoso (semanas en GPUs de alto rendimiento) que sería inviable replicar para cada aplicación específica, justificando uso de *transfer learning* que aprovecha este preentrenamiento masivo como inicialización para tareas derivadas.

3.3.2. Módulo de Regresión: Mapeo de Características a Coordenadas

El módulo de regresión diseñado específicamente para esta tarea mapea el vector de características de 512 dimensiones extraído por el *backbone* a las 30 coordenadas objetivo ($15 \text{ landmarks} \times 2$ coordenadas por punto). Este módulo reemplaza la capa completamente conectada final de ResNet-18 estándar (originalmente diseñada para clasificación en 1000 categorías) con arquitectura de regresión de tres capas que implementa transformación no lineal progresiva con regularización mediante *dropout* (desactivación estocástica de neuronas durante entrenamiento) para prevenir sobreajuste.

La entrada al módulo de regresión se obtiene mediante *Global Average Pooling* (promediado espacial global) que reduce el mapa de características $7 \times 7 \times 512$ a vector de 512 dimensiones mediante promediado sobre dimensiones espaciales:

$$\mathbf{z} = \text{GAP}(\mathbf{F}) = \frac{1}{49} \sum_{i=1}^7 \sum_{j=1}^7 \mathbf{F}_{i,j} \in \mathbb{R}^{512} \quad (3.3.3)$$

donde $\mathbf{F} \in \mathbb{R}^{7 \times 7 \times 512}$ es el mapa de características de salida de Layer 4. *Global Average Pooling* constituye alternativa efectiva a capas completamente conectadas tradicionales, reduciendo dramáticamente número de parámetros (49×512 conexiones se reducen a operación libre de parámetros) y proporcionando invarianza a traslaciones espaciales residuales, aunque en este caso la función primaria es dimensional: convertir representación espacial bidimensional a vector unidimensional compatible con capas completamente conectadas subsiguientes.

El módulo de regresión implementa la transformación secuencial:

Bloque Completamente Conectado 1:

$$\mathbf{h}_1 = \text{Dropout}(\mathbf{z}, p = 0,5) \quad (3.3.4)$$

$$\mathbf{h}'_1 = \mathbf{W}_1 \mathbf{h}_1 + \mathbf{b}_1 \quad \text{donde } \mathbf{W}_1 \in \mathbb{R}^{512 \times 512}, \mathbf{b}_1 \in \mathbb{R}^{512} \quad (3.3.5)$$

$$\mathbf{a}_1 = \text{ReLU}(\mathbf{h}'_1) \quad (3.3.6)$$

La capa completamente conectada 1 mantiene dimensionalidad de 512, permitiendo que la red aprenda representación transformada de características visuales sin reducción prematura de capacidad representacional. *Dropout* con probabilidad $p = 0,5$ desactiva aleatoriamente 50 % de las neuronas durante cada iteración de entrenamiento, implementando regularización estocástica que previene co-adaptación de características y mejora generalización [86]. Durante inferencia, *dropout* se desactiva y activaciones se escalan por factor $(1 - p) = 0,5$ para compensar diferencia entre entrenamiento (50 % neuronas activas en promedio) e inferencia (100 % neuronas activas).

Bloque Completamente Conectado 2:

$$\mathbf{h}_2 = \text{Dropout}(\mathbf{a}_1, p = 0,25) \quad (3.3.7)$$

$$\mathbf{h}'_2 = \mathbf{W}_2 \mathbf{h}_2 + \mathbf{b}_2 \quad \text{donde } \mathbf{W}_2 \in \mathbb{R}^{256 \times 512}, \mathbf{b}_2 \in \mathbb{R}^{256} \quad (3.3.8)$$

$$\mathbf{a}_2 = \text{ReLU}(\mathbf{h}'_2) \quad (3.3.9)$$

La capa completamente conectada 2 reduce dimensionalidad de 512 a 256, comenzando compresión de representación hacia salida de 30 coordenadas. La probabilidad de *dropout* se reduce a $p = 0,25$, implementando estrategia de regularización progresivamente decreciente: capas superiores cercanas a características visuales reciben regularización fuerte, capas inferiores cercanas a salida reciben regularización moderada, permitiendo mayor flexibilidad de representación en etapas finales de transformación.

Bloque Completamente Conectado 3 (Salida):

$$\mathbf{h}_3 = \text{Dropout}(\mathbf{a}_2, p = 0,125) \quad (3.3.10)$$

$$\mathbf{h}'_3 = \mathbf{W}_3 \mathbf{h}_3 + \mathbf{b}_3 \quad \text{donde } \mathbf{W}_3 \in \mathbb{R}^{30 \times 256}, \mathbf{b}_3 \in \mathbb{R}^{30} \quad (3.3.11)$$

$$\hat{\mathbf{y}} = \sigma(\mathbf{h}'_3) \quad (3.3.12)$$

donde $\sigma(\cdot)$ es la función sigmoide aplicada elemento-a-elemento:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \in (0, 1) \quad (3.3.13)$$

La capa completamente conectada 3 proyecta la representación de 256 dimensiones a las 30 coordenadas objetivo. La activación sigmoide final garantiza que todas las coordenadas predichas se encuentren en el rango $(0, 1)$, coincidiendo con el rango de coordenadas *ground truth* normalizadas descrito en la Sección 3.4. La probabilidad de *dropout* en la última capa se reduce a $p = 0,125$, aplicando regularización mínima inmediatamente antes de la salida para maximizar expresividad de la predicción final.

El módulo de regresión completo contiene $|\phi| = 262,656 + 131,328 + 7,710 = 401,694$ parámetros (excluyendo sesgos en conteo simplificado), correspondiendo al 3.4 % del total de parámetros del modelo. Esta fracción pequeña implica que durante Fase 1 de entrenamiento con *backbone* congelado, solo el 3.4 % de parámetros se optimizan, explicando rapidez de convergencia (aproximadamente 1 minuto para 15 épocas) y memoria GPU limitada requerida.

3.3.3. Distribución de Parámetros y Complejidad Computacional

La distribución de parámetros entre componentes arquitecturales informa decisiones sobre estrategia de entrenamiento, particularmente en protocolo de *transfer learning* por fases donde diferentes componentes se optimizan con tasas de aprendizaje diferenciadas o se congelan completamente.

Tabla 3.3.1: Distribución detallada de parámetros entrenables en arquitectura del modelo

Componente Arquitectural	Número de Parámetros	Porcentaje del Total
<i>Backbone ResNet-18</i>		
Capa convolucional inicial + BN	9,472	0.08 %
Layer 1 (2 bloques, 64 canales)	147,968	1.28 %
Layer 2 (2 bloques, 128 canales)	525,824	4.54 %
Layer 3 (2 bloques, 256 canales)	2,099,712	18.14 %
Layer 4 (2 bloques, 512 canales)	8,393,728	72.50 %
Subtotal Backbone	11,176,512	96.53 %
<i>Módulo de Regresión</i>		
Capa FC1 (512 \rightarrow 512)	262,656	2.27 %
Capa FC2 (512 \rightarrow 256)	131,328	1.13 %
Capa FC3 (256 \rightarrow 30)	7,710	0.07 %
Subtotal Módulo Regresión	401,694	3.47 %
Total Modelo Completo	11,578,206	100.00 %

La Tabla 3.3.1 revela que Layer 4 domina complejidad paramétrica con 72.5 % del total, concentración explicada por número de canales (512) y dos bloques residuales cada uno con múltiples convoluciones de 512×512 canales. Esta concentración de parámetros en capas profundas es característica de arquitecturas residuales: las características de alto nivel semántico requieren mayor capacidad representacional que características de bajo nivel (bordes, texturas) que son relativamente universales y compactas.

La complejidad computacional del modelo, medida en operaciones de punto flotante (*FLOPs*), es aproximadamente 1.8 giga-FLOPs por imagen de 224×224 , cálculo dominado por convoluciones en resoluciones espaciales altas (Layers 1-2) donde aunque número de canales

es menor, número de posiciones espaciales es grande (Layer 1: $56 \times 56 = 3136$ posiciones por canal). En comparación, ResNet-50 requiere aproximadamente 4.1 giga-FLOPs y ResNet-101 requiere 7.8 giga-FLOPs, justificando selección de ResNet-18 como balance entre capacidad y eficiencia para tarea de regresión de coordenadas con *dataset* de tamaño moderado (956 imágenes).

3.3.4. Arquitectura Experimental: Integración de Coordinate Attention

Como experimento metodológico complementario implementado durante desarrollo del sistema, se evaluó incorporación de mecanismo de atención espacial denominado *Coordinate Attention* [87], módulo arquitectural diseñado para mejorar sensibilidad posicional de redes convolucionales mediante descomposición de información espacial en atención horizontal y vertical separadas. La motivación teórica para este experimento surge del reconocimiento que localización precisa de *landmarks* requiere sensibilidad fina a posiciones absolutas en la imagen, aspecto que convoluciones estándar capturan solo implícitamente a través de receptive fields (campos receptivos) que agregan información espacial local sin codificación explícita de coordenadas globales.

El módulo *Coordinate Attention* se inserta entre Layer 4 del *backbone* ResNet-18 y la capa de *Global Average Pooling*, procesando el mapa de características $\mathbf{F} \in \mathbb{R}^{7 \times 7 \times 512}$ mediante atención selectiva que amplifica características en posiciones espaciales informativas mientras suprime características en posiciones irrelevantes. El módulo implementa tres operaciones secuenciales:

Pooling Direccional: El mapa de características se agrega separadamente a lo largo de dimensiones horizontal y vertical:

$$z_h^c(i) = \frac{1}{W} \sum_{j=0}^{W-1} F^c(i, j) \quad (\text{pooling horizontal, preserva altura}) \quad (3.3.14)$$

$$z_w^c(j) = \frac{1}{H} \sum_{i=0}^{H-1} F^c(i, j) \quad (\text{pooling vertical, preserva anchura}) \quad (3.3.15)$$

donde $H = 7$, $W = 7$ son dimensiones espaciales, c indexa canales, y las salidas $z_h \in \mathbb{R}^{H \times C}$ y $z_w \in \mathbb{R}^{W \times C}$ codifican perfiles de activación promediados a lo largo de cada fila y columna respectivamente. Esta descomposición direccional captura información posicional sin colapsar completamente estructura espacial como hace *Global Average Pooling* estándar.

Codificación Compartida: Los perfiles direccionales se concatenan y procesan mediante

convolución 1D compartida seguida de activación:

$$\mathbf{f} = \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}([\mathbf{z}_h; \mathbf{z}_w]))) \quad (3.3.16)$$

donde $[\mathbf{z}_h; \mathbf{z}_w]$ denota concatenación y la convolución 1×1 reduce canales de 512 a $512/32 = 16$ mediante factor de reducción $r = 32$, implementando cuello de botella que fuerza compresión de información posicional en representación compacta.

Generación de Atención: La representación compartida se divide y procesa mediante convoluciones separadas para generar mapas de atención direccionales:

$$\mathbf{a}_h = \sigma(\text{Conv}_{1 \times 1}(\mathbf{f}_h)) \quad \text{donde } \mathbf{a}_h \in \mathbb{R}^{H \times C} \quad (3.3.17)$$

$$\mathbf{a}_w = \sigma(\text{Conv}_{1 \times 1}(\mathbf{f}_w)) \quad \text{donde } \mathbf{a}_w \in \mathbb{R}^{W \times C} \quad (3.3.18)$$

donde σ es sigmoide que normaliza atención a $(0, 1)$. Los mapas de atención se aplican multiplicativamente al mapa de características original:

$$\mathbf{F}'(i, j, c) = \mathbf{F}(i, j, c) \times \mathbf{a}_h(i, c) \times \mathbf{a}_w(j, c) \quad (3.3.19)$$

El modelo con *Coordinate Attention* fue entrenado siguiendo protocolo idéntico a Fase 2 (70 épocas, tasas de aprendizaje diferenciadas, *Wing Loss*), pero los resultados experimentales demostraron que la complejidad arquitectural adicional no proporcionó beneficio medible en métricas de localización. El análisis de estos resultados negativos, detallado en el Capítulo ??, sugiere que para tarea de regresión de coordenadas con *dataset* de tamaño moderado, la capacidad representacional de ResNet-18 estándar es suficiente y adición de mecanismos de atención introduce riesgo de sobreajuste que contrarresta potenciales beneficios de sensibilidad posicional mejorada. Esta observación es consistente con principio general de parsimonia arquitectural: complejidad adicional solo beneficia cuando capacidad base es insuficiente y datos de entrenamiento son abundantes, condiciones no satisfechas en este trabajo.

La siguiente sección describe el *pipeline* completo de procesamiento de datos que transforma radiografías crudas y coordenadas anotadas en tensores normalizados compatibles con la arquitectura descrita.

3.4. Pipeline de preprocesamiento y aumentación de datos

La Sección 3.3 especificó la arquitectura neuronal implementada: ResNet-18 preentrenada con módulo de regresión especializado, diseñada para procesar tensores normalizados de dimensión $224 \times 224 \times 3$ con estadísticas específicas de ImageNet. La presente sección describe exhaustivamente el *pipeline* (secuencia de procesamiento) completo que transforma radiografías de tórax en formato digital crudo desde su adquisición clínica hasta tensores adecuadamente normalizados para inferencia mediante la red neuronal, preservando simultáneamente las correspondencias geométricas precisas entre coordenadas de *landmarks* (puntos de referencia anatómicos) en espacio de imagen original y espacio de entrada de la red. Este *pipeline* constituye componente crítico de la metodología: transformaciones geométricas incorrectas o normalizaciones inapropiadas comprometerían irremediablemente la capacidad del modelo para localizar estructuras anatómicas con precisión sub-píxel, independientemente de la sofisticación arquitectural o estrategia de entrenamiento implementadas.

El diseño del *pipeline* de datos enfrenta múltiples restricciones simultáneas que deben reconciliarse cuidadosamente. Primero, compatibilidad con arquitecturas preentrenadas en ImageNet: ResNet-18 estándar fue entrenada sobre $1,3 \times 10^6$ imágenes RGB naturales normalizadas con estadísticas específicas (medias $\mu = [0,485, 0,456, 0,406]$ y desviaciones estándar $\sigma = [0,229, 0,224, 0,225]$ por canal), requiriendo conversión de radiografías monocromáticas a representación pseudocromática y normalización coherente con distribución de activaciones esperada por capas convolucionales iniciales [4, 12]. Segundo, preservación de precisión geométrica: cada transformación espacial aplicada a la imagen (redimensionamiento, rotación, reflexión) debe replicarse matemáticamente sobre coordenadas de *landmarks* mediante transformaciones afines inversas correctamente parametrizadas, garantizando correspondencia exacta entre píxeles de entrada y coordenadas de supervisión durante entrenamiento. Tercero, aumentación de variabilidad sin corrupción anatómica: transformaciones de *data augmentation* (aumentación de datos) deben incrementar robustez del modelo ante variabilidad clínica realista (posicionamiento del paciente, diferencias en técnica radiográfica) sin generar configuraciones anatómicamente imposibles que confundirían el aprendizaje de restricciones geométricas. La estrategia implementada, documentada en esta sección, balancea estas restricciones mediante secuencia cuidadosamente ordenada de transformaciones determinísticas y estocásticas [88].

El *pipeline* se estructura en dos etapas funcionalmente distintas. La etapa de preprocesamiento determinístico aplica transformaciones idénticas a todas las muestras tanto en entrenamiento como en inferencia: conversión de espacio de color, redimensionamiento estandarizado, y normalización según estadísticas de ImageNet. Esta etapa garantiza compatibilidad con

representaciones preentrenadas y homogeneidad de entrada. La etapa de aumentación estocástica se aplica exclusivamente durante entrenamiento, introduciendo variabilidad controlada mediante transformaciones geométricas (reflexión horizontal, rotación limitada) y fotométricas (ajustes de brillo y contraste) aplicadas aleatoriamente con probabilidades calibradas. Esta separación permite reproducibilidad perfecta en evaluación y validación mientras maximiza variabilidad durante aprendizaje, siguiendo principios establecidos de regularización mediante transformaciones de datos [12, 88].

3.4.1. Preprocesamiento determinístico

El preprocesamiento determinístico transforma radiografías crudas desde formato de adquisición clínica a representación normalizada esperada por ResNet-18, mediante secuencia de tres operaciones aplicadas consistentemente a cada muestra.

Conversión de espacio de color

Las radiografías digitales de tórax en el conjunto de datos descrito en la Sección 3.2 se almacenan como imágenes monocromáticas de un solo canal de intensidad, codificando información de transmisión de rayos X en escala de grises de 8 bits ($I \in [0, 255]$). Sin embargo, ResNet-18 preentrenada en ImageNet espera tensores tricromáticos de entrada con tres canales RGB $(R, G, B) \in \mathbb{R}^{224 \times 224 \times 3}$, reflejando la naturaleza de las imágenes fotográficas naturales sobre las cuales fue entrenada. Esta incompatibilidad dimensional requiere conversión explícita del espacio de color monocromático al espacio pseudocromático RGB.

La estrategia de conversión implementada replica el canal de intensidad monocromática a través de los tres canales RGB, generando imagen pseudocromática acromática:

$$\begin{aligned} R(i, j) &= I(i, j), \\ G(i, j) &= I(i, j), \\ B(i, j) &= I(i, j), \end{aligned} \tag{3.4.1}$$

donde $I(i, j)$ denota la intensidad del píxel en posición (i, j) de la radiografía monocromática original, y $R(i, j)$, $G(i, j)$, $B(i, j)$ representan los valores de los canales rojo, verde y azul en la representación tricromática resultante. Esta transformación preserva completamente la información radiográfica original (no introduce contenido espurio ni descarta información diagnóstica) mientras satisface la restricción dimensional de arquitecturas preentrenadas en imágenes naturales.

La justificación de esta estrategia simple de replicación de canal, en contraste con esquemas

más sofisticados de pseudocolorización basados en *colormaps* (mapas de color especializados para visualización médica), se fundamenta en dos consideraciones. Primero, compatibilidad con normalización de ImageNet: la replicación uniforme garantiza que, tras normalización con estadísticas de ImageNet, las activaciones en capas convolucionales iniciales permanezcan dentro del régimen de valores para el cual los filtros preentrenados fueron optimizados, facilitando *transfer learning* (aprendizaje por transferencia) efectivo [4]. Segundo, preservación de linealidad radiométrica: la relación lineal entre intensidad de píxel y atenuación de rayos X, fundamental para interpretación radiográfica, se mantiene sin distorsión no lineal que introduciría *colormaps* complejos. Estudios empíricos sobre *transfer learning* en dominio médico validan esta estrategia, demostrando que replicación simple de canal produce resultados comparables o superiores a esquemas de pseudocolorización elaborados cuando se combina con normalización apropiada [22].

La conversión de espacio de color se implementa mediante la función `cv2.cvtColor` de OpenCV [81] con parámetro `cv2.COLOR_GRAY2RGB`, ejecutada inmediatamente tras carga de imagen desde disco. Esta operación tiene costo computacional despreciable ($\mathcal{O}(N)$ donde $N = 299 \times 299$ es el número de píxeles) y complejidad de memoria $3\times$ respecto a imagen monocromática original.

Redimensionamiento y transformación de coordenadas

Las radiografías en el conjunto de datos poseen resolución espacial uniforme de 299×299 píxeles (Tabla 3.2.1 en Sección 3.2), mientras que ResNet-18 estándar procesa entradas de dimensión 224×224 píxeles, resolución establecida como estándar en competencias ImageNet [12]. Esta discrepancia dimensional requiere operación de redimensionamiento que reduce resolución espacial mediante interpolación, acompañada de transformación compensatoria de coordenadas de *landmarks* que preserva correspondencias geométricas.

El redimensionamiento de imagen se realiza mediante interpolación bilineal, aplicando transformación de escalamiento uniforme isotrópico:

$$\mathbf{I}'(i', j') = \text{BilinearInterpolation} \left(\mathbf{I}, \frac{i' \cdot W_{orig}}{W_{target}}, \frac{j' \cdot H_{orig}}{H_{target}} \right), \quad (3.4.2)$$

donde \mathbf{I} denota la imagen RGB de dimensión $W_{orig} \times H_{orig} = 299 \times 299$, \mathbf{I}' es la imagen redimensionada de dimensión $W_{target} \times H_{target} = 224 \times 224$, e $(i', j') \in [0, W_{target}) \times [0, H_{target})$ son coordenadas en espacio de imagen de salida. La interpolación bilineal fue seleccionada por su balance óptimo entre calidad de reconstrucción (superior a interpolación por vecino más cercano) y eficiencia computacional (superior a interpolación bicúbica), siendo estándar en *pipelines* de visión computacional [81].

Dado que el problema formulado es regresión directa de coordenadas normalizadas $(x_k, y_k) \in [0, 1]^2$ para cada *landmark* $k \in \{1, \dots, 15\}$, expresadas como fracciones relativas a dimensiones de imagen, el redimensionamiento espacial no requiere transformación explícita de coordenadas de supervisión. Las coordenadas normalizadas son invariantes ante cambios de escala uniforme:

$$\left(\frac{x_{pixel}}{W}, \frac{y_{pixel}}{H} \right) = \left(\frac{x_{pixel} \cdot s}{W \cdot s}, \frac{y_{pixel} \cdot s}{H \cdot s} \right) \quad \forall s > 0, \quad (3.4.3)$$

donde $s = W_{target}/W_{orig} = 224/299 \approx 0,749$ es el factor de escalamiento. Esta propiedad constituye ventaja significativa de la formulación de regresión de coordenadas normalizadas respecto a regresión de coordenadas absolutas en píxeles, eliminando necesidad de transformaciones compensatorias complejas ante variaciones en resolución de entrada. La normalización de coordenadas a rango $[0, 1]$ fue implementada durante anotación del conjunto de datos (Sección 3.2), permitiendo compatibilidad inmediata con múltiples resoluciones de procesamiento.

El redimensionamiento se implementa mediante la función `cv2.resize` de OpenCV con parámetro de interpolación `cv2.INTER_LINEAR`, aplicada tras conversión de espacio de color. La reducción de resolución de $299^2 = 89401$ píxeles a $224^2 = 50176$ píxeles (reducción del 44 %) disminuye sustancialmente demanda computacional de capas convolucionales subsecuentes sin degradación observable de capacidad de localización, dado que la resolución efectiva para tareas de detección de *landmarks* anatómicos en imágenes de 224×224 permanece suficiente para precisión sub-píxel cuando se combinan representaciones jerárquicas profundas con funciones de pérdida especializadas [15].

Normalización según estadísticas de ImageNet

La normalización de intensidades de píxeles constituye componente crítico del *pipeline* de preprocesamiento, transformando valores de píxeles RGB desde su rango original $[0, 255]$ (enteros de 8 bits) a distribución centrada y escalada compatible con estadísticas de activación aprendidas por capas convolucionales de ResNet-18 durante preentrenamiento en ImageNet [12].

El procedimiento de normalización se realiza en dos etapas secuenciales. Primero, conversión a rango de punto flotante $[0, 1]$ mediante división por 255:

$$\tilde{\mathbf{I}}(i, j, c) = \frac{\mathbf{I}'(i, j, c)}{255}, \quad c \in \{R, G, B\}, \quad (3.4.4)$$

donde \mathbf{I}' denota la imagen redimensionada con valores enteros en $[0, 255]$, y $\tilde{\mathbf{I}}$ representa la imagen en formato de punto flotante. Segundo, estandarización canal-específica mediante

sustracción de media y división por desviación estándar de ImageNet:

$$\mathbf{I}_{norm}(i, j, c) = \frac{\tilde{\mathbf{I}}(i, j, c) - \mu_c}{\sigma_c}, \quad (3.4.5)$$

donde $\mu = [\mu_R, \mu_G, \mu_B] = [0,485, 0,456, 0,406]$ y $\sigma = [\sigma_R, \sigma_G, \sigma_B] = [0,229, 0,224, 0,225]$ son las medias y desviaciones estándar computadas sobre el conjunto de entrenamiento de ImageNet ILSVRC-2012 [12], valores estándar utilizados universalmente en *transfer learning* con arquitecturas preentrenadas en ImageNet.

Esta normalización canal-específica garantiza que, para cada canal c , la distribución de activaciones de entrada posea media aproximadamente cero y varianza unitaria cuando se promedian sobre el conjunto de datos. Aunque las radiografías pseudocromáticas producidas por replicación de canal (Ecuación 3.4.1) poseen estadísticas idénticas en los tres canales ($\mu_R = \mu_G = \mu_B$ y $\sigma_R = \sigma_G = \sigma_B$ localmente para cada imagen individual), la aplicación de parámetros de normalización diferenciados por canal de ImageNet introduce asimetría deliberada que mejora compatibilidad con filtros convolucionales preentrenados. Estos filtros aprendieron patrones visuales sensibles a variaciones cromáticas específicas de imágenes naturales, y la normalización canal-específica preserva la estructura de covarianza entre canales que caracteriza el espacio de representación aprendido durante preentrenamiento [4].

La normalización se implementa mediante transformaciones de PyTorch: conversión inicial a tensor con `torch.from_numpy`, permutación de dimensiones desde formato OpenCV ($H \times W \times C$) a formato PyTorch ($C \times H \times W$) mediante `permute(2, 0, 1)`, conversión a punto flotante con `float()`, división por 255, y aplicación de normalización mediante `torchvision.transforms.Normalize` con medias y desviaciones estándar especificadas [80]. El tensor normalizado resultante $\mathbf{I}_{norm} \in \mathbb{R}^{3 \times 224 \times 224}$ constituye la entrada estándar al modelo durante entrenamiento e inferencia.

3.4.2. Aumentación estocástica de datos

La aumentación de datos mediante transformaciones estocásticas constituye técnica fundamental de regularización en aprendizaje profundo supervisado, particularmente crítica en dominio médico donde conjuntos de datos anotados son inherentemente limitados por el costo prohibitivo de anotación experta [88]. La estrategia implementada aplica transformaciones geométricas y fotométricas aleatorias durante entrenamiento, expandiendo artificialmente la diversidad del conjunto de datos de 956 muestras anotadas (Sección 3.2) mediante generación implícita de variantes transformadas de cada radiografía original.

El diseño del protocolo de aumentación enfrenta restricción fundamental impuesta por la naturaleza de la tarea: a diferencia de clasificación de imágenes donde transformaciones como

recortes aleatorios (*random crops*) y escalamientos no uniformes son admisibles, la tarea de regresión de *landmarks* requiere que cada transformación geométrica aplicada a la imagen se replique exactamente sobre las coordenadas de supervisión mediante transformación afín inversa matemáticamente consistente. Transformaciones que no preservan correspondencias geométricas (como recortes asimétricos que eliminan *landmarks* del campo de visión) corrompen irremediabilmente la supervisión, impidiendo aprendizaje. Esta restricción limita las transformaciones admisibles a aquellas geométricamente invertibles: reflexiones, rotaciones, traslaciones, y escalamientos uniformes cuyos parámetros son conocidos exactamente [88].

El protocolo implementado incorpora tres categorías de transformaciones estocásticas, aplicadas secuencialmente con probabilidades calibradas para balancear incremento de variabilidad contra preservación de realismo anatómico.

Reflexión horizontal

La reflexión horizontal constituye transformación de aumentación más frecuentemente aplicada (probabilidad $p_{flip} = 0,70$), explotando la simetría bilateral aproximada de la anatomía torácica humana. Una radiografía de tórax reflejada horizontalmente permanece anatómicamente plausible y diagnósticamente válida, representando simplemente una adquisición con orientación lateral invertida.

La transformación de reflexión horizontal se define matemáticamente como:

$$\mathbf{I}_{flip}(i, j) = \mathbf{I}(W - 1 - i, j), \quad (3.4.6)$$

donde $W = 224$ es el ancho de imagen, (i, j) son coordenadas en imagen original, y $(W - 1 - i, j)$ son coordenadas reflejadas respecto al eje vertical central. Las coordenadas normalizadas de *landmarks* se transforman mediante reflexión correspondiente:

$$x'_k = 1 - x_k, \quad y'_k = y_k, \quad k \in \{1, \dots, 15\}, \quad (3.4.7)$$

donde $(x_k, y_k) \in [0, 1]^2$ son coordenadas normalizadas originales del *landmark* k , y (x'_k, y'_k) son coordenadas transformadas tras reflexión.

Adicionalmente, la reflexión horizontal requiere intercambio de identidades entre *landmarks* que forman pares simétricos bilaterales. Como se definió en la Sección 3.2, el conjunto de datos incluye cinco pares de *landmarks* bilateralmente simétricos: $\mathcal{P}_{sym} = \{(2, 3), (4, 5), (6, 7), (11, 12), (13, 14)\}$, correspondientes a estructuras anatómicas emparejadas (ápices pulmonares, ángulos costofrénicos, hilios, etc.). Tras reflexión horizontal, la identidad de *landmarks* emparejados debe intercambiarse para

mantener consistencia anatómica:

$$(x'_i, y'_i) \leftrightarrow (x'_j, y'_j) \quad \forall (i, j) \in \mathcal{P}_{sym}. \quad (3.4.8)$$

La implementación de reflexión horizontal utiliza `torch.flip` con parámetro `dims=[2]` para invertir dimensión espacial horizontal del tensor de imagen, y permutación explícita de índices de coordenadas para intercambio de pares simétricos. La alta probabilidad de aplicación ($p = 0,70$) garantiza que el modelo observe tanto configuraciones anatómicas originales como reflejadas con frecuencia balanceada, promoviendo invariancia ante orientación lateral y mejorando capacidad de generalización a variabilidad de posicionamiento clínico.

Rotación aleatoria limitada

La rotación aleatoria dentro de rango angular limitado modela variabilidad en posicionamiento del paciente durante adquisición radiográfica, donde ligeras inclinaciones son inevitables en práctica clínica. La transformación se aplica con probabilidad $p_{rot} = 0,30$ (menos frecuente que reflexión para evitar exceso de transformaciones compuestas que degradarían calidad de imagen), muestreando ángulo de rotación uniformemente desde intervalo $\theta \sim \mathcal{U}(-15^\circ, +15^\circ)$.

La transformación de rotación centrada en el centro de imagen se define mediante matriz de rotación afín:

$$\mathbf{R}(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & t_x \\ \sin \theta & \cos \theta & t_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (3.4.9)$$

donde las componentes de traslación compensatoria t_x y t_y garantizan rotación centrada en punto $(W/2, H/2)$ de imagen. Las coordenadas normalizadas de *landmarks* se transforman mediante aplicación de rotación inversa centrada en $(0,5, 0,5)$, punto central en espacio de coordenadas normalizadas:

$$\begin{bmatrix} x'_k - 0,5 \\ y'_k - 0,5 \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_k - 0,5 \\ y_k - 0,5 \end{bmatrix}. \quad (3.4.10)$$

La limitación del rango angular a $\pm 15^\circ$ responde a dos consideraciones. Primero, realismo clínico: rotaciones superiores a 15° son infrecuentes en radiografías de tórax de calidad diagnóstica estándar, dado que protocolos de posicionamiento radiográfico buscan alineación precisa del paciente con el detector. Segundo, preservación de visibilidad de *landmarks*: rotaciones excesivas podrían desplazar *landmarks* periféricos (ápices pulmonares, ángulos

costofrénicos) fuera del campo de visión tras rotación, corrompiendo supervisión. El rango de $\pm 15^\circ$ balancea incremento de robustez ante variabilidad de posicionamiento con preservación de validez anatómica.

La implementación utiliza `torchvision.transforms.functional.affine` para aplicar transformación afín a imagen, y rotación matricial explícita sobre coordenadas normalizadas mediante operaciones tensoriales de PyTorch. La interpolación bilineal se emplea para reconstrucción de imagen rotada, con relleno de regiones externas mediante valor cero (*padding* negro), consistente con fondo típico de radiografías digitales.

Ajustes fotométricos

Los ajustes fotométricos modelan variabilidad en técnica radiográfica (variaciones en kilovoltaje pico, miliamperaje-segundo, tiempo de exposición) y procesamiento posterior (ajustes de ventana y nivel en sistemas PACS), que afectan contraste y brillo aparente de radiografías sin alterar geometría anatómica. Estas transformaciones se aplican con probabilidad $p_{photo} = 0,50$, ajustando brillo (mediante suma aditiva) y contraste (mediante multiplicación) dentro de rangos calibrados.

El ajuste de brillo se define como traslación aditiva uniforme de intensidades:

$$\mathbf{I}_{bright}(i, j, c) = \mathbf{I}(i, j, c) + \beta, \quad \beta \sim \mathcal{U}(-0,2, +0,2), \quad (3.4.11)$$

donde β es el factor de brillo muestreado uniformemente desde intervalo $[-0,2, +0,2]$ en espacio normalizado $[0, 1]$. El ajuste de contraste se implementa como escalamiento multiplicativo centrado en intensidad media:

$$\mathbf{I}_{contrast}(i, j, c) = \alpha \cdot \mathbf{I}(i, j, c), \quad \alpha \sim \mathcal{U}(0,8, 1,2), \quad (3.4.12)$$

donde α es el factor de contraste muestreado desde intervalo $[0,8, 1,2]$, permitiendo reducción o incremento del 20 % en contraste aparente.

Crucialmente, los ajustes fotométricos no requieren transformación de coordenadas de *landmarks*, dado que preservan completamente la geometría espacial de imagen: ningún píxel cambia de posición, solo su intensidad. Esta propiedad contrasta con transformaciones geométricas (reflexión, rotación) que requieren compensación de coordenadas. Los ajustes fotométricos se aplican tras normalización de ImageNet descrita en la Sección 3.4.1, operando sobre tensores normalizados antes de ingreso a la red neuronal.

La implementación utiliza `torchvision.transforms.ColorJitter` con parámetros `brightness=0.2` y `contrast=0.2`, aplicados con probabilidad 0.5 mediante envoltura en

RandomApply. Los valores de brillo y contraste tras ajuste se recortan (*clipping*) al rango válido de tensores normalizados para evitar valores atípicos extremos que desestabilizarían gradientes durante retropropagación [80].

3.4.3. Orden de aplicación y composición de transformaciones

Las transformaciones de preprocesamiento y aumentación descritas en las subsecciones previas se aplican mediante *pipeline* secuencial implementado como composición de funciones en PyTorch [80]. El orden de aplicación es crítico: transformaciones no conmutan, y secuencias incorrectas producirían inconsistencias entre imágenes procesadas y coordenadas transformadas.

El *pipeline* completo de entrenamiento aplica transformaciones en el siguiente orden estrictamente especificado:

1. Carga de radiografía monocromática desde disco (formato PNG de 8 bits).
2. Conversión de espacio de color: monocromático \rightarrow RGB pseudocromático (Ecuación 3.4.1).
3. Redimensionamiento mediante interpolación bilineal: $299 \times 299 \rightarrow 224 \times 224$ (Ecuación 3.4.2).
4. Conversión a tensor PyTorch con permutación de dimensiones.
5. Normalización a rango $[0, 1]$ mediante división por 255.
6. Normalización según estadísticas de ImageNet (Ecuación 3.4.5).
7. Aplicación estocástica de reflexión horizontal con $p = 0,70$ (Ecuaciones 3.4.6–3.4.8).
8. Aplicación estocástica de rotación con $p = 0,30$, ángulo $\theta \sim \mathcal{U}(-15^\circ, +15^\circ)$ (Ecuaciones 3.4.9–3.4.10).
9. Aplicación estocástica de ajustes fotométricos con $p = 0,50$ (Ecuaciones 3.4.11–3.4.12).

Durante inferencia (validación y evaluación en conjunto de prueba), únicamente los pasos determinísticos (1–6) se aplican, omitiendo completamente transformaciones estocásticas de aumentación. Esta separación garantiza reproducibilidad perfecta de predicciones durante evaluación, requisito fundamental para comparación rigurosa de rendimiento entre modelos y reportes de métricas estandarizadas.

La composición de transformaciones estocásticas geométricas (reflexión y rotación) puede aplicarse simultáneamente con probabilidades independientes, produciendo ocasionalmente muestras transformadas mediante ambas operaciones. La probabilidad de aplicación conjunta

es $p_{flip} \cdot p_{rot} = 0,70 \times 0,30 = 0,21$ (21 % de muestras), mientras que la probabilidad de al menos una transformación geométrica es $1 - (1 - p_{flip})(1 - p_{rot}) = 1 - 0,30 \times 0,70 = 0,79$ (79 % de muestras). Esta composición estocástica expande significativamente la diversidad efectiva del conjunto de entrenamiento: cada radiografía original de 956 disponibles puede presentarse en múltiples configuraciones transformadas a lo largo de las épocas de entrenamiento, reduciendo sobreajuste mediante exposición continua a variantes no idénticas [88].

3.4.4. Síntesis del pipeline

El *pipeline* de preprocesamiento y aumentación documentado en esta sección reconcilia exitosamente los requerimientos simultáneos de compatibilidad con arquitecturas preentrenadas (mediante normalización de ImageNet), preservación de precisión geométrica (mediante transformaciones afines matemáticamente consistentes sobre coordenadas de *landmarks*), e incremento de robustez ante variabilidad clínica (mediante aumentación estocástica controlada). La separación funcional entre preprocesamiento determinístico (aplicado uniformemente en entrenamiento e inferencia) y aumentación estocástica (exclusiva de entrenamiento) garantiza reproducibilidad en evaluación mientras maximiza regularización durante aprendizaje, siguiendo principios establecidos de diseño de *pipelines* de visión computacional [12, 88].

La implementación completa del *pipeline* se encapsula en clase personalizada `ChestXrayDataset` derivada de `torch.utils.data.Dataset`, que gestiona carga de imágenes, aplicación de transformaciones, y generación de pares (imagen transformada, coordenadas transformadas) durante iteración de entrenamiento. Esta abstracción modular facilita experimentación con variaciones del protocolo de aumentación y garantiza consistencia de procesamiento a través de todas las fases de entrenamiento descritas en la Sección 3.5.

La siguiente sección describe exhaustivamente la estrategia de entrenamiento progresivo en cuatro fases que incorpora gradualmente restricciones geométricas inspiradas en conocimiento anatómico, construyendo sobre la arquitectura especificada en la Sección 3.3 y operando sobre datos procesados mediante el *pipeline* documentado en la presente sección.

3.5. Estrategia de Entrenamiento Progresivo

La incorporación de conocimiento anatómico mediante restricciones geométricas diferenciables constituye un paradigma promisorio para mejorar consistencia estructural de predicciones de *landmarks* en imágenes médicas. La estrategia de entrenamiento desarrollada en este trabajo implementa este paradigma mediante un protocolo progresivo en cuatro fases que incorpora gradualmente funciones de pérdida geométricamente restringidas, comenzando con optimización estándar mediante Error Cuadrático Medio (*Mean Squared Error*, MSE) para establecer una línea base, transitando a *Wing Loss* para mejorar precisión sub-píxel, agregando *Symmetry Loss* para imponer consistencia bilateral, y finalmente incorporando *Distance Preservation Loss* para garantizar proporciones anatómicas válidas. Cada fase se construye sobre la anterior mediante inicialización con pesos óptimos de la fase previa, estrategia de *warm-start* (inicio cálido) que acelera convergencia y previene degradación de desempeño al introducir términos de pérdida adicionales.

Esta organización en fases progresivas, en lugar de entrenamiento directo con la función de pérdida completa desde el inicio, se fundamenta en observaciones empíricas previas sobre dificultad de optimización de funciones de pérdida multi-objetivo complejas: el entrenamiento simultáneo con múltiples términos de pérdida geométrica desde inicialización aleatoria frecuentemente resulta en inestabilidad numérica, convergencia prematura a mínimos locales de calidad inferior, o dificultad en balancear magnitudes relativas de gradientes provenientes de diferentes términos. La incorporación gradual permite al modelo primero establecer predicciones aproximadamente correctas mediante supervisión MSE estándar, luego refinar precisión mediante *Wing Loss* que proporciona gradientes más informativos en régimen de error pequeño, posteriormente mejorar consistencia geométrica mediante *Symmetry Loss*, y finalmente incorporar restricciones de proporciones anatómicas mediante *Distance Preservation Loss*, secuencia que guía la optimización a través de paisaje de pérdida complejo de manera controlada.

3.5.1. Fase 1: Entrenamiento del Módulo de Regresión con Backbone Congelado

La primera fase implementa el protocolo estándar de *transfer learning* en dos etapas: congelar completamente los pesos del *backbone* preentrenado y entrenar únicamente el módulo de regresión añadido, permitiendo que las capas superiores aprendan a mapear representaciones visuales de ImageNet a coordenadas de *landmarks* anatómicos sin perturbar las características de bajo nivel ya aprendidas. Esta estrategia conservadora es particularmente apropiada cuando

el *dataset* objetivo es pequeño (< 1000 imágenes) y el riesgo de sobreajuste es alto: entrenar todos los 11.6 millones de parámetros desde inicialización aleatoria con solo 669 imágenes de entrenamiento resultaría inevitablemente en memorización de datos de entrenamiento sin capacidad de generalización.

El modelo de Fase 1 se define formalmente como $f_\theta = h_\phi \circ g_\psi$, donde $g_\psi : \mathbb{R}^{224 \times 224 \times 3} \rightarrow \mathbb{R}^{512}$ representa el *backbone* ResNet-18 que mapea imágenes a vectores de características de 512 dimensiones con parámetros ψ inicializados desde ImageNet y mantenidos fijos ($\nabla_\psi \mathcal{L} = 0$ forzado), y $h_\phi : \mathbb{R}^{512} \rightarrow \mathbb{R}^{30}$ representa el módulo de regresión de tres capas completamente conectadas con parámetros ϕ inicializados aleatoriamente mediante inicialización Kaiming [89], esquema que escala pesos iniciales según número de conexiones para garantizar estabilidad numérica durante propagación hacia adelante y retropropagación de gradientes.

Configuración de Fase 1:

- **Función de pérdida:** MSE estándar sobre coordenadas normalizadas

$$\mathcal{L}_{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{30} \sum_{i=1}^{30} (\hat{y}_i - y_i)^2 \quad (3.5.1)$$

donde $\hat{\mathbf{y}} \in [0, 1]^{30}$ son coordenadas predichas (salida Sigmoid) y $\mathbf{y} \in [0, 1]^{30}$ son coordenadas *ground truth* normalizadas.

- **Optimizador:** Adam [90] con parámetros estándar ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$)
- **Tasa de aprendizaje:** $\alpha = 1 \times 10^{-3}$ (constante durante toda la fase)
- **Regularización L2:** *Weight decay* $\lambda = 1 \times 10^{-4}$ aplicado solo a parámetros ϕ del módulo de regresión
- **Tamaño de batch:** 32 imágenes (compromiso entre eficiencia computacional y estabilidad de gradientes)
- **Número de épocas:** 15 (suficiente para convergencia del módulo de regresión sin sobreajuste)
- **Parámetros entrenables:** $|\phi| \approx 400,000$ (solo módulo de regresión), correspondiente al 3.4 % del total

El entrenamiento de Fase 1 emplea MSE como función de pérdida inicial por simplicidad y estabilidad: MSE proporciona gradientes bien comportados sin singularidades, superficie de pérdida convexa localmente que facilita convergencia desde inicialización aleatoria, y minimización directa de discrepancia euclidiana entre predicciones y *ground truth*. Aunque MSE presenta limitaciones conocidas para localización sub-píxel (penalización

uniforme independiente de magnitud de error, sesgo hacia promedio de distribución en presencia de outliers), estas desventajas son menos críticas en fase inicial donde objetivo es establecer aproximación razonable al mapeo imágenes→coordenadas antes de refinamientos posteriores.

El protocolo específico de Fase 1 procede mediante iteración sobre *mini-batches* de 32 imágenes extraídos aleatoriamente del conjunto de entrenamiento, computando predicciones mediante propagación hacia adelante a través del modelo $\hat{\mathbf{y}}^{(b)} = f_{\theta}(\mathbf{X}^{(b)})$ donde $\mathbf{X}^{(b)} \in \mathbb{R}^{32 \times 3 \times 224 \times 224}$ es el *batch* de imágenes preprocesadas, calculando pérdida MSE promediada sobre el *batch* $\mathcal{L}^{(b)} = \frac{1}{32} \sum_{i=1}^{32} \mathcal{L}_{MSE}(\hat{\mathbf{y}}_i^{(b)}, \mathbf{y}_i^{(b)})$, computando gradientes mediante retropropagación $\nabla_{\phi} \mathcal{L}^{(b)}$ (solo respecto a parámetros del módulo de regresión, gradientes del *backbone* son descartados), y actualizando parámetros mediante regla de Adam que combina momentum de primer y segundo orden para convergencia estable. Después de procesar todos los *mini-batches* del conjunto de entrenamiento (una época), el modelo se evalúa sobre el conjunto de validación para monitoreo de convergencia, guardando el *checkpoint* (punto de control) del modelo con menor pérdida de validación observada hasta el momento.

El entrenamiento de Fase 1 típicamente completa en aproximadamente 1 minuto en la configuración de *hardware* empleada (GPU AMD RX 6600), tiempo reducido explicado por el pequeño número de parámetros entrenables (400K vs 11.6M totales) y épocas limitadas (15), suficientes para convergencia del módulo de regresión sin necesitar ajuste fino extenso del *backbone*. El modelo resultante establece una línea base funcional que captura correspondencia aproximada entre apariencia visual de radiografías y posiciones de *landmarks*, aunque con precisión limitada debido a las limitaciones inherentes de MSE para localización sub-píxel, aspecto abordado en fases subsiguientes mediante funciones de pérdida especializadas.

3.5.2. Fase 2: Fine-Tuning Completo con Wing Loss

La segunda fase desbloquea todos los parámetros del modelo, permitiendo optimización de la arquitectura completa mediante *fine-tuning* que adapta representaciones visuales preentrenadas en ImageNet a características específicas de radiografías de tórax, simultáneamente introduciendo *Wing Loss* como función de pérdida especializada para localización sub-píxel de *landmarks*. *Wing Loss* [15] proporciona gradientes más informativos que MSE en régimen de error pequeño mediante transición suave entre comportamiento logarítmico cerca de error cero (gradiente grande, aceleración de convergencia final) y comportamiento lineal para errores grandes (robustez ante *outliers*), característica demostrada empíricamente en detección de *landmarks* faciales y extendida exitosamente a dominio médico en trabajos recientes.

El modelo de Fase 2 mantiene la arquitectura idéntica a Fase 1 ($f_{\theta} = h_{\phi} \circ g_{\psi}$), pero todos los parámetros $\theta = \{\psi, \phi\}$ son ahora entrenables con tasas de aprendizaje diferenciadas: el

backbone ψ recibe tasa de aprendizaje reducida para preservar parcialmente conocimiento de ImageNet, mientras el módulo de regresión ϕ mantiene tasa de aprendizaje estándar para adaptación rápida. Esta estrategia de tasas de aprendizaje diferenciadas implementa el principio de *discriminative learning rates* que reconoce que capas inferiores (características genéricas de bajo nivel) requieren ajuste mínimo, mientras capas superiores (características específicas de tarea) necesitan adaptación sustancial.

Configuración de Fase 2:

- **Función de pérdida:** *Wing Loss* con parámetros $\omega = 10,0$, $\epsilon = 2,0$

$$\mathcal{L}_{wing}(x) = \begin{cases} \omega \times \ln\left(1 + \frac{|x|}{\epsilon}\right) & \text{si } |x| < \omega \\ |x| - C & \text{si } |x| \geq \omega \end{cases} \quad (3.5.2)$$

donde $x = \hat{y}_i - y_i$ es el error de predicción por coordenada, $\omega = 10,0$ es el umbral de transición (expresado en escala normalizada $[0, 1]$, equivalente a $\approx 2,24$ píxeles en imagen de 224×224), $\epsilon = 2,0$ controla curvatura en régimen logarítmico, y $C = \omega - \omega \ln(1 + \omega/\epsilon) \approx 3,906$ es constante de continuidad. La pérdida total se promedia sobre las 30 coordenadas:

$$\mathcal{L}_{total}^{(P2)} = \frac{1}{30} \sum_{i=1}^{30} \mathcal{L}_{wing}(\hat{y}_i - y_i) \quad (3.5.3)$$

- **Optimizador:** Adam con grupos de parámetros separados
 - Parámetros *backbone* ψ : $\alpha_{back} = 2 \times 10^{-5}$ (tasa reducida 50×)
 - Parámetros módulo regresión ϕ : $\alpha_{head} = 2 \times 10^{-4}$ (tasa estándar)
- **Regularización L2:** *Weight decay* $\lambda = 5 \times 10^{-5}$ (reducido vs Fase 1 para permitir mayor flexibilidad)
- **Scheduler de tasa de aprendizaje:** CosineAnnealingLR [91] con período $T_{max} = 70$ épocas y tasa mínima $\eta_{min} = 2 \times 10^{-6}$, implementando decaimiento suave según

$$\alpha_t = \eta_{min} + \frac{1}{2}(\alpha_0 - \eta_{min}) \left(1 + \cos\left(\frac{t\pi}{T_{max}}\right)\right) \quad (3.5.4)$$

donde t es el número de época actual y α_0 es la tasa de aprendizaje inicial (α_{back} o α_{head} según grupo de parámetros). Este *scheduler* proporciona decaimiento gradual que facilita convergencia a mínimos de alta calidad.

- **Tamaño de batch:** 8 imágenes (reducido vs Fase 1 para estabilidad al optimizar 11.6M parámetros)

- **Número de épocas:** 70 (entrenamiento extenso para convergencia completa)
- **Early stopping:** Paciencia de 15 épocas sin mejora en pérdida de validación, deteniendo entrenamiento anticipadamente si el modelo deja de mejorar
- **Inicialización:** Pesos $\theta_{P2}^{init} = \theta_{P1}^{best}$ (*warm-start* desde mejor *checkpoint* de Fase 1)
- **Parámetros entrenables:** $|\theta| = 11,578,206$ (arquitectura completa)

La selección de *Wing Loss* sobre alternativas como L1 o Smooth L1 se fundamenta en su comportamiento de gradiente adaptativo: para errores pequeños ($|x| < \omega$), el gradiente $\partial \mathcal{L}_{wing} / \partial x = \frac{\omega}{\epsilon + |x|} \cdot \text{sign}(x)$ escala inversamente con error, proporcionando fuerza mayor cuando la predicción está muy cerca del *ground truth*, acelerando convergencia final a precisión sub-píxel. Para errores grandes ($|x| \geq \omega$), el gradiente se satura a $\text{sign}(x)$, proporcionando robustez ante *outliers* similar a L1. Los valores de hiperparámetros $\omega = 10,0$ y $\epsilon = 2,0$ fueron establecidos por Feng et al. [15] basándose en experimentos extensos en detección facial y adoptados aquí sin modificación, constituyendo configuración estándar en literatura de localización de *landmarks*.

El protocolo de Fase 2 implementa entrenamiento estándar con dos grupos de parámetros, donde el optimizador Adam mantiene estadísticas de momentum separadas para cada grupo y aplica tasas de aprendizaje diferenciadas. La reducción de tamaño de *batch* de 32 a 8 es necesaria por limitaciones de memoria GPU (8GB VRAM) al propagar gradientes a través de toda la arquitectura ResNet-18, aunque *batch* de 8 mantiene estimación de gradiente suficientemente estable mediante acumulación de estadísticas de momentum de Adam. El *scheduler* CosineAnnealingLR proporciona decaimiento suave de tasa de aprendizaje que evita oscilaciones en fases finales de entrenamiento, transicionando gradualmente de exploración con tasa alta a refinamiento con tasa baja.

La estrategia de *early stopping* monitorea pérdida de validación después de cada época, manteniendo registro del mejor valor observado y contador de épocas sin mejora. Cuando el contador alcanza paciencia de 15 épocas (aproximadamente 20 % del máximo de 70 épocas), el entrenamiento se detiene anticipadamente, previniendo sobreajuste prolongado al conjunto de entrenamiento. El modelo final de Fase 2 corresponde al *checkpoint* con menor pérdida de validación, no al modelo de la última época, implementando principio de selección de modelo basada en desempeño de generalización en lugar de ajuste a entrenamiento.

3.5.3. Fase 3: Incorporación de Symmetry Loss para Consistencia Bilateral

La tercera fase introduce restricciones de simetría bilateral mediante *Symmetry Loss*, función de pérdida geométrica que penaliza inconsistencias entre posiciones de *landmarks* pareados a través del eje mediastínico. Como se fundamentó en la Sección 3.2.4, la simetría bilateral es una invariante anatómica de la estructura torácica que puede explotarse como supervisión adicional: pares de *landmarks* correspondientes a estructuras izquierda-derecha (ápices, hila, bases pulmonares) deben presentar reflexión aproximada respecto al eje vertical definido por el mediastino, restricción que proporciona señal de aprendizaje complementaria a la supervisión de coordenadas punto-a-punto.

La función *Symmetry Loss* implementada compara posiciones predichas de pares simétricos con sus reflexiones esperadas, calculando discrepancia mediante distancia euclidiana. Para cada par $(i, j) \in \mathcal{P}_{sym}$ definido en Ecuación 3.2.1, se computa la posición esperada del *landmark* derecho j como reflexión del *landmark* izquierdo i a través del eje mediastínico $x = x_{axis}$ dado por Ecuación 3.2.2, y se penaliza desviación de esta predicción. La pérdida se formula bidireccionalmente (izquierda→derecha y derecha→izquierda) para tratar ambos lados simétricamente:

$$\mathcal{L}_{symmetry}(\hat{\mathbf{y}}) = \frac{1}{2|\mathcal{P}_{sym}|} \sum_{(i,j) \in \mathcal{P}_{sym}} \left[\|\hat{\mathbf{p}}_j - \text{Mirror}(\hat{\mathbf{p}}_i, x_{axis})\|_2 + \|\hat{\mathbf{p}}_i - \text{Mirror}(\hat{\mathbf{p}}_j, x_{axis})\|_2 \right] \quad (3.5.5)$$

donde $\hat{\mathbf{p}}_k = [\hat{x}_k, \hat{y}_k]^T$ representa las coordenadas 2D predichas del *landmark* k , $|\mathcal{P}_{sym}| = 5$ es el número de pares simétricos, y la operación de reflexión especular se define como:

$$\text{Mirror}(\mathbf{p}, x_{axis}) = \begin{bmatrix} 2x_{axis} - p_x \\ p_y \end{bmatrix} \quad (3.5.6)$$

reflejando la coordenada horizontal a través de $x = x_{axis}$ mientras preservando la coordenada vertical. El eje x_{axis} se calcula dinámicamente para cada predicción usando Ecuación 3.2.2 con las coordenadas predichas de *landmarks* mediastínicos, permitiendo que el eje de simetría se adapte a la imagen específica en lugar de asumir eje fijo en el centro de la imagen, lo cual sería inadecuado para radiografías con rotación o descentrado del paciente.

La función de pérdida total de Fase 3 combina *Wing Loss* con *Symmetry Loss* mediante suma

ponderada:

$$\mathcal{L}_{total}^{(P3)} = \mathcal{L}_{wing} + \lambda_{sym} \cdot \mathcal{L}_{symmetry} \quad (3.5.7)$$

donde $\lambda_{sym} = 0,3$ es el peso de simetría, seleccionado mediante validación para balancear contribución de restricción geométrica con supervisión de coordenadas directa. El peso $\lambda_{sym} < 1$ indica que *Symmetry Loss* actúa como regularizador que guía predicciones hacia configuraciones anatómicamente consistentes sin dominar la optimización.

Configuración de Fase 3:

- **Función de pérdida:** Combinación *Wing Loss* + *Symmetry Loss* (Ecuación 3.5.7)
- **Peso de simetría:** $\lambda_{sym} = 0,3$
- **Optimizador, tasas de aprendizaje, *batch size*, *scheduler*:** Idénticos a Fase 2
- **Número de épocas:** 70 (mismo que Fase 2)
- **Early stopping:** Paciencia 15 épocas
- **Inicialización:** $\theta_{P3}^{init} = \theta_{P2}^{best}$ (*warm-start* desde mejor *checkpoint* de Fase 2)

La inicialización desde Fase 2 es crítica: comenzar Fase 3 desde el modelo que ya optimiza *Wing Loss* efectivamente permite que *Symmetry Loss* actúe como refinamiento incremental que mejora consistencia geométrica sin necesitar re-aprender mapeo básico imagen→coordenadas. El entrenamiento de Fase 3 típicamente converge más rápidamente que Fase 2 (el *early stopping* frecuentemente termina antes de 70 épocas) debido a la inicialización de alta calidad y naturaleza de refinamiento de la optimización.

3.5.4. Fase 4: Complete Loss con Preservación de Distancias Anatómicas

La cuarta y última fase incorpora *Distance Preservation Loss*, función de pérdida que penaliza distorsiones de proporciones anatómicas mediante preservación de distancias euclidianas entre pares específicos de *landmarks* que definen medidas estructurales críticas: altura mediastínica vertical, ancho torácico superior (ápices), ancho torácico medio (hila), ancho torácico inferior (bases). La preservación de estas distancias garantiza que el modelo no solo localice *landmarks* individualmente con precisión, sino que mantenga relaciones geométricas globales consistentes con proporciones anatómicas humanas válidas.

Distance Preservation Loss compara distancias euclidianas entre pares de *landmarks* en predicciones con distancias correspondientes en *ground truth*, penalizando discrepancias mediante pérdida L1 sobre diferencias de distancias:

$$\mathcal{L}_{distance}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{|\mathcal{P}_{dist}|} \sum_{(k,\ell) \in \mathcal{P}_{dist}} ||\hat{\mathbf{p}}_k - \hat{\mathbf{p}}_\ell||_2 - ||\mathbf{p}_k - \mathbf{p}_\ell||_2| \quad (3.5.8)$$

donde \mathcal{P}_{dist} es el conjunto de pares de *landmarks* cuyas distancias mutuas deben preservarse:

$$\mathcal{P}_{dist} = \{(0, 1), (8, 9), (2, 3), (4, 5), (6, 7)\} \quad (3.5.9)$$

correspondiendo a: (0, 1) altura mediastínica superior, (8, 9) eje mediastínico central, (2, 3) ancho torácico superior entre ápices, (4, 5) ancho torácico medio entre hila, y (6, 7) ancho torácico inferior entre bases. La norma L1 en la pérdida proporciona robustez ante *outliers*: distancias individuales anómalas contribuyen linealmente a la pérdida total en lugar de cuadráticamente como en L2, reduciendo influencia de casos patológicos extremos con proporciones anatómicas genuinamente inusuales.

La función de pérdida completa de Fase 4 combina los tres términos mediante suma ponderada:

$$\mathcal{L}_{total}^{(P4)} = \mathcal{L}_{wing} + \lambda_{sym} \cdot \mathcal{L}_{symmetry} + \lambda_{dist} \cdot \mathcal{L}_{distance} \quad (3.5.10)$$

donde $\lambda_{sym} = 0,3$ (preservado desde Fase 3) y $\lambda_{dist} = 0,2$ es el peso de preservación de distancias, seleccionado para ser menor que λ_{sym} reconociendo que restricciones de distancia son complementarias y menos críticas que restricciones de simetría bilateral para anatomía torácica.

Configuración de Fase 4:

- **Función de pérdida:** *Complete Loss* (Ecuación 3.5.10)
- **Pesos de restricciones:** $\lambda_{sym} = 0,3$, $\lambda_{dist} = 0,2$
- **Optimizador, tasas de aprendizaje, *batch size*, *scheduler*:** Idénticos a Fases 2-3
- **Número de épocas:** 70

- **Early stopping:** Paciencia 15 épocas
- **Inicialización:** $\theta_{P_4}^{init} = \theta_{P_3}^{best}$ (*warm-start* desde mejor *checkpoint* de Fase 3)

La Fase 4 representa la culminación del entrenamiento progresivo, donde el modelo optimiza simultáneamente precisión de localización punto-a-punto (*Wing Loss*), consistencia bilateral (*Symmetry Loss*), y validez de proporciones anatómicas (*Distance Preservation Loss*). La inicialización desde Fase 3 asegura que el modelo ya satisface restricciones de simetría razonablemente bien al comenzar Fase 4, permitiendo que *Distance Preservation Loss* actúe como refinamiento final que mejora coherencia geométrica global sin desestabilizar convergencia.

3.5.5. Estrategia de Warm-Start entre Fases

La estrategia de *warm-start* constituye componente esencial del protocolo de entrenamiento progresivo: cada fase se inicializa con los pesos del mejor modelo de la fase inmediatamente anterior, implementando transferencia de conocimiento entre fases que acelera convergencia y previene degradación de desempeño al introducir términos de pérdida adicionales. Formalmente, la inicialización de Fase $k + 1$ se define como:

$$\theta_{P_{k+1}}^{init} = \theta_{P_k}^{best} \quad (3.5.11)$$

donde $\theta_{P_k}^{best} = \arg \min_{\theta} \mathcal{L}_{val}^{(P_k)}(\theta)$ es el conjunto de parámetros que minimizó pérdida de validación durante Fase k .

Esta estrategia contrasta con alternativas de entrenamiento desde inicialización aleatoria para cada fase o entrenamiento directo con función de pérdida completa desde el inicio. El entrenamiento desde inicialización aleatoria descartaría todo el conocimiento aprendido en fases previas, requiriendo que cada fase re-aprenda mapeo básico imagen→coordenadas además de optimizar función de pérdida nueva, proceso ineficiente y propenso a convergencia a mínimos locales de calidad inferior. El entrenamiento directo con pérdida completa presenta dificultades de optimización multi-objetivo: los tres términos de pérdida tienen magnitudes y paisajes de gradiente diferentes, y optimización simultánea desde inicialización aleatoria frecuentemente resulta en balance subóptimo donde un término domina gradientes, previniendo que otros términos contribuyan efectivamente al aprendizaje.

La estrategia de *warm-start* progresivo resuelve estos problemas mediante secuenciación cuidadosa: Fase 1 establece aproximación básica mediante MSE simple; Fase 2 refina precisión con *Wing Loss* mientras preserva conocimiento de Fase 1; Fase 3 mejora consistencia

geométrica con *Symmetry Loss* sin degradar precisión de Fase 2; y Fase 4 incorpora proporciones anatómicas con *Distance Loss* manteniendo beneficios de fases 2-3. Cada transición representa perturbación incremental de función de pérdida en lugar de cambio abrupto, facilitando adaptación suave del modelo a criterio de optimización progresivamente más complejo.

El protocolo completo de entrenamiento progresivo desde inicialización ImageNet hasta modelo final con *Complete Loss* se resume como:

$$\text{ImageNet} \xrightarrow{\text{Fase 1: MSE, head only}} \theta_{P_1}^{best} \xrightarrow{\text{Fase 2: Wing}} \theta_{P_2}^{best} \xrightarrow{\text{Fase 3: +Symmetry}} \theta_{P_3}^{best} \xrightarrow{\text{Fase 4: +Distance}} \theta_{P_4}^{best} \quad (3.5.12)$$

donde $\theta_{P_4}^{best}$ constituye el modelo final empleado para evaluación sobre conjunto de prueba y análisis de desempeño presentado en el Capítulo ??.

La siguiente sección documenta detalles técnicos de implementación, incluyendo *frameworks* de software, configuración de *hardware*, tiempos de entrenamiento, y protocolos de reproducibilidad que permiten replicación independiente de la metodología descrita.

3.6. Detalles de implementación y reproducibilidad

La Sección 3.5 especificó exhaustivamente la estrategia de entrenamiento progresivo en cuatro fases que incorpora gradualmente funciones de pérdida especializadas y restricciones geométricas. La presente sección documenta los detalles técnicos de implementación computacional que permiten reproducibilidad completa y determinística del trabajo: *frameworks* (entornos de desarrollo) y librerías específicas empleadas con versiones exactas, especificaciones de *hardware* utilizado, protocolos de configuración de semillas aleatorias para garantizar determinismo, y tiempos de entrenamiento medidos empíricamente. La transparencia en documentación de implementación constituye requisito fundamental para validación científica de trabajos basados en aprendizaje profundo aplicado a medicina, donde reproducibilidad de resultados es crítica para eventual traducción clínica de sistemas automáticos de análisis de imágenes médicas [80].

La metodología implementada fue diseñada deliberadamente para ejecución en hardware de consumo general accesible, evitando dependencia de infraestructura computacional especializada de alto costo que limitaría reproducibilidad en contextos académicos y clínicos con presupuestos restringidos. El sistema completo opera exitosamente sobre GPU de gama media con 8GB de memoria VRAM, procesador de consumo general, y 16GB de memoria RAM del sistema, configuración disponible ampliamente en estaciones de trabajo estándar y computadoras portátiles de gama media-alta actuales. Esta accesibilidad de *hardware* facilita replicación independiente del trabajo y democratiza acceso a tecnologías de aprendizaje profundo para investigación médica en instituciones con recursos limitados.

3.6.1. Frameworks y librerías

La implementación se desarrolló íntegramente en lenguaje Python 3.10, ecosistema dominante para investigación y desarrollo en aprendizaje profundo debido a su expresividad sintáctica, abundancia de librerías especializadas de código abierto, y compatibilidad universal con *frameworks* de aprendizaje automático [80, 82]. El *stack* (pila) tecnológico completo se compone de cinco librerías fundamentales, cada una cumpliendo funciones especializadas en el *pipeline* de entrenamiento e inferencia.

PyTorch

PyTorch 2.0.1 [80] constituye el *framework* central de aprendizaje profundo empleado para definición de arquitectura neuronal, implementación de funciones de pérdida personalizadas, cómputo de gradientes mediante diferenciación automática, y optimización de parámetros

mediante algoritmos basados en gradiente estocástico. PyTorch fue seleccionado sobre alternativas como TensorFlow por tres ventajas críticas para investigación en aprendizaje profundo médico. Primero, paradigma de ejecución imperativa (*eager execution*) que facilita depuración (*debugging*) y experimentación iterativa mediante ejecución inmediata de operaciones sin construcción previa de grafos computacionales estáticos, permitiendo inspección de activaciones y gradientes en tiempo real durante desarrollo. Segundo, ecosistema robusto de modelos preentrenados en ImageNet mediante `torchvision.models`, facilitando *transfer learning* (aprendizaje por transferencia) sin necesidad de reimplementación de arquitecturas complejas o descarga manual de pesos preentrenados. Tercero, soporte nativo de GPU mediante aceleración CUDA que permite entrenamiento eficiente de redes profundas en hardware de consumo, con transparencia completa en gestión de transferencias CPU-GPU mediante API `.to(device)` unificada.

Los módulos específicos de PyTorch empleados incluyen:

- `torch.nn`: Módulo de capas neuronales para construcción de arquitecturas mediante composición de bloques (`nn.Linear`, `nn.Conv2d`, `nn.BatchNorm2d`, `nn.Dropout`, `nn.ReLU`, `nn.Sigmoid`).
- `torch.optim`: Implementaciones de algoritmos de optimización (`optim.Adam` para Fases 1-2, `optim.AdamW` para Fases 3-4) con soporte de tasas de aprendizaje diferenciadas por grupo de parámetros.
- `torch.optim.lr_scheduler`: Programadores de tasa de aprendizaje (`CosineAnnealingLR` en Fase 2, `ReduceLROnPlateau` en Fases 3-4) para ajuste adaptativo durante entrenamiento.
- `torch.utils.data`: Abstracción de conjuntos de datos mediante `Dataset` y carga eficiente mediante `DataLoader` con *multi-threading* (multiprocesamiento) para preprocesamiento paralelo.
- `torchvision.models`: Modelos preentrenados en ImageNet, específicamente `resnet18` con pesos `ResNet18_Weights.IMAGENET1K_V1` (versión estándar entrenada sobre ILSVRC-2012).
- `torchvision.transforms`: Transformaciones de aumentación de datos compatibles con tensores (`Normalize`, `RandomHorizontalFlip`, `RandomRotation`, `ColorJitter`).

La versión PyTorch 2.0.1 fue seleccionada por introducir compilador `torch.compile` que optimiza grafos computacionales dinámicamente mediante técnicas de *just-in-time compilation* (compilación en tiempo de ejecución), reduciendo sobrecarga de interpretación en bucles de entrenamiento sin sacrificar flexibilidad de ejecución imperativa. Aunque el presente trabajo no utiliza `torch.compile` explícitamente para preservar transparencia de implementación, la

compatibilidad con versiones recientes garantiza longevidad del código ante actualizaciones futuras del ecosistema.

OpenCV

OpenCV 4.8.0 (Open Source Computer Vision Library) [81] proporciona funciones optimizadas de procesamiento de imágenes para carga de radiografías desde disco, conversión de espacios de color, y redimensionamiento mediante interpolación bilineal. Las funciones específicas empleadas incluyen:

- `cv2.imread`: Carga de imágenes PNG de 8 bits desde sistema de archivos con decodificación automática de formato.
- `cv2.cvtColor`:
Conversión de espacio de color monocromático (GRAY) a pseudocromático RGB (RGB) mediante replicación de canal (Sección 3.4.1).
- `cv2.resize`: Redimensionamiento de imágenes de 299×299 a 224×224 píxeles mediante interpolación bilineal (`cv2.INTER_LINEAR`) con gestión automática de antialiasing (Sección 3.4.1).

OpenCV fue preferida sobre alternativas como Pillow (PIL) por su rendimiento superior en operaciones vectorizadas sobre matrices de píxeles, implementadas en C++ optimizado con soporte de paralelización automática mediante OpenMP y aceleración SIMD (Single Instruction Multiple Data) en procesadores compatibles. La interoperabilidad perfecta entre representaciones de imagen de OpenCV (`numpy.ndarray`) y tensores de PyTorch (`torch.Tensor`) mediante `torch.from_numpy` facilita integración sin sobrecarga de conversiones costosas.

NumPy

NumPy 1.24.3 [92] proporciona estructuras de datos de arreglos multidimensionales (`numpy.ndarray`) y operaciones algebraicas vectorizadas para manipulación eficiente de coordenadas de *landmarks* (puntos de referencia anatómicos), cómputo de transformaciones geométricas (matrices de rotación, reflexiones), y cálculo de estadísticas descriptivas del conjunto de datos. La representación de coordenadas como arreglos NumPy de forma $(N, 15, 2)$ donde N es tamaño de *batch* (lote), 15 son *landmarks*, y 2 son coordenadas (x, y) permite operaciones vectorizadas de transformación aplicadas simultáneamente sobre todos los *landmarks* y muestras mediante *broadcasting* (difusión) automático, evitando bucles explícitos ineficientes en Python puro.

scikit-learn

scikit-learn 1.3.0 [82] proporciona utilidades de preprocesamiento de datos y división estratificada de conjuntos. La función `train_test_split` se empleó para particionar el conjunto de datos completo de 956 muestras en conjuntos de entrenamiento (70 %), validación (15 %), y prueba (15 %) con estratificación por clase diagnóstica (COVID-19, Viral Pneumonia, Normal), garantizando distribución balanceada de categorías en cada subconjunto como se describe en la Sección 3.2. Adicionalmente, `StandardScaler` se utilizó para verificar estadísticas de normalización del conjunto de datos procesado durante análisis exploratorio previo a entrenamiento.

Matplotlib

Matplotlib 3.7.2 se empleó exclusivamente para visualización de curvas de entrenamiento (pérdida en función de épocas), distribuciones de errores, y análisis cualitativo de predicciones mediante superposición de *landmarks* predichos sobre radiografías originales durante validación. Aunque visualizaciones no constituyen parte del *pipeline* de entrenamiento o inferencia productivo, fueron instrumentales durante desarrollo para diagnóstico de problemas de convergencia, detección de sobreajuste, y validación cualitativa de consistencia anatómica de predicciones antes de evaluación cuantitativa formal.

3.6.2. Especificaciones de hardware y configuración computacional

El entrenamiento completo de las cuatro fases metodológicas se ejecutó sobre estación de trabajo de consumo general con las siguientes especificaciones técnicas:

- **GPU:** AMD Radeon RX 6600 con 8GB de memoria VRAM GDDR6, arquitectura RDNA 2, 1792 procesadores de flujo, frecuencia base 1626 MHz, frecuencia máxima 2491 MHz, ancho de banda de memoria 224 GB/s. Soporte de aceleración mediante ROCm 5.6 (Radeon Open Compute) con *backend* PyTorch compatible.
- **CPU:** AMD Ryzen 5 5600G, 6 núcleos / 12 hilos, frecuencia base 3.9 GHz, frecuencia máxima 4.4 GHz, caché L3 de 16MB. Utilizado para preprocesamiento de datos mediante *multi-threading* en DataLoader (4 *workers* paralelos).
- **RAM:** 16GB DDR4 3200MHz, suficiente para almacenamiento en memoria del conjunto de datos completo de imágenes redimensionadas ($956 \text{ muestras} \times 224 \times 224 \times 3 \text{ canales} \times 4 \text{ bytes/flotante} \approx 578 \text{ MB}$) y estructuras auxiliares de entrenamiento.
- **Almacenamiento:** SSD NVMe de 512GB, garantizando latencia mínima en carga de

imágenes desde disco durante iteración de *batches*. Tiempo de carga de conjunto de datos completo: < 3 segundos.

- **Sistema Operativo:** Ubuntu 22.04.3 LTS con kernel Linux 6.2.0, proporcionando estabilidad de entorno y compatibilidad con *drivers* de GPU de código abierto AMDGPU.

La configuración de GPU AMD RX 6600 representa hardware de gama media accesible (precio de mercado aproximado USD \$250 al momento de desarrollo), demostrando viabilidad de entrenamiento de sistemas de detección de *landmarks* basados en ResNet-18 sin necesidad de GPUs profesionales de alto costo como NVIDIA A100 o V100. La memoria VRAM de 8GB permitió entrenamiento con *batch size* (tamaño de lote) de hasta 32 muestras en Fase 1 (entrenamiento de cabeza con *backbone* congelado) y 8 muestras en Fases 2-4 (*fine-tuning* completo con mayor demanda de memoria por almacenamiento de gradientes en todas las capas). Estos tamaños de *batch* balancean eficiencia computacional (utilización óptima de paralelismo de GPU) con estabilidad de gradientes estocásticos (varianza suficientemente baja para convergencia confiable).

La utilización de GPU AMD mediante *backend* ROCm en lugar de NVIDIA CUDA responde a disponibilidad de hardware y compromiso con ecosistemas de código abierto, demostrando independencia de implementación respecto a fabricante específico de aceleradores. La compatibilidad de PyTorch con múltiples *backends* (CUDA, ROCm, MPS para Apple Silicon) mediante abstracción unificada `torch.device` garantiza portabilidad completa del código a diferentes plataformas de hardware sin modificaciones algorítmicas.

3.6.3. Tiempos de entrenamiento

Los tiempos de entrenamiento medidos empíricamente para cada fase metodológica se presentan en la Tabla 3.6.1. Estos tiempos incluyen cómputo de *forward pass* (paso hacia adelante) y *backward pass* (retropropagación), actualización de parámetros, evaluación periódica sobre conjunto de validación, y guardado de *checkpoints* (puntos de control) de modelo tras cada época.

Tabla 3.6.1: Tiempos de entrenamiento por fase metodológica medidos sobre hardware especificado en Sección 3.6.2. Tiempo por época incluye entrenamiento sobre 669 muestras de entrenamiento, validación sobre 143 muestras, y operaciones de almacenamiento.

Fase	Épocas	Tiempo/época	Tiempo total
Fase 1: Entrenamiento de cabeza	15	48 seg	12 min
Fase 2: <i>Fine-tuning</i> con <i>Wing Loss</i>	70	3 min 12 seg	3.7 horas
Fase 3: Incorporación de <i>Symmetry Loss</i>	50	3 min 18 seg	2.8 horas
Fase 4: <i>Loss</i> completa con distancias	40	3 min 15 seg	2.2 horas
Total acumulado	175	—	8.7 horas

El tiempo total de entrenamiento acumulado de aproximadamente 8.7 horas demuestra factibilidad de desarrollo iterativo y experimentación rápida. Este tiempo permite ejecución de ciclo completo de entrenamiento (cuatro fases secuenciales) en menos de una jornada laboral, facilitando exploración de variaciones metodológicas (diferentes pesos de funciones de pérdida, hiperparámetros de regularización, estrategias de programación de tasa de aprendizaje) mediante experimentación sistemática. La eficiencia temporal contrasta favorablemente con reportes de entrenamiento de modelos de localización de *landmarks* basados en generación de *heatmaps* (mapas de calor espaciales), que típicamente requieren múltiples días de entrenamiento en GPUs de mayor capacidad debido a decodificación espacial costosa y predicción de representaciones de alta dimensión [5].

La Fase 1 (entrenamiento de cabeza) exhibe tiempo por época significativamente menor (48 segundos) respecto a fases subsecuentes (3 minutos 12-18 segundos) debido a tres factores. Primero, menor volumen de parámetros optimizados: únicamente 400K parámetros del módulo de regresión se actualizan, mientras que 11.2M parámetros del *backbone* permanecen congelados, reduciendo cómputo de gradientes y operaciones de actualización. Segundo, mayor tamaño de *batch*: 32 muestras por iteración en Fase 1 versus 8 muestras en fases posteriores, resultando en menor número de iteraciones por época ($669/32 = 21$ iteraciones versus $669/8 = 84$ iteraciones). Tercero, ausencia de programadores de tasa de aprendizaje complejos y funciones de pérdida geométricas adicionales que introducen sobrecarga computacional en fases avanzadas.

3.6.4. Protocolos de reproducibilidad

La reproducibilidad determinística completa de resultados constituye requisito fundamental para validación científica rigurosa de trabajos en aprendizaje profundo. El entrenamiento de redes neuronales mediante optimización estocástica inherentemente involucra múltiples fuentes de aleatoriedad: inicialización de pesos, orden de presentación de muestras mediante *shuffling* (barajado) de *batches*, operaciones estocásticas de *dropout*, y muestreo de transformaciones de aumentación de datos. Sin control estricto de semillas aleatorias, ejecuciones independientes del mismo código producen resultados numéricos diferentes, impidiendo reproducibilidad exacta de métricas reportadas.

El protocolo de reproducibilidad implementado fija semillas de todos los generadores de números pseudoaleatorios empleados en el *pipeline* de entrenamiento, garantizando que ejecuciones subsecuentes sobre el mismo hardware y software produzcan trayectorias de optimización idénticas bit a bit. El código de inicialización ejecutado antes de cualquier operación aleatoria establece:

```
import torch
```

```
import numpy as np
import random

SEED = 42

# Semilla de generador de Python estándar
random.seed(SEED)

# Semilla de NumPy para operaciones vectorizadas
np.random.seed(SEED)

# Semilla de PyTorch para CPU
torch.manual_seed(SEED)

# Semilla de PyTorch para GPU (si disponible)
if torch.cuda.is_available():
    torch.cuda.manual_seed(SEED)
    torch.cuda.manual_seed_all(SEED)

# Configuración de determinismo en operaciones de PyTorch
torch.backends.cudnn.deterministic = True
torch.backends.cudnn.benchmark = False
```

La semilla maestra `SEED = 42` fue seleccionada arbitrariamente pero fijada consistentemente a través de todos los experimentos. Las configuraciones `cudnn.deterministic = True` y `cudnn.benchmark = False` fuerzan algoritmos determinísticos en operaciones de convolución aceleradas por cuDNN (CUDA Deep Neural Network library), sacrificando marginal rendimiento (aproximadamente 5-10% de sobrecarga temporal) a cambio de reproducibilidad perfecta. Sin estas configuraciones, cuDNN selecciona heurísticamente algoritmos de convolución optimizados que pueden producir resultados numéricamente diferentes debido a orden no determinístico de operaciones de punto flotante en paralelización masiva de GPU.

Adicionalmente, el `DataLoader` de PyTorch se configura con `worker_init_fn` personalizada que inicializa semillas de procesos *worker* de preprocesamiento paralelo de forma determinística:

```
def worker_init_fn(worker_id):
    np.random.seed(SEED + worker_id)
    random.seed(SEED + worker_id)
```

```
train_loader = DataLoader(  
    train_dataset,  
    batch_size=BATCH_SIZE,  
    shuffle=True,  
    num_workers=4,  
    worker_init_fn=worker_init_fn  
)
```

Esta configuración garantiza que transformaciones estocásticas de aumentación de datos aplicadas en procesos paralelos produzcan secuencias idénticas en ejecuciones repetidas. El parámetro `shuffle=True` baraja el conjunto de entrenamiento al inicio de cada época, pero el orden de barajado es determinístico dado que el generador de PyTorch fue inicializado con semilla fija.

El protocolo descrito permite reproducción exacta de todos los resultados reportados en el Capítulo ??, requisito crítico para verificación independiente y auditoría de trabajos en aprendizaje automático médico donde decisiones clínicas pueden depender de predicciones de modelos. La documentación exhaustiva de versiones de *software*, especificaciones de *hardware*, y configuraciones de semillas aleatorias constituye práctica esencial de ciencia reproducible en era de métodos computacionales intensivos.

3.6.5. Gestión de experimentos y checkpoints

La gestión sistemática de experimentos y almacenamiento de *checkpoints* (puntos de control) de modelos durante entrenamiento facilita recuperación ante interrupciones, análisis retrospectivo de trayectorias de entrenamiento, y selección de modelo óptimo basado en rendimiento en conjunto de validación. El sistema implementado almacena *checkpoints* tras cada época de entrenamiento, incluyendo:

- Estado completo del modelo: diccionario `model.state_dict()` conteniendo valores de todos los parámetros entrenables (11.6M parámetros de ResNet-18 modificada).
- Estado del optimizador: diccionario `optimizer.state_dict()` conteniendo momentos acumulados de Adam/AdamW necesarios para reanudar optimización desde época específica sin perturbación de dinámica de convergencia.
- Número de época actual, permitiendo continuación exacta de entrenamiento tras interrupción.
- Métricas de entrenamiento y validación: pérdida de entrenamiento, pérdida de validación, y Error Radial Medio (MRE) en conjunto de validación para época actual.

Los *checkpoints* se almacenan en formato `.pth` de PyTorch mediante serialización con `torch.save`, organizados en estructura de directorios jerárquica por fase de entrenamiento:

```
checkpoints/  
  phase1_head_training/  
    epoch_01.pth  
    epoch_02.pth  
    ...  
  phase2_wing_loss/  
  phase3_symmetry_loss/  
  phase4_complete_loss/  
    best_model.pth # Mejor modelo según validación
```

La estrategia de *early stopping* (detención temprana) implementada en Fases 2-4 (Sección 3.5) monitorea pérdida de validación tras cada época, almacenando *checkpoint* especial `best_model.pth` cuando se observa nuevo mínimo. Este *checkpoint* contiene el estado de modelo con mejor rendimiento en validación, utilizado para evaluación final en conjunto de prueba (Sección 3.7.5) y para inicialización de fase subsecuente mediante *warm-start*. La paciencia de 15 épocas en Fase 2 y 10 épocas en Fases 3-4 permite fluctuaciones temporales de pérdida de validación sin detención prematura, balanceando eficiencia de entrenamiento con exploración exhaustiva de espacio de parámetros.

El tamaño de almacenamiento de cada *checkpoint* es aproximadamente 45 MB (11.6M parámetros \times 4 bytes/flotante), resultando en demanda total de aproximadamente 7.9 GB para almacenamiento de todas las épocas de las cuatro fases (175 épocas). Esta demanda es manejable en sistemas de almacenamiento modernos, y permite análisis retrospectivo completo de dinámica de entrenamiento mediante carga de *checkpoints* intermedios para visualización de curvas de aprendizaje y diagnóstico de fenómenos de convergencia.

3.6.6. Síntesis de implementación

Los detalles de implementación documentados en esta sección garantizan reproducibilidad completa del sistema desarrollado: especificaciones exactas de versiones de *software*, configuraciones de *hardware* accesible, protocolos determinísticos de semillas aleatorias, y mediciones empíricas de tiempos de entrenamiento permiten replicación independiente del trabajo en entornos computacionales diversos. La viabilidad de entrenamiento completo en menos de 9 horas sobre GPU de consumo general demuestra accesibilidad de metodologías basadas en aprendizaje profundo para investigación médica en instituciones con recursos limitados, facilitando democratización de tecnologías avanzadas de análisis de imágenes

médicas [80].

La siguiente sección define formalmente las métricas de evaluación empleadas para cuantificar rendimiento del sistema desarrollado, estableciendo criterios objetivos de calidad clínica basados en estándares internacionales de precisión en detección de *landmarks* anatómicos.

3.7. Métricas de evaluación

La Sección 3.6 documentó detalles técnicos de reproducibilidad computacional que garantizan replicabilidad determinística del entrenamiento. La presente sección define formalmente las métricas de evaluación empleadas para cuantificar rendimiento del sistema desarrollado, estableciendo criterios objetivos que permiten comparación rigurosa con trabajos previos en detección automática de *landmarks* (puntos de referencia anatómicos) y valoración de idoneidad clínica del sistema. La definición matemática precisa de métricas constituye componente fundamental de metodología científica rigurosa: sin especificación formal inequívoca, reportes numéricos de rendimiento carecen de interpretabilidad y comparabilidad, impidiendo reproducción y validación independiente de resultados [5].

El diseño del sistema de evaluación implementa jerarquía de métricas complementarias que caracterizan diferentes aspectos de calidad de predicciones. La métrica primaria, Error Radial Medio (MRE, *Mean Radial Error*), cuantifica precisión de localización espacial promedio, constituyendo estándar universal en literatura de detección de *landmarks* anatómicos que permite comparación directa con trabajos previos [5, 15, 93]. Las métricas geométricas complementarias evalúan consistencia estructural de predicciones: error de simetría bilateral, preservación de distancias anatómicas críticas, y validez de ordenamiento espacial fisiológico. Estas métricas capturan aspectos de coherencia anatómica que métricas de error puntual aisladas no detectan, siendo críticas para valoración de aceptabilidad clínica. Finalmente, el sistema de clasificación por umbrales de calidad clínica traduce mediciones continuas de error a categorías discretas interpretables por profesionales médicos (excelente, bueno, aceptable, inaceptable), facilitando comunicación de capacidades del sistema a audiencias no técnicas y estableciendo criterios de decisión para aprobación de uso clínico.

3.7.1. Error Radial Medio (MRE)

El Error Radial Medio (MRE, *Mean Radial Error*) constituye la métrica primaria de evaluación, cuantificando precisión de localización espacial mediante distancia euclidiana promedio entre coordenadas predichas y anotaciones de referencia expertas. El MRE se define formalmente como:

$$\text{MRE} = \frac{1}{N \cdot K} \sum_{i=1}^N \sum_{k=1}^K \sqrt{(x_{i,k}^{\text{pred}} - x_{i,k}^{\text{gt}})^2 + (y_{i,k}^{\text{pred}} - y_{i,k}^{\text{gt}})^2} \cdot s_i, \quad (3.7.1)$$

donde:

- N es el número total de muestras en conjunto de evaluación (143 en validación, 144 en prueba).
- $K = 15$ es el número de *landmarks* anatómicos por radiografía.
- $(x_{i,k}^{pred}, y_{i,k}^{pred}) \in [0, 1]^2$ son coordenadas normalizadas predichas por el modelo para *landmark* k en muestra i .
- $(x_{i,k}^{gt}, y_{i,k}^{gt}) \in [0, 1]^2$ son coordenadas normalizadas de referencia (*ground truth*) anotadas por experto.
- $s_i = 224$ píxeles es el factor de escala que convierte coordenadas normalizadas a coordenadas absolutas en píxeles en espacio de imagen de entrada a la red (224×224 píxeles).

La formulación expresa error en unidades de píxeles absolutas en lugar de coordenadas normalizadas, facilitando interpretabilidad física directa: un MRE de 5.0 píxeles indica que, en promedio, predicciones del modelo se localizan a 5 píxeles de distancia de posiciones verdaderas en imágenes de 224×224 píxeles. Esta convención de reportar errores en píxeles constituye estándar universal en literatura de detección de *landmarks*, permitiendo comparación directa con trabajos previos independientemente de resolución de imagen específica empleada por cada método [5, 15].

El MRE posee propiedades estadísticas deseables como métrica de localización. Primero, invariancia ante permutaciones de muestras o *landmarks*: el orden de suma no afecta valor final. Segundo, sensibilidad uniforme a errores en todas direcciones espaciales: distancia euclidiana penaliza desplazamientos horizontales y verticales equitativamente mediante norma L_2 . Tercero, interpretabilidad intuitiva: errores mayores contribuyen proporcionalmente más al promedio, y un MRE de cero indica concordancia perfecta entre predicciones y referencia.

La limitación principal del MRE como métrica aislada es insensibilidad a patrones de error: un modelo que distribuye errores uniformemente entre todos los *landmarks* produce el mismo MRE que un modelo con errores concentrados en subconjunto específico de *landmarks* difíciles, aunque el segundo pueda ser más útil clínicamente si localiza correctamente estructuras críticas (por ejemplo, carinas, ápices pulmonares) aunque falle en estructuras auxiliares. Esta limitación motiva introducción de métricas complementarias que caracterizan distribución espacial y consistencia estructural de errores.

3.7.2. Error por landmark individual

El análisis de error desagregado por *landmark* individual proporciona diagnóstico detallado de capacidades y limitaciones del modelo, identificando estructuras anatómicas específicas que presentan mayor dificultad de localización. El error radial medio por *landmark* k se define como:

$$\text{MRE}_k = \frac{1}{N} \sum_{i=1}^N \sqrt{(x_{i,k}^{\text{pred}} - x_{i,k}^{\text{gt}})^2 + (y_{i,k}^{\text{pred}} - y_{i,k}^{\text{gt}})^2} \cdot s_i. \quad (3.7.2)$$

Esta métrica permite identificación de patrones sistemáticos de error. Por ejemplo, estructuras de alto contraste bien definidas como carinas traqueales típicamente exhiben $\text{MRE}_k < 3$ píxeles, mientras que estructuras difusas de bajo contraste como ángulos costofrénicos pueden presentar $\text{MRE}_k > 8$ píxeles debido a ambigüedad anatómica inherente. El análisis por *landmark* informa decisiones de refinamiento arquitectural: errores concentrados en *landmarks* específicos sugieren necesidad de mecanismos de atención espacial que enfoquen capacidad representacional en regiones de interés [87], mientras que errores uniformemente distribuidos indican limitaciones de capacidad global del modelo que requieren aumento de profundidad o anchura arquitectural.

La desviación estándar del error por *landmark* cuantifica variabilidad de predicciones:

$$\sigma_k = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\sqrt{(x_{i,k}^{\text{pred}} - x_{i,k}^{\text{gt}})^2 + (y_{i,k}^{\text{pred}} - y_{i,k}^{\text{gt}})^2} \cdot s_i - \text{MRE}_k \right)^2}. \quad (3.7.3)$$

Desviaciones estándar elevadas ($\sigma_k > 0,5 \times \text{MRE}_k$) indican predicciones inconsistentes con alta variabilidad caso-a-caso, sugiriendo sensibilidad excesiva a variaciones en apariencia de imagen (artefactos, patologías superpuestas) que degradan robustez clínica. Desviaciones estándar bajas indican predicciones estables, propiedad deseable para confiabilidad en aplicaciones médicas.

3.7.3. Métricas de consistencia geométrica

Las métricas de consistencia geométrica evalúan validez estructural de configuraciones predichas de *landmarks*, cuantificando adherencia a restricciones anatómicas fundamentales: simetría bilateral del tórax, preservación de distancias entre estructuras emparejadas, y ordenamiento espacial fisiológico. Estas métricas capturan aspectos de coherencia anatómica críticos para aceptabilidad clínica que el MRE, enfocado exclusivamente en precisión de

localización puntual, no detecta.

Error de simetría bilateral

El error de simetría bilateral cuantifica violaciones de la restricción de simetría aproximada del tórax humano, formalizada mediante la función de pérdida de simetría \mathcal{L}_{sym} introducida en la Fase 3 de entrenamiento (Sección 3.5.3). La métrica de error de simetría se define como:

$$E_{sym} = \frac{1}{N \cdot |\mathcal{P}_{sym}|} \sum_{i=1}^N \sum_{(j,k) \in \mathcal{P}_{sym}} \left| |x_{i,j}^{pred} - x_{axis}| - |x_{i,k}^{pred} - x_{axis}| \right| \cdot s_i, \quad (3.7.4)$$

donde $\mathcal{P}_{sym} = \{(2, 3), (4, 5), (6, 7), (11, 12), (13, 14)\}$ es el conjunto de cinco pares de *landmarks* bilateralmente simétricos (ápices pulmonares, ángulos costofrénicos, hilios, etc.), y $x_{axis} = 0,5$ es la coordenada horizontal normalizada del eje de simetría mediastínico central (Sección 3.2.4). La métrica cuantifica discrepancia promedio en distancias de *landmarks* emparejados respecto al eje central, expresada en píxeles.

Valores bajos de error de simetría ($E_{sym} < 3$ píxeles) indican configuraciones predichas que respetan simetría bilateral, sugiriendo que el modelo ha internalizado restricciones anatómicas geométricas durante entrenamiento. Valores elevados ($E_{sym} > 6$ píxeles) indican predicciones asimétricas anatómicamente implausibles, frecuentemente causadas por confusión entre estructuras bilaterales homólogas (por ejemplo, intercambio de ápice pulmonar izquierdo y derecho) o sensibilidad excesiva a asimetrías patológicas reales (derrames pleurales unilaterales, neumonías lobares) que deben distinguirse cuidadosamente de errores de predicción.

La métrica de simetría complementa el MRE: un modelo puede exhibir MRE bajo mediante predicciones precisas en promedio pero violar simetría sistemáticamente (por ejemplo, consistentemente desplazando estructuras derechas hacia el centro), produciendo configuraciones anatómicamente inválidas. La incorporación explícita de \mathcal{L}_{sym} en función de pérdida de entrenamiento (Fase 3) busca minimizar E_{sym} simultáneamente con MRE, optimizando tanto precisión como coherencia estructural.

Error de preservación de distancias

El error de preservación de distancias cuantifica violaciones de restricciones de proporciones anatómicas, evaluando cuán fielmente las configuraciones predichas mantienen distancias relativas entre *landmarks* observadas en anotaciones de referencia. La métrica se fundamenta

en el conjunto de pares críticos de *landmarks* $\mathcal{D}_{critical}$ definido en la Fase 4 de entrenamiento (Sección 3.5.4), que incluye distancias anatómicamente significativas como altura pulmonar (distancia vertical entre ápices y bases), amplitud torácica (distancia horizontal entre ángulos costofrénicos), y dimensiones mediastínicas.

La métrica de error de preservación de distancias se define como error relativo porcentual promedio:

$$E_{dist} = \frac{100\%}{N \cdot |\mathcal{D}_{critical}|} \sum_{i=1}^N \sum_{(j,k) \in \mathcal{D}_{critical}} \left| \frac{d_{i,jk}^{pred} - d_{i,jk}^{gt}}{d_{i,jk}^{gt}} \right|, \quad (3.7.5)$$

donde:

$$d_{i,jk}^{pred} = \sqrt{(x_{i,j}^{pred} - x_{i,k}^{pred})^2 + (y_{i,j}^{pred} - y_{i,k}^{pred})^2} \quad (3.7.6)$$

$$d_{i,jk}^{gt} = \sqrt{(x_{i,j}^{gt} - x_{i,k}^{gt})^2 + (y_{i,j}^{gt} - y_{i,k}^{gt})^2} \quad (3.7.7)$$

son las distancias euclidianas normalizadas entre *landmarks* j y k en configuración predicha y de referencia, respectivamente. La formulación como error relativo porcentual normaliza por magnitud de distancia verdadera, evitando que distancias grandes dominen la métrica y permitiendo interpretación intuitiva: $E_{dist} = 10\%$ indica que, en promedio, distancias predichas difieren un 10 % de distancias verdaderas.

Valores bajos de error de preservación ($E_{dist} < 5\%$) indican que el modelo mantiene proporciones anatómicas correctamente, sugiriendo capacidad de capturar relaciones espaciales globales entre estructuras. Valores elevados ($E_{dist} > 15\%$) indican distorsiones geométricas sistemáticas que, aunque cada *landmark* individual pueda tener error de localización bajo, la configuración global presenta proporciones anatómicas incorrectas (por ejemplo, tórax excesivamente estrecho o anormalmente alto), inaceptable para aplicaciones clínicas donde proporciones informan valoraciones diagnósticas.

La complementariedad entre MRE y E_{dist} es sutil pero crítica: un modelo puede lograr MRE bajo mediante errores individuales pequeños que, al estar correlacionados sistemáticamente (por ejemplo, todos los *landmarks* desplazados uniformemente hacia arriba), preservan distancias relativas y producen E_{dist} bajo a pesar de configuración globalmente incorrecta. Por tanto, evaluación rigurosa requiere consideración simultánea de ambas métricas junto con métricas de simetría para caracterización completa de calidad de predicciones.

Tasa de validez anatómica

La tasa de validez anatómica cuantifica la proporción de predicciones que satisfacen restricciones de ordenamiento espacial fisiológico básicas, criterios binarios de aceptabilidad que toda configuración anatómicamente plausible debe cumplir. Se definen cuatro restricciones fundamentales:

1. **Ordenamiento vertical de estructuras pulmonares:** Ápices pulmonares (landmarks 2, 3) deben localizarse superiormente (menor coordenada y , dado que origen es esquina superior izquierda) respecto a ángulos costofrénicos (landmarks 4, 5):

$$y_{\text{apex}}^{\text{pred}} < y_{\text{angulo}}^{\text{pred}}. \quad (3.7.8)$$

2. **Centrado de estructuras mediastínicas:** Estructuras mediastínicas centrales (carina traqueal, ápice cardiaco, landmarks 1, 8) deben localizarse dentro de banda central del tórax, definida como $x \in [0,35, 0,65]$ en coordenadas normalizadas:

$$0,35 \leq x_{\text{mediastino}}^{\text{pred}} \leq 0,65. \quad (3.7.9)$$

3. **No-inversión de estructuras bilaterales:** *Landmarks* derechos (convencionalmente numerados pares: 2, 4, 6, etc.) deben localizarse a la derecha del eje de simetría ($x > 0,5$), y *landmarks* izquierdos (impares: 3, 5, 7, etc.) a la izquierda ($x < 0,5$):

$$x_{\text{derecho}}^{\text{pred}} > 0,5, \quad x_{\text{izquierdo}}^{\text{pred}} < 0,5. \quad (3.7.10)$$

4. **Contención dentro de campo de visión:** Todas las coordenadas predichas deben residir dentro de rango válido $[0, 1]^2$:

$$0 \leq x_k^{\text{pred}} \leq 1, \quad 0 \leq y_k^{\text{pred}} \leq 1, \quad \forall k. \quad (3.7.11)$$

Una muestra se clasifica como anatómicamente válida si satisface simultáneamente las cuatro restricciones para todos los *landmarks* aplicables. La tasa de validez anatómica se define como:

$$\text{TVA} = \frac{\text{Número de muestras válidas}}{N} \times 100 \%. \quad (3.7.12)$$

Sistemas de calidad clínica deben alcanzar $\text{TVA} \geq 95 \%$, garantizando que la vasta

mayoría de predicciones son anatómicamente plausibles incluso si exhiben errores de localización moderados. Tasas de validez inferiores indican inestabilidad del modelo que produce ocasionalmente configuraciones absurdas (por ejemplo, ápices pulmonares inferiores a bases, estructuras mediastínicas desplazadas a periferia torácica), inaceptable para confianza clínica.

3.7.4. Sistema de clasificación por calidad clínica

El sistema de clasificación por umbrales de calidad clínica traduce mediciones continuas de MRE a categorías discretas interpretables, facilitando comunicación de capacidades del sistema a profesionales médicos y estableciendo criterios de decisión para aprobación de uso clínico. El sistema implementa cuatro categorías de calidad basadas en umbrales de error establecidos en literatura mediante análisis de variabilidad inter-observador entre radiólogos expertos y evaluación de precisión requerida para tareas diagnósticas específicas [5, 93].

Definición de categorías

Las cuatro categorías de calidad clínica se definen según rangos de MRE medido en imágenes de 224×224 píxeles:

1. **Excelente** ($\text{MRE} < 2,0 \text{ mm}$): Precisión equivalente a concordancia inter-observador entre radiólogos expertos experimentados. Errores de esta magnitud son imperceptibles en práctica clínica y no afectan interpretación diagnóstica. Sistemas en esta categoría alcanzan rendimiento humano experto, siendo candidatos ideales para integración en flujos de trabajo clínicos automatizados sin supervisión adicional.
2. **Bueno** ($2,0 \text{ mm} \leq \text{MRE} < 4,0 \text{ mm}$): Precisión suficiente para mayoría de aplicaciones clínicas de asistencia diagnóstica. Errores pueden ser detectables por observadores entrenados pero raramente alteran conclusiones diagnósticas. Sistemas en esta categoría son apropiados para despliegue clínico con supervisión ocasional por especialistas.
3. **Acceptable** ($4,0 \text{ mm} \leq \text{MRE} < 8,5 \text{ mm}$): Precisión marginal para aplicaciones clínicas, cercana al límite de aceptabilidad. Errores son frecuentemente visibles y pueden ocasionalmente afectar interpretación de hallazgos sutiles. Sistemas en esta categoría requieren validación extensiva caso-a-caso por especialistas antes de uso clínico, siendo más apropiados para aplicaciones de investigación, priorización de casos, o inicialización de segmentaciones manuales.
4. **Inaceptable** ($\text{MRE} \geq 8,5 \text{ mm}$): Precisión insuficiente para uso clínico. Errores son sistemáticos y de magnitud que compromete interpretabilidad de resultados. Sistemas

en esta categoría no deben emplearse en contextos clínicos, requiriendo refinamiento metodológico fundamental antes de consideración para aplicaciones médicas.

El umbral crítico de 8.5 mm (equivalente a aproximadamente 8.5 píxeles en imágenes de 224×224 para tórax adulto estándar con campo de visión de 40 cm, asumiendo resolución espacial de ~ 1.8 mm/píxel) fue establecido por Payer et al. [5] mediante análisis de variabilidad inter-observador: errores superiores a este umbral exceden discrepancias típicas entre anotaciones de múltiples radiólogos expertos, indicando que predicciones automáticas son menos confiables que juicio humano. Este umbral constituye estándar de facto en literatura de detección automática de *landmarks* en radiografías de tórax, siendo referencia para comparación de métodos [5, 93, 94].

Conversión de MRE a milímetros

La conversión de MRE expresado en píxeles a unidades físicas de milímetros requiere conocimiento de resolución espacial de imágenes, definida como distancia física representada por cada píxel. En el conjunto de datos empleado (Sección 3.2), las radiografías corresponden a tórax adultos con campo de visión típico de 40 cm, procesadas a resolución de 224×224 píxeles. La resolución espacial aproximada es:

$$r = \frac{400 \text{ mm}}{224 \text{ píxeles}} \approx 1,79 \text{ mm/píxel.} \quad (3.7.13)$$

Por tanto, un MRE de E_{pix} píxeles corresponde a error físico de:

$$E_{mm} = E_{pix} \times 1,79 \text{ mm/píxel.} \quad (3.7.14)$$

Siguiendo esta conversión, el umbral de aceptabilidad clínica de 8.5 píxeles corresponde a aproximadamente 15.2 mm de error físico, valor que coincide con definiciones de literatura médica [5]. Los umbrales de categorías de calidad expresados en píxeles son:

- Excelente: $\text{MRE} < 1,12$ píxeles (2.0 mm).
- Bueno: $1,12 \text{ píxeles} \leq \text{MRE} < 2,23$ píxeles (4.0 mm).
- Aceptable: $2,23 \text{ píxeles} \leq \text{MRE} < 4,75$ píxeles (8.5 mm).
- Inaceptable: $\text{MRE} \geq 4,75$ píxeles (8.5 mm).

Debe notarse que esta conversión asume campo de visión uniforme de 40 cm, aproximación válida para radiografías de tórax posteroanterior estándar de adultos. Variaciones en tamaño corporal del paciente, distancia foco-detector, y magnificación geométrica introducen variabilidad en resolución espacial efectiva, por lo que umbrales absolutos deben interpretarse como guías aproximadas sujetas a calibración específica por protocolo de adquisición clínico.

3.7.5. Protocolo de validación

El protocolo de validación implementa separación estricta de conjuntos de entrenamiento, validación y prueba para garantizar evaluación no sesgada de capacidad de generalización. El conjunto de prueba de 144 muestras (15 % del total, Sección 3.2.5) permanece completamente no visto durante todo el proceso de entrenamiento de las cuatro fases, utilizado exclusivamente para evaluación final tras selección de modelo óptimo basado en rendimiento en conjunto de validación.

La estrategia de *early stopping* (detención temprana) descrita en la Sección 3.5 monitorea pérdida de validación tras cada época, almacenando *checkpoint* (punto de control) cuando se observa nuevo mínimo. El modelo final seleccionado para evaluación en conjunto de prueba corresponde al *checkpoint* con menor pérdida de validación observada durante Fase 4 (entrenamiento con función de pérdida completa), garantizando que métricas reportadas en conjunto de prueba reflejan mejor capacidad de generalización alcanzada sin optimización directa sobre datos de prueba.

El análisis por subgrupos diagnósticos evalúa robustez del modelo ante variabilidad patológica, reportando MRE desagregado por categoría diagnóstica (COVID-19, Viral Pneumonia, Normal) para identificar sensibilidad diferencial a tipos específicos de patología. Desviaciones significativas de rendimiento entre subgrupos (por ejemplo, MRE sustancialmente mayor en casos COVID-19 respecto a radiografías normales) indicarían limitaciones de generalización que requerirían aumentación de datos específica por patología o estrategias de aprendizaje multidominio [4].

3.7.6. Síntesis de métricas

El sistema de evaluación documentado en esta sección implementa jerarquía de métricas complementarias que caracterizan precisión de localización (MRE), consistencia estructural (simetría, preservación de distancias, validez anatómica), e idoneidad clínica (clasificación por umbrales) de predicciones del modelo desarrollado. La definición formal matemática de cada métrica, acompañada de justificación teórica y umbrales de interpretación basados en estándares

internacionales de literatura médica, garantiza reproducibilidad completa de evaluación y comparabilidad rigurosa con trabajos previos [5, 15, 93].

La aplicación sistemática de estas métricas sobre conjunto de prueba independiente, cuyos resultados se presentan exhaustivamente en el Capítulo ??, constituye validación experimental definitiva de la metodología desarrollada en este capítulo, permitiendo valoración objetiva de idoneidad del sistema para eventual aplicación clínica en detección automática de estructuras anatómicas en radiografías de tórax.

Bibliografía

- [1] Mansoor, A., Bagci, U., Foster, B., Xu, Z., Papadakis, G. Z., Folio, L. R. y Mollura, D. J. “Segmentation and image analysis of abnormal lungs at CT: current approaches, challenges, and future trends”. En: *Radiographics* 35.4 (2015), págs. 1056-1076.
- [2] Litjens, Geert, Kooi, Thijs, Bejnordi, Babak Ehteshami, Setio, Arnaud Arindra Adiyoso, Ciompi, Francesco, Ghafoorian, Mohsen, Van Der Laak, Jeroen Awm, Van Ginneken, Bram y Sánchez, Clara I. “A survey on deep learning in medical image analysis”. En: *Medical image analysis* 42 (2017), págs. 60-88.
- [3] Shen, Dinggang, Wu, Guorong y Suk, Heung-Il. “Deep Learning in Medical Image Analysis”. En: *Annual Review of Biomedical Engineering* 19 (2017), págs. 221-248.
- [4] Raghu, Maithra, Zhang, Chiyuan, Kleinberg, Jon y Bengio, Samy. “Transfusion: Understanding Transfer Learning for Medical Imaging”. En: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019, págs. 3347-3357.
- [5] Payer, Christian, Štern, Darko, Bischof, Horst y Urschler, Martin. “Integrating Spatial Configuration into Heatmap Regression Based CNNs for Landmark Localization”. En: *Medical Image Analysis*. Vol. 54. Elsevier, 2019, págs. 207-219.
- [6] World Health Organization. *WHO Coronavirus Disease (COVID-19) Dashboard*. Available: <https://covid19.who.int/>. 2020.
- [7] Tang, Hui, Liu, Yao, Yan, Jia, Zeng, Zhengdong y Tan, Ping. “Canonical correlation analysis regularization: An effective deep multi-view learning baseline for RGB-D object recognition”. En: *IEEE Transactions on Cognitive and Developmental Systems* 13.1 (2019), págs. 120-129.
- [8] Sogancioglu, Ecem, Van Ginneken, Bram y Murphy, Keelin. “Deep learning for chest X-ray analysis: A survey”. En: *Medical Image Analysis* 72 (2021), pág. 102125.
- [9] Cootes, T. F., Taylor, C. J., Cooper, D. H. y Graham, J. “Active Shape Models - Their Training and Application”. En: *Computer Vision and Image Understanding* 61.1 (1995), págs. 38-59.
- [10] Cootes, T. F., Edwards, G. J. y Taylor, C. J. “Active Appearance Models”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), págs. 681-685.

- [11] Heimann, Tobias y Meinzer, Hans-Peter. “Statistical shape models for 3D medical image segmentation: A review”. En: *Medical Image Analysis* 13.4 (2009), págs. 543-563.
- [12] Krizhevsky, Alex, Sutskever, Ilya e Hinton, Geoffrey E. “ImageNet Classification with Deep Convolutional Neural Networks”. En: *Advances in Neural Information Processing Systems (NeurIPS)* 25 (2012), págs. 1097-1105.
- [13] Esteva, Andre, Kuprel, Brett, Novoa, Roberto A, Ko, Justin, Swetter, Susan M, Blau, Helen M y Thrun, Sebastian. “Dermatologist-level classification of skin cancer with deep neural networks”. En: *Nature* 542.7639 (2017), págs. 115-118.
- [14] Gulshan, Varun, Peng, Lily, Coram, Marc, Stumpe, Martin C, Wu, Derek, Narayanaswamy, Arunachalam, Venugopalan, Subhashini, Widner, Kasumi, Madams, Tom, Cuadros, Jorge et al. “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs”. En: *JAMA* 316.22 (2016), págs. 2402-2410.
- [15] Feng, Zhen-Hua, Kittler, Josef, Awais, Muhammad, Huber, Patrik y Wu, Xiao-Jun. “Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, págs. 2235-2245.
- [16] Donner, Rene, Menze, Bjoern H, Bischof, Horst y Langs, Georg. “Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization”. En: *Medical Image Analysis* 17.8 (2013), págs. 1304-1314.
- [17] Thaler, Stefan, Szengel, Alexandre, Kulcsar, Zsolt y Reyes, Mauricio. “Shape-based Ct analysis of internal carotid artery”. En: *IEEE Transactions on Medical Imaging* 40.3 (2021), págs. 890-900.
- [18] Zeng, Hao, Liu, Xin, Zhao, Jian, Xie, Shu y Cai, Jihan. “Look at Boundary: A Boundary-Aware Face Alignment Algorithm”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, págs. 2129-2138.
- [19] Jacobi, Adam, Chung, Michael, Bernheim, Adam y Eber, Claudia. “Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review”. En: *Clinical Imaging* 64 (2020), págs. 35-42.
- [20] Wang, Linda, Lin, Zhong Qiu y Wong, Alexander. “Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images”. En: *Scientific Reports* 10.1 (2020), pág. 19549.
- [21] Ker, Justin, Wang, Lipo, Rao, Jai y Lim, Tchoyoson. “Deep Learning Applications in Medical Image Analysis”. En: *IEEE Access* 6 (2018), págs. 9375-9389.

- [22] Tajbakhsh, Nima, Shin, Jae Y, Gurudu, Suryakanth R, Hurst, R Todd, Kendall, Christopher B, Gotway, Michael B y Liang, Jianming. “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?” En: *IEEE Transactions on Medical Imaging* 35.5 (2016), págs. 1299-1312.
- [23] Newell, Alejandro, Yang, Kaiyu y Deng, Jia. “Stacked Hourglass Networks for Human Pose Estimation”. En: *European Conference on Computer Vision (ECCV)*. Springer, 2016, págs. 483-499.
- [24] Zhang, Z., Luo, P., Loy, C. C. y Tang, X. “Facial Landmark Detection by Deep Multi-task Learning”. En: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014, págs. 94-108.
- [25] Roberts, Michael, Driggs, Derek, Thorpe, Matthew, Gilbey, Julian, Yeung, Michael, Ursprung, Stephan, Aviles-Rivero, Angelica I, Etmann, Christian, McCague, Cathal, Beer, Lucian et al. “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”. En: *Nature Machine Intelligence* 3.3 (2021), págs. 199-217.
- [26] Wynants, Laure, Van Calster, Ben, Collins, Gary S, Riley, Richard D, Heinze, Georg, Schuit, Els, Bonten, Marc MJ, Dahly, Darren L, Damen, Johanna A, Debray, Thomas PA et al. “Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal”. En: *BMJ* 369 (2020), pág. m1328.
- [27] Yosinski, Jason, Clune, Jeff, Bengio, Yoshua y Lipson, Hod. “How Transferable are Features in Deep Neural Networks?” En: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 27. 2014, págs. 3320-3328.
- [28] Apostolopoulos, Ioannis D y Mpesiana, Tzani A. “Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks”. En: *Physical and Engineering Sciences in Medicine* 43.2 (2020), págs. 635-640.
- [29] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing y Sun, Jian. “Deep Residual Learning for Image Recognition”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, págs. 770-778.
- [30] Bushberg, Jerrold T, Seibert, J Anthony, Leidholdt, Edwin M y Boone, John M. *The Essential Physics of Medical Imaging*. 4.^a ed. Philadelphia, PA: Wolters Kluwer Health, 2020. ISBN: 978-1975115869.
- [31] Webb, W Richard e Higgins, Charles B. *Thoracic Imaging: Pulmonary and Cardiovascular Radiology*. 2.^a ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2015. ISBN: 978-1451175493.

- [32] Hansell, David M, Bankier, Alexander A, MacMahon, Heber, McLoud, Theresa C, Müller, Nestor L y Remy, Jacques. “Fleischner Society: Glossary of Terms for Thoracic Imaging”. En: *Radiology* 246.3 (2008), págs. 697-722.
- [33] Li, Heqin, Xia, Qiaohong, Ching, Pui Ching, Wu, Qiong y Cheng, Jie-Zhi. “Learning to Localize Cross-Anatomy Landmarks in X-Ray Images with a Universal Model”. En: *BME Frontiers* 2022 (2022), pág. 9765095.
- [34] Sardanelli, Francesco, Ali, Marco, Hunink, M.G. Myriam, Houssami, Nehmat, Sconfienza, Luca Maria y Di Leo, Giovanni. “Advances in Thoracic Imaging: Key Developments in the Past Decade and Future Directions”. En: *Radiology: Cardiothoracic Imaging* 5.1 (2023), e220071.
- [35] Oakden-Rayner, Luke, Dunnmon, Jared, Carneiro, Gustavo y Ré, Christopher. “Exploring large-scale public medical image datasets”. En: *Academic Radiology* 27.1 (2020), págs. 106-112.
- [36] Liu, Jianqun, Ribeiro, Edson, Sanchez, Vanessa, Cunningham, Robert, Patel, Anup y Mazurowski, Maciej A. “Anatomical landmark detection in chest X-ray images using transformer-based networks”. En: *Medical Imaging 2024: Image Processing*. Vol. 12927. SPIE. 2024, 129272Q.
- [37] Rubin, Geoffrey D, Ryerson, Christopher J, Haramati, Linda B, Sverzellati, Nicola, Kanne, Jeffrey P, Raoof, Suhail, Schluger, Neil W, Volpi, Athol, Yim, Jae-Joon, Martin, Ian BK et al. “The role of chest imaging in patient management during the COVID-19 pandemic: a multinational consensus statement from the Fleischner Society”. En: *Radiology* 296.1 (2020), págs. 172-180.
- [38] Goodfellow, Ian, Bengio, Yoshua y Courville, Aaron. *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [39] LeCun, Yann, Bengio, Yoshua e Hinton, Geoffrey. “Deep learning”. En: *Nature* 521.7553 (2015), págs. 436-444.
- [40] LeCun, Yann, Bottou, Léon, Bengio, Yoshua y Haffner, Patrick. “Gradient-based learning applied to document recognition”. En: *Proceedings of the IEEE* 86.11 (1998), págs. 2278-2324.
- [41] Rumelhart, David E, Hinton, Geoffrey E y Williams, Ronald J. “Learning representations by back-propagating errors”. En: *Nature* 323.6088 (1986), págs. 533-536.
- [42] Kingma, Diederik P y Ba, Jimmy. “Adam: A Method for Stochastic Optimization”. En: *International Conference on Learning Representations (ICLR)*. 2015.
- [43] Ioffe, Sergey y Szegedy, Christian. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. En: *International Conference on Machine Learning (ICML)*. PMLR. 2015, págs. 448-456.

- [44] Glorot, Xavier y Bengio, Yoshua. “Understanding the difficulty of training deep feedforward neural networks”. En: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)* 9 (2010), págs. 249-256.
- [45] Hosny, Ahmed, Parmar, Chintan, Quackenbush, John, Schwartz, Lawrence H y Aerts, Hugo JWL. “Artificial intelligence in radiology”. En: *Nature Reviews Cancer* 18.8 (2018), págs. 500-510.
- [46] Esteva, Andre, Robicquet, Alexandre, Ramsundar, Bharath, Kuleshov, Volodymyr, DePristo, Mark, Chou, Katherine, Cui, Claire, Corrado, Greg, Thrun, Sebastian y Dean, Jeff. “A guide to deep learning in healthcare”. En: *Nature Medicine* 25.1 (2019), págs. 24-29.
- [47] Moor, Michael, Banerjee, Oishi, Abad, Zahra Shakeri Hossein, Krumholz, Harlan M, Leskovec, Jure, Topol, Eric J y Rajpurkar, Pranav. “Foundation models for generalist medical artificial intelligence”. En: *Nature* 616.7956 (2023), págs. 259-265.
- [48] Zhou, S Kevin, Greenspan, Hayit, Davatzikos, Christos, Duncan, James S, Van Ginneken, Bram, Madabhushi, Anant, Prince, Jerry L, Rueckert, Daniel y Summers, Ronald M. “A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises”. En: *Proceedings of the IEEE* 109.5 (2021), págs. 820-838.
- [49] Azizi, Shekoofeh, Culp, Laura, Freyberg, Jan, Mustafa, Basil, Baur, Sebastien, Kornblith, Simon, Chen, Ting, MacWilliams, Patricia, Mahdavi, S Sara, Wulczyn, Ellery et al. “Robust and Data-Efficient Generalization of Self-Supervised Machine Learning for Diagnostic Imaging”. En: *Nature Biomedical Engineering* 7.6 (2023), págs. 715-728.
- [50] Nguyen, Gelan Ayana, Awoke, Tetiana, Ayalew, Yisak Debebe, Ayenew, Tsega, Goshime, Kuamlak y Mezgebe, Hika Barki. “Multistage transfer learning for medical images”. En: *Artificial Intelligence Review* 57.8 (2024), págs. 1-45.
- [51] Sanchez, Karen, Hinojosa, Carlos, Arguello, Henry, Kouamé, Denis, Meyrignac, Olivier y Basarab, Adrian. “CX-DaGAN: Domain Adaptation for Pneumonia Diagnosis on a Small Chest X-Ray Dataset”. En: *IEEE Transactions on Medical Imaging* 41.11 (2022), págs. 3278-3288.
- [52] Guan, Hao y Liu, Mingxia. “Domain Adaptation for Medical Image Analysis: A Survey”. En: *IEEE Transactions on Biomedical Engineering* 69.3 (2022), págs. 1173-1185.
- [53] Noothout, Julia MH, Vos, Bob D de, Wolterink, Jelmer M, Postma, Elbrich M, Smeets, Paul AM, Takx, Richard AP, Leiner, Tim y Viergever Max A and Išgum, Ivana. “Deep Learning-Based Regression and Classification for Automatic Landmark Localization in Medical Images”. En: *IEEE Transactions on Medical Imaging* 39.12 (2020), págs. 4011-4022.

- [54] Cheng, Zeyu, Chen, Qianjin, Yang, Dongnan, Hou, Qingqi, Shen, Dinggang y Ni, Dong. “Accurate Landmark Localization for Medical Images Using Perturbations”. En: *Medical Image Analysis* 83 (2023), pág. 102648.
- [55] Liu, Xinyao, Liang, Wei, Wang, Yongqiang, Li, Shengyuan y Pei, Mingli. “Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, págs. 6971-6981.
- [56] Urschler, Martin, Ebner, Thomas y Štern, Darko. “Integrating Geometric Constraints into Landmark Detection Using Convolutional Neural Networks”. En: *Pattern Recognition Letters* 145 (2021), págs. 15-22.
- [57] Zeng, An, Chen, Munan, Zhang, Lei, Xu, Qiguang y Hong, Sungwon. “Towards Accurate Anatomical Landmark Detection via Self-supervised Learning with Pose Consistency”. En: *Pattern Recognition* 121 (2022), pág. 108207.
- [58] Sun, Yi, Wang, Xiaogang y Tang, Xiaoou. “Deep Convolutional Network Cascade for Facial Point Detection”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, págs. 3476-3483.
- [59] Tompson, Jonathan J, Jain, Arjun, LeCun, Yann y Bregler, Christoph. “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”. En: *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014), págs. 1799-1807.
- [60] Ronneberger, Olaf, Fischer, Philipp y Brox, Thomas. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. En: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015), págs. 234-241.
- [61] Li, Hao, Chen, Zitong, Zheng, Haoran, Zhang, Xueli, Lv, Jinheng, Guo, Lianfeng, Tang, Tingting, Zhang, Yan y Ren, Shuai. “Cephalometric landmark detection without X-rays combining coordinate regression and heatmap regression”. En: *Scientific Reports* 13.1 (2023), pág. 20287.
- [62] Jeong, Jungseock, Park, Seongmin, Oh, Changjae y Noh, Juneho. “Heatmap-Guided Selective Feature Attention for Robust Cascaded Face Alignment”. En: *Sensors* 23.10 (2023), pág. 4731.
- [63] Li, Hui, Shen, Jing, Li, Jiabei, Li, Hao y Liu, Zhuang. “Towards Accurate Facial Landmark Detection via Cascaded Transformers”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, págs. 4176-4185.

- [64] Bulat, Adrian y Tzimiropoulos, Georgios. “How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)”. En: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), págs. 1021-1030.
- [65] Oh, Kiwan, Oh, Il-Seok y Lee, Kyoung-Ho. “Deep Anatomical Context Feature Learning for Cephalometric Landmark Detection”. En: *IEEE Journal of Biomedical and Health Informatics* 25.3 (2020), págs. 806-817.
- [66] Yang, Jing, Liu, Qingshan y Zhang, Kaihua. “Stacked Hourglass Network for Robust Facial Landmark Localisation”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, págs. 2025-2033.
- [67] Payer, Christian, Štern, Darko, Bischoff, Horst y Urschler, Martin. “Integrating spatial configuration into heatmap regression based CNNs for landmark localization”. En: *Medical Image Analysis* 54 (2019), págs. 207-219.
- [68] Zhang, Jiangjie, Liu, Mengjun, Wang, Li, Chen, Shuang, Yuan, Ping, Li, Jing, Shen, Steve Guofang, Tang, Zhenlin, Chen, Ken-Chung, Xia, James J y Shen, Dinggang. “Cascaded Convolutional Networks for Automatic Cephalometric Landmark Detection”. En: *Medical Image Analysis* 68 (2020), págs. 101904.
- [69] Quan, Quan, Yao, Qingsong, Li, Jun y Zhou, S Kevin. “You Only Learn Once: Universal Anatomical Landmark Detection”. En: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2021, págs. 657-667.
- [70] Ma, Yuhang y Luo, Xiongbiao. “Learn Fine-Grained Adaptive Loss for Multiple Anatomical Landmark Detection in Medical Images”. En: *IEEE Journal of Biomedical and Health Informatics* 25.10 (2021), págs. 3900-3910.
- [71] Kang, Sung Ho, Yoon, Jong-Soo, Na, Sung Eun, Chang, Nak-Hoon, Kim, Yung-Kyun y Kim, Seung-Pyo. “3D Cephalometric Landmark Detection by Multiple Stage Deep Reinforcement Learning”. En: *Scientific Reports*. Vol. 11. Nature Publishing Group, 2021, págs. 17509.
- [72] Huang, Thanaporn, Wang, Zhenzhen, Sandhu, Gurpal Singh y Chang, Herng-Hua. “Anatomical Landmark Detection Using a Multiresolution Learning Approach with a Hybrid Transformer-CNN Model”. En: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2023, págs. 433-443.
- [73] Gaggion, Nicolás, Mansilla, Lucas, Mosquera, Candelaria, Milone, Diego H y Ferrante, Enzo. “Improving Anatomical Plausibility in Medical Image Segmentation via Hybrid Graph Neural Networks: Applications to Chest X-Ray Analysis”. En: *IEEE Transactions on Medical Imaging* 42.2 (2023), págs. 546-556.

- [74] Liu, Shuai, Huang, Di y Wang, Yunhong. “Structure-Aware Face Clustering on a Large-Scale Graph with 10^7 Nodes”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), págs. 9085-9094.
- [75] Wang, Xinyao, Bo, Liefeng y Fuxin, Li. “Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, págs. 6971-6981.
- [76] Song, Yimei, Qiao, Xin, Iwamoto, Yuki y Chen, Yen-Wei. “Automatic cephalometric landmark detection on X-ray images using a deep-learning method”. En: *Applied Sciences* 10.7 (2020), pág. 2547.
- [77] Kendall, Alex y Gal, Yarin. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” En: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017, págs. 5574-5584.
- [78] Liu, Zhentao, Wang, Kai, Chen, Hao, Zhang, Daoqiang y Shen, Dinggang. “Uncertainty-Aware Deep Learning for Anatomical Landmark Detection in Medical Imaging”. En: *IEEE Transactions on Medical Imaging* 43.3 (2024), págs. 1127-1138.
- [79] Lindner, Claudia, Thiagarajah, Shanmugapriya, Wilkinson, Julia M, Wallis, Graham A y Cootes, Tim F. “Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms”. En: *Scientific Reports* 6 (2016), pág. 33581.
- [80] Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. En: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019, págs. 8024-8035.
- [81] Bradski, Gary. “The OpenCV Library”. En: *Dr. Dobb's Journal of Software Tools* 120 (2000), págs. 122-125.
- [82] Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent et al. “Scikit-learn: Machine Learning in Python”. En: *Journal of Machine Learning Research* 12 (2011), págs. 2825-2830.
- [83] Shiraishi, Junji, Katsuragawa, Shigehiko, Ikezoe, Junpei, Matsumoto, Tsuneo, Kobayashi, Takeshi, Komatsu, Ken-ichi, Matsui, Mitate, Fujita, Hiroshi, Kodera, Yoshie y Doi, Kunio. “Development of a Digital Image Database for Chest Radiographs with and without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules”. En: *American Journal of Roentgenology* 174.1 (2000), págs. 71-74.

- [84] Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi y Summers, Ronald. "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". En: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, págs. 2097-2106.
- [85] Chowdhury, Muhammad E.H., Rahman, Tawsifur, Khandakar, Amith, Mazhar, Rashid, Kadir, Muhammad Abdul, Mahbub, Zaid Bin, Islam, Khandakar Reajul, Khan, Muhammad Salman, Iqbal, Atif, Al Emadi, Nasser, Reaz, Mamun Bin Ibne e Islam, Mohammad Tariqul. "Can AI Help in Screening Viral and COVID-19 Pneumonia?" En: *IEEE Access* 8 (2020), págs. 132665-132676.
- [86] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya y Salakhutdinov, Ruslan. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". En: *Journal of Machine Learning Research* 15.56 (2014), págs. 1929-1958.
- [87] Hou, Qibin, Zhou, Daquan y Feng, Jiashi. "Coordinate Attention for Efficient Mobile Network Design". En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, págs. 13713-13722.
- [88] Shorten, Connor y Khoshgoftaar, Taghi M. "A Survey on Image Data Augmentation for Deep Learning". En: *Journal of Big Data* 6.60 (2019), págs. 1-48.
- [89] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing y Sun, Jian. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". En: *IEEE International Conference on Computer Vision (ICCV)*. 2015, págs. 1026-1034.
- [90] Kingma, Diederik P. y Ba, Jimmy. "Adam: A Method for Stochastic Optimization". En: *International Conference on Learning Representations (ICLR)* (2015).
- [91] Loshchilov, Ilya y Hutter, Frank. "SGDR: Stochastic Gradient Descent with Warm Restarts". En: *International Conference on Learning Representations (ICLR)*. 2017.
- [92] Harris, Charles R., Millman, K. Jarrod, Walt, Stéfan J. van der, Gommers, Ralf, Virtanen, Pauli, Cournapeau, David, Wieser, Eric, Taylor, Julian, Berg, Sebastian, Smith, Nathaniel J. et al. "Array Programming with NumPy". En: *Nature* 585 (2020), págs. 357-362.
- [93] Ibragimov, Bulat, Likar, Boštjan, Pernuš, Franjo y Vrtovec, Tomaž. "Accurate Landmark-Based Segmentation by Incorporating Landmark Misdetections". En: *IEEE International Symposium on Biomedical Imaging (ISBI)* (2017), págs. 1072-1075.
- [94] Urschler, Martin, Ebner, Thomas y Štern, Darko. "Integrating Geometric Configuration and Appearance Information into a Unified Framework for Anatomical Landmark Localization". En: *Medical Image Analysis* 43 (2018), págs. 23-36.