

UNIVERSIDAD DE GUANAJUATO

CAMPUS IRAPUATO-SALAMANCA
DIVISIÓN DE INGENIERÍAS

A Deep Learning Framework For Automatic Cephalometric Landmark Localization Through Gaussian Heatmap Regression.

TESIS

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN INGENIERÍA ELÉCTRICA

(Instrumentación y sistemas digitales)

PRESENTA:

Ing. Simon Tshibengabu Kalonji

DIRECTORES:

Dr. Juan Gabriel Aviña Cervantes

M.I. Fernando Daniel Hernández Gutiérrez

Salamanca, Guanajuato, México. Octubre, 2026

Salamanca, Gto., a 08 de septiembre del 2026.

MTRO. JUAN SIGFRIDO LÓPEZ CUSTODIO
COORDINADOR DE ASUNTOS ESCOLARES
P R E S E N T E.

Por medio de la presente, se otorga autorización para proceder a los trámites de impresión, empastado de tesis y titulación al alumno(a) ***Simon Tshibengabu Kalonji*** del ***Programa de Maestría Ingeniería Eléctrica (Instrumentación y Sistemas Digitales)*** y cuyo número de ***NUA*** es: 148898 del cual soy director. El título de la tesis es: *A Deep Learning Framework For Automatic Cephalometric Landmark Localization Through Gaussian Heatmap Regression*

Hago constar que he revisado dicho trabajo y he tenido comunicación con los sinodales asignados para la revisión de la tesis, por lo que no hay impedimento alguno para fijar la fecha de examen de titulación.

A T E N T A M E N T E.

Dr. Juan Gabriel Aviña Cervantes
DIRECTOR DE TESIS
SECRETARIO

Hernandez Gutierrez Fernando Daniel

M.I. Fernando Daniel Hernández Gutiérrez
DIRECTOR DE TESIS

Dr. Oscar Gerardo Ibarra Manzano
PRESIDENTE

Dr. Juan Pablo Ignacio Ramírez Paredes
VOCAL

Dedication

To my dear mother, Perpetue Ngoya Lukusa, who left us far too soon and now rests in peace: I so wished you were here. But the laws of nature were stricter towards you and stronger than my will.

Personal Acknowledgments

I thank Dr. Juan Gabriel Aviña Cervantes for agreeing to supervise this master's thesis. I also express my gratitude to Mr. Fernando Daniel Hernández Gutiérrez for his unwavering support and for the time he generously dedicated to me whenever I needed him.

I would like to pay tribute to my immediate family, especially my wife, Georgette Kapinga, and my five children, whose love, patience, and encouragement have been a constant source of strength throughout this journey.

I sincerely thank my father, Albert Kalonji, my older brother, Philippe Mbuyi, and my sister, Jeannette Mbombo for their support, advice, and encouragement.

I would also like to express my deepest gratitude to my very dear boss, the current General Manager of Société Générale de Télédistribution "S.G.T S.A", where I work; to Dr. Elisée Ilunga Mbuyamba, current Director General of my alma mater, ISTA/Kinshasa, and to his wife, Arlette Mbelu; to my close and dear friend Edgard Tshishimbi Kanyinda; as well as to all my other friends (Dr. Blaise, the maestros Francisco, Karine, and Tere), fellow graduates, and the Mexican families who have supported me throughout this journey.

Finally, I thank the members of the examining committee for their constructive feedback and valuable suggestions.

Institutional Acknowledgments

I express my sincere gratitude to the University of Guanajuato, especially the Division of Engineering at the Irapuato-Salamanca Campus (DICIS), for accepting my application to pursue the master's degree and for providing me with the university education I sought. This research work received support from the following institutions.:

- *Department of Electronics Engineering, University of Guanajuato, NUA: 148898.*



- *Secretariat of Science, Humanities, Technology and Innovation (SECIHTI), for the financial support provided, grant CVU: 1321998.*



- *University of Guanajuato, for the support provided through the Institutional Call for Scientific Research, project CIIC-318/2026.*

Abstract

Accurate cephalometric landmark localization is essential for orthodontic diagnosis, craniofacial assessment, and treatment planning. Yet reliable deep-learning performance on limited annotated datasets depends not only on network architecture but also on landmark representation, coordinate decoding, physical error computation, and validation design. This thesis develops a reproducible framework for the automatic localization of 19 anatomical landmarks in lateral cephalometric radiographs.

The framework integrates a lightweight two-level U-Net with Gaussian heatmap regression, differentiable Soft-Argmax coordinate extraction, and image-specific anisotropic pixel-to-millimeter conversion. Evaluation used the PKU Cephalometric Dataset of 102 radiographs. The Holdout 70/30 protocol was defined as the primary evaluation, with 72 radiographs for model development and 30 fully isolated radiographs for final testing; a K-Fold 70/30 heatmap ensemble on the same 30-image external cohort and a Nested Leave-One-Out Cross-Validation (Nested LOOCV) procedure, providing one strictly out-of-sample prediction per radiograph, served as complementary validation.

Under the primary protocol, the framework achieved a Mean Radial Error (MRE) of 2.1663 mm, with Successful Detection Rates of 80.35%, 87.37%, 91.40%, and 95.79% at thresholds of 2.0, 2.5, 3.0, and 4.0 mm, respectively. The K-Fold ensemble and Nested LOOCV yielded MREs of 2.2423 mm and 2.2541 mm (95% CI [1.2232, 3.2850]; median landmark error 1.1148 mm). Paired statistical analyzes on the 30 common radiographs revealed no statistically significant differences between protocols, supporting robust performance.

The main contribution is methodological rather than architectural: established components are integrated into a coherent pipeline combining spatial prediction, continuous coordinate estimation, physically consistent error measurement, strict external evaluation, and matched statistical analysis. These results establish validation design as an integral component of deep-learning-based cephalometric localization on small annotated medical-imaging datasets.

Keywords: Cephalometric landmark localization; Gaussian heatmap regression; U-Net; Soft-Argmax; Holdout validation; Nested cross-validation; PKU dataset; Medical image analysis.

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.2 Justification . . . 3

1.3 Research Problem . . . 4

1.4 Research Questions . . . 5

1.5 Scope of this Study . . . 5

1.6 Research Objectives . . . 6

1.6.1 Specific Objectives . . . 6

1.7 Research Hypotheses . . . 7

1.8 Main Contributions . . . 7

1.9 Manuscript Organization . . . 8

1.10 Chapter Summary . . . 9

2 State of the Art **10**

2.1 Lateral Cephalometric Radiography . . . 10

2.2 Cephalometric Analysis in Orthodontics . . . 12

2.3 Cephalometric Analysis by Manual Localization . . . 14

2.4 Classical Landmark Detection Approaches . . . 15

2.4.1 Knowledge-Based Methods and Contour Detection . . . 16

2.4.2 Shape-Based Models: ASM and AAM . . . 17

2.4.3 Random Forests and Classical Machine Learning . . . 17

2.5 Deep Learning for Medical Imaging . . . 18

2.5.1 Artificial Neural Networks . . . 18

2.5.2 Biological Inspiration for CNNs . . . 19
2.5.3 From Fully Connected Networks to CNNs . . . 21
2.5.4 The Convolution Operation . . . 23
2.5.5 Activation Functions and Nonlinear Processing . . . 25
2.5.6 Pooling and Spatial Downsampling . . . 27
2.5.7 Hierarchical Feature Learning . . . 29
2.5.8 How Convolutional Neural Networks Work . . . 30
2.5.9 Advantages and Limitations of CNNs . . . 32
2.5.10 CNN-Based Landmark Localization . . . 33
2.6 Heatmap-Based Landmark Localization . . . 34
2.6.1 Coordinate Regression versus Heatmap Regression . . . 34
2.6.2 Gaussian Heatmap Representation . . . 35
2.7 U-Net Architectures for Segmentation and Landmark Localization . . . 36
2.7.1 Original U-Net Architecture . . . 37
2.7.2 Contracting and Expanding Paths . . . 38
2.7.3 Skip Connections and Preservation of Spatial Information . . . 39
2.8 Recent Developments in Deep Learning for Cephalometric Landmark Localization . . . 40
2.8.1 Evolution of Deep Learning Architectures . . . 40
2.8.2 Emerging Methodological Trends . . . 41
2.8.3 Performance Comparison . . . 42
2.8.4 Remaining Challenges . . . 43
2.8.5 Motivation for the Proposed Methodology . . . 44
2.9 Differentiable Coordinate Extraction . . . 45
2.9.1 Existing Coordinate Extraction Strategies . . . 46
2.9.2 Mathematical Formulation of Soft-Argmax . . . 46
2.10 Image Preprocessing and Loss Function . . . 47
2.11 Evaluation Metrics and Validation Protocol . . . 48
2.12 Chapter Conclusion . . . 49

3 General Methodology **50**
3.1 Introduction . . . 50

3.2 Heatmap-Based Landmark Localization Framework . . . 51
3.3 Computational Environment . . . 52
3.4 Methodological Workflow . . . 53
3.5 Dataset Overview . . . 55
3.6 Data Preprocessing . . . 56
3.6.1 Image Resizing and Intensity Scaling . . . 56
3.6.2 Landmark Coordinate Transformation . . . 56
3.6.3 Anisotropic Conversion from Resized Pixels to Millimeters . . . 57
3.6.4 Landmark Heatmap Generation . . . 58
3.7 Architecture of the Proposed Model . . . 59
3.7.1 Detailed Network Structure . . . 59
3.7.2 Output Dimensions and Basic Operations . . . 60
3.8 Heatmap Decoding to Coordinates . . . 62
3.8.1 Classical Argmax . . . 62
3.8.2 Soft-Argmax . . . 62
3.9 Learning Strategy . . . 64
3.9.1 Primary Holdout 70/30 Experiment . . . 64
3.9.2 Complementary K-Fold 70/30 Experiment . . . 64
3.9.3 Complementary Nested LOOCV Partitioning . . . 65
3.9.4 Nested LOOCV Training Procedure . . . 65
3.9.5 Training Hyperparameters . . . 66
3.9.6 Heatmap Regression Loss . . . 66
3.9.7 Training-Duration Selection and Refitting . . . 67
3.10 Aggregation of Predictions across Protocols . . . 68
3.11 Quantitative Evaluation Protocol . . . 68
3.11.1 Mean Radial Error . . . 69
3.11.2 Successful Detection Rate . . . 69
3.11.3 Root Mean Squared Error . . . 69
3.11.4 Standard Deviation of Localization Error . . . 70
3.11.5 Per-Image and Per-Landmark Evaluation . . . 70
3.11.6 Statistical Reporting of Evaluation Results . . . 71

3.12 Qualitative Analysis and Visualization . . . 71
3.12.1 Overlay Visualization of Predictions . . . 71
3.12.2 Visual Monitoring of Training Dynamics . . . 71
3.12.3 Graphical Performance Analysis . . . 72
3.13 Organization of the Experimental Set-up . . . 72
3.13.1 Reproducibility Considerations . . . 73
3.14 Scientific Rationale of the Methodology . . . 74
3.15 Chapter Conclusion . . . 75

4 Results and Discussion **76**
4.1 Introduction . . . 76
4.2 Experimental Configuration . . . 76
4.3 Primary Results: Holdout 70/30 . . . 77
4.3.1 Experimental Configuration . . . 77
4.3.2 Global Performance . . . 78
4.3.3 Image-Level Results . . . 79
4.3.4 Landmark-Level Results . . . 80
4.3.5 Graphical Analysis . . . 80
4.3.6 Interpretation . . . 81
4.4 Additional Validation Results . . . 88
4.4.1 K-Fold 70/30 Ensemble . . . 89
4.4.2 Nested Leave-One-Out Cross-Validation . . . 93
4.5 Inter-Protocol Comparison . . . 96
4.5.1 Global Performance Comparison . . . 96
4.5.2 Paired Statistical Comparison . . . 97
4.5.3 Landmark-Specific Errors and Extreme Cases . . . 98
4.5.4 Comparison with the State of the Art . . . 99
4.5.5 Methodological and Clinical Interpretation . . . 101
4.6 Chapter Summary . . . 102

5 General Conclusions **104**
5.1 Introduction . . . 104

5.2 Summary of the Work Performed . . . 104
5.3 Main Findings . . . 105
5.4 Main Contributions of the Study . . . 107
5.4.1 Integrated Localization Pipeline . . . 107
5.4.2 Continuous Coordinate Extraction . . . 107
5.4.3 Image-Specific Anisotropic Evaluation . . . 107
5.4.4 Three-Level Validation Hierarchy with Matched Statistical Analysis . . . 107
5.4.5 Reproducible and Traceable Workflow . . . 108
5.5 Limitations of the Study . . . 108
5.6 Future Work and Perspectives . . . 109
5.6.1 Independent External Validation . . . 109
5.6.2 Larger and More Diverse Populations . . . 109
5.6.3 Analysis of Difficult Cases and Landmarks . . . 109
5.6.4 Prediction Uncertainty and Failure Detection . . . 109
5.6.5 Controlled Ablation Studies . . . 109
5.6.6 Toward Clinical Evaluation . . . 110
5.7 Final Conclusion . . . 110

Bibliography **118**

List of Figures

1.1 Representative examples of malocclusion types and prevalence 2

1.2 Head movement simulation system with a mounted dry skull, prepared for X-ray acquisition using a cephalostat [Martin et al., 2025]. 3

1.3 Manual cephalometric landmark annotation and reference-plane construction 4

2.1 Lateral radiograph of a human skull, commonly referred to as a lateral cephalogram [Izzetti et al., 2026]. 11

2.2 Spatial distribution of the 19 cephalometric landmarks considered in this study [Sumon et al., 2025]. 12

2.3 Example of conventional manual localization and tracing of cephalometric landmarks on a lateral radiograph [Bansal et al., 2012]. 14

2.4 Methodological families for automatic cephalometric landmark localization 16

2.5 Architecture of an artificial neuron 19

2.6 Hierarchical visual-processing stages from retina to hypercomplex cells 20

2.7 Architecture of a fully connected network applied to a 4×4 image 22

2.8 Principles of convolutional layers 22

2.9 Two-dimensional cross-correlation with a 3×3 Sobel filter 24

2.10 Effect of padding on the spatial dimensions of the output feature map 25

2.11 Selected activation functions used in convolutional neural networks 27

2.12 Illustration of max-pooling, average-pooling, and stochastic pooling applied to a 4×4 input matrix using a 2×2 pooling window and a stride of 2. 28

2.13 Hierarchical feature learning in convolutional neural networks 30

2.14 General processing pipeline of a CNN for cephalometric image analysis 31

2.15 Principle of heatmap-based landmark localization 34

2.16 Gaussian target heatmaps for different values of the standard deviation σ 36

2.17 Original U-Net architecture 37

3.1 Hardware and software architecture of the experimental environment. 52

3.2 Overall workflow of the cephalometric landmark-localization framework 54

3.3 Two-level U-Net architecture used for cephalometric landmark localization 60

4.1 Image-level MRE values for the Holdout 70/30 final test set. 81

4.2 Distribution of the landmark-level radial errors in the Holdout 70/30 protocol. 81

4.3 Successful Detection Rate as a function of the error threshold in the Holdout 70/30 protocol. . . 82

4.4 Distribution of radial errors across the 19 landmarks in the Holdout 70/30 protocol. 82

4.5 Violin plots of the landmark-level radial errors in the Holdout 70/30 protocol. 83

4.6 Heatmap of the landmark-level radial errors in the Holdout 70/30 protocol. 83

4.7 Landmark-specific performance represented as a radar chart in the Holdout 70/30 protocol. . . . 84

4.8 Ground-truth versus predicted landmark coordinates in the Holdout 70/30 protocol. 85

4.9 Training curve of the final model refitted on the 72 development images. 86

4.10 Qualitative localization overlays for four representative test images from the primary Holdout 70/30 protocol: **(a)** ID 63, **(b)** ID 90, **(c)** ID 45, and **(d)** ID 18. 87

4.11 Predicted inner-validation heatmap at six epoadd for a representative outer fold 95

List of Tables

2.1 Definition of the 19 cephalometric landmarks considered in this study 13

2.2 Mapping between visual-processing stages and their CNN counterparts 21

2.3 Structural comparison between a fully connected network and a CNN 23

2.4 Comparison of common activation functions 28

2.5 Comparison of common pooling operations, including their mathematical principles, advantages, and limitations. 29

2.6 Main advantages and limitations of conventional convolutional neural networks for medical image analysis. 32

2.7 Representative deep learning approaches for cephalometric landmark localization 43

3.1 Software dependencies and CUDA configuration used in the implementation. 53

3.2 Composition and output dimensions of the functional blocks in the proposed U-Net. 61

3.3 Main hyperparameters shared by the primary Holdout 70/30 protocol and the complementary K-Fold 70/30 and Nested LOOCV experiments. 67

4.1 Summary of the experimental configuration and validation protocols. 77

4.2 Experimental configuration of the primary Holdout 70/30 protocol. 78

4.3 Global performance on the 30-image Holdout 70/30 final test set. 78

4.4 Image-level results for the Holdout 70/30 final test set. 79

4.5 Landmark-level results for the Holdout 70/30 final test set. 80

4.6 Landmark-level localization results for test image ID 63. 88

4.7 Landmark-level localization results for test image ID 90. 89

4.8 Landmark-level localization results for test image ID 45. 90

4.9 Landmark-level localization results for test image ID 18. 91

4.10 Fold-wise training and validation configuration for the K-Fold 70/30 ensemble. 91

4.11 Global performance of the K-Fold 70/30 ensemble on the 30-image final test set. 92

4.12 Landmark-level results for the K-Fold 70/30 ensemble on the 30-image final test set. 92

4.13 Overall performance of the Nested LOOCV protocol across the 102 held-out images. 93

4.14 Landmark-level results for the Nested LOOCV protocol across the 102 held-out images. 94

4.15 Global comparison of the evaluation protocols. Medians and standard deviations are computed at the image level. 96

4.16 Statistical comparison of the evaluation protocols on the 30 common radiographs 97

4.17 Comparison with representative methods reported in the literature 100

Chapter 1

Introduction

Cephalometric analysis remains a central component of orthodontic diagnosis, craniofacial growth assessment, and treatment planning [Durão et al., 2013; Junaid et al., 2022]. Although three-dimensional imaging modalities are increasingly available, lateral cephalometric radiographs remain important in clinical practice because they are standardized, affordable, and widely accessible. The reliability of cephalometric analysis, however, depends critically on the accurate localization of anatomical landmarks, a task that is time-consuming and subject to substantial observer variability when performed manually [Baumrind y Frantz, 1971].

During the past decade, deep learning has advanced medical image analysis and created new opportunities for automated cephalometric landmark localization. Nevertheless, recent reviews have shown that high predictive performance alone does not guaranty clinical usefulness. Robustness, evaluation rigor, reproducibility, and interpreting errors in clinically meaningful units remain important unresolved issues [Schwendicke et al., 2021; Mesquita et al., 2023]. This thesis responds to these concerns by developing an integrated deep learning framework that combines spatially informed prediction, Gaussian heatmap representation, differentiable coordinate extraction, rigorous validation, and clinically interpretable performance assessment.

1.1 Motivation

Malocclusions and maxillofacial deformities are growth-related abnormalities that affect the facial bones, jaws, and dentofacial structures [Chen et al., 2014]. As illustrated in Figure 1.1, these abnormalities may occur in the sagittal, vertical, or transverse dimension and may influence occlusal function, facial balance, speech, mastication, and facial appearance. Because they affect both function and morphology, accurate anatomical assessment is essential for diagnosis and treatment planning.

Among the imaging modalities used in orthodontic diagnosis, lateral cephalometric radiography remains one of the most widely adopted. It provides a standardized two-dimensional representation of the craniofacial complex and has a long history of clinical use [Broadbent, 1931]. Cephalometry is a quantitative radiographic technique in which anatomical landmarks are identified on standardized lateral skull images, after which angular



Figure 1.1: *Representative examples of different types of malocclusion and their reported prevalence in children, adapted from [Mendoza-Oropeza et al., 2014]:* ***(a,b)*** *anterior crossbite (12.8 %);* ***(c,d)*** *anterior open bite (35.1 %);* ***(e)*** *upper occlusal view associated with a horizontal overbite (7.4 %);* ***(f)*** *edge-to-edge bite;* ***(g)*** *vertical overbite;* ***(h)*** *upper anterior crowding (19.6 %); and* ***(i)*** *lower anterior crowding (26.4 %).*

and linear measurements are derived for diagnosis, treatment planning, growth monitoring, and post-treatment assessment [Durão et al., 2013; Junaid et al., 2022]. Its clinical value lies in its ability to translate anatomical structures into measurable variables that support the assessment of skeletal deformities, dental relationships, and soft-tissue balance [Jacobson y Jacobson, 2006]. Consequently, cephalometric analysis has become an integral part of orthodontic practice.

The reliability of cephalometric analysis is only as strong as the landmark localization on which it relie. Some landmarks have clear anatomical boundaries, whereas others lie in regions with weak contrast, poorly defined contours, or overlapping structures [Baumrind y Frantz, 1971]. Because lateral cephalometric radiography projects three-dimensional anatomy onto a two-dimensional image, several anatomical structures may overlap within the same region, complicating landmark identification.

As shown in Figure 1.2, landmark identification may remain sensitive to patient movement, image quality, anatomical variability, and observer experience, even when image acquisition is standardized with a cephalostat [Martin et al., 2025; Baumrind y Frantz, 1971].



Figure 1.2: Head movement simulation system with a mounted dry skull, prepared for X-ray acquisition using a cephalostat [Martin et al., 2025].

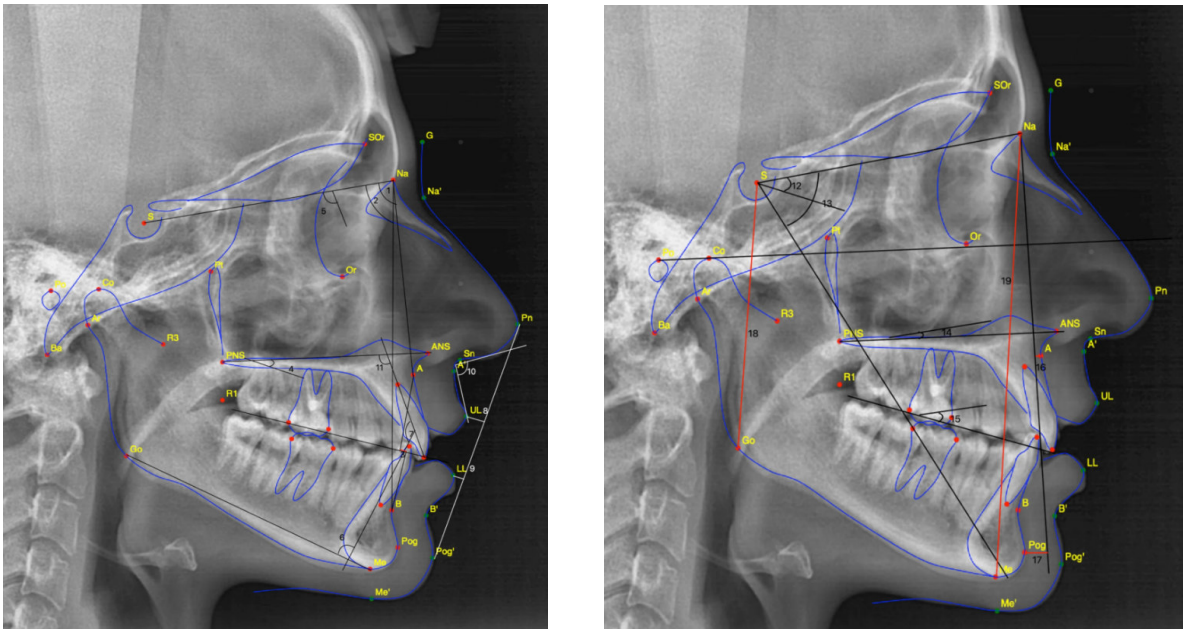
1.2 Justification

This research is motivated by both clinical needs and methodological limitations identified in the current literature. Clinically, manual cephalometric landmark annotation is laborious and subject to observer-dependent variability [Singh et al., 2025]. Accurate annotation requires training, experience, and sustained attention, particularly when many landmarks must be identified or when assessments are repeated throughout treatment.

Figure 1.3 illustrates the dense network of landmark points and intersecting reference planes required for a single diagnostic assessment. It also conveys the visual complexity and cognitive load associated with manual tracing.

Some landmarks also exhibit substantial intra- and inter-observer variability, and even expert annotations are unlikely to be perfectly uniform [Hendrickx et al., 2024]. This variability can reduce the reproducibility of subsequent cephalometric measurements and supports developing automated methods that provide faster, more consistent localization [Baumrind y Frantz, 1971].

Methodologically, recent advances in deep learning have improved automatic cephalometric analysis [Schwendicke et al., 2021], but several limitations remain. First, landmarks differ in their detectability, and global average scores may conceal poor performance on clinically important landmarks with indistinct anatomical boundaries [Durão et al., 2015; Hendrickx et al., 2024]. Second, many studies rely on relatively small medical imaging datasets, increasing the risk of overfitting and making performance estimates sensitive



(a): 1. SNA 2. SNB 3. Wits (distance between the A perpendicular and the B perpendicular on the occlusal plane) 4. MMA 5. ISN 6. IMPA 7. II 8. E-line upper lip 9. E-line lower lip 10. NLA 11. IPP

(b): 12. SN-Mand 13. Y-axis 14. SNP 15. SNO 16. McNamara to point A 17. McNamara to Pogonion 18. PFH (posterior facial height) 19. AFH (anterior facial height) [Jarabak ratio is PFH/AFH]

Figure 1.3: *Example of manual cephalometric landmark annotation and reference-plane construction on a lateral cephalogram. Hard-tissue landmarks (S, N, Na, Or, Po, Ar, Ba, ANS, A, B, Pog, Me, Go) and soft-tissue landmarks (Sn, UL, LL, Pn, G, N', A', B', Pog', Me') are identified and connected to form reference planes and angular measurements used in orthodontic diagnosis. The illustration highlights the density of anatomical structures that must be located accurately and, consequently, the complexity and subjectivity inherent in manual tracing [Zaheer et al., 2024].*

to data partitioning and validation protocols [Nan et al., 2023; Dot et al., 2022]. Third, the reported errors are not always clinically interpretable, especially if performance metrics are expressed solely in pixels rather than millimeters [Schwendicke et al., 2021; Mesquita et al., 2023].

The quality of a cephalometric localization pipeline therefore depends on more than the network architecture. It also depends on how landmarks are represented during training, how coordinates are recovered from model outputs, how image scaling is handled, and how performance is evaluated under limited data availability [Dot et al., 2022; Schwendicke et al., 2021]. A reliable framework should integrate these interdependent components rather than treat the neural network as an isolated solution.

1.3 Research Problem

Deep learning-based methods have advanced automatic cephalometric landmark localization, yet the literature still lacks a comprehensive account of how to design a framework that is accurate, robust, clinically interpretable,

and experimentally reliable [Sitaras et al., 2025]. The challenge extends beyond predicting landmark coordinates from radiographic images. It also involves selecting an appropriate target representation, extracting coordinates consistently, converting localization errors into clinically meaningful units, and evaluating generalization rigorously when only a limited number of annotated images is available [Hendrickx et al., 2024; Dot et al., 2022].

The research problem addressed in this thesis is therefore formulated as follows:

> *How can a deep learning-based localization framework be designed to achieve accurate spatial predictions, rigorous evaluation under limited data availability, reproducible experimentation, and clinically meaningful performance assessment?*

This thesis addresses the problem by jointly examining target representation, coordinate extraction, network design, training strategy, and evaluation. The primary evaluation uses the Holdout 70/30 protocol, which reserves 30 radiographs as a fully isolated held-out test set while model development proceeds on the remaining 72 radiographs. The K-Fold 70/30 protocol, with heatmap-level ensemble inference, and the Nested Leave-One-Out Cross-Validation protocol, with 102 outer folds, are included as additional validation and robustness analyzes of the same localization framework rather than as alternative primary protocols.

1.4 Research Questions

The general research problem gives rise to three research questions that guide the design and evaluation of the proposed localization framework:

1. How can dense Gaussian landmark heatmaps be transformed into continuous coordinate estimates through a mathematically consistent decoding procedure?
2. Can a lightweight two-level U-Net achieve sufficient representational capacity for accurate Gaussian heatmap regression on a limited cephalometric dataset?
3. How can a primary Holdout 70/30 protocol, complemented by a K-Fold 70/30 ensemble evaluation and a Nested Leave-One-Out Cross-Validation (Nested LOOCV) robustness analysis, characterize generalization and performance stability under limited data availability?

These questions correspond to the three research hypotheses of Section 1.7 and are addressed according to the objectives in Section 1.6.1.

1.5 Scope of this Study

This thesis focuses on automatically localizing cephalometric landmarks in two-dimensional lateral cephalometric radiographs using supervised deep learning. The scope is limited to static image analysis and excludes real-time

deployment, multimodal image fusion, and three-dimensional reconstruction.

The proposed localization framework is not intended to replace clinical expertise. Instead, it supports clinical decision-making by improving the consistency, reproducibility, and efficiency of landmark localization. The methodology centers on Gaussian heatmap representation, continuous coordinate estimation, a lightweight U-Net architecture, a primary Holdout 70/30 evaluation supported by complementary K-Fold and Nested LOOCV experiments, and clinically meaningful performance assessment. Accordingly, the study prioritizes methodological rigor, interpretability, and reproducibility over architectural complexity or novelty alone.

1.6 Research Objectives

The primary objective of this research is to adapt a U-Net-based framework for cephalometric landmark localization by coordinating Gaussian heatmap regression, continuous coordinate decoding, image scaling, model training, internal training-duration selection, refitting, validation, and evaluation in clinically meaningful units while retaining a lightweight network architecture.

1.6.1 Specific Objectives

1. To implement a differentiable Soft-Argmax coordinate-decoding module that transforms dense landmark heatmaps into continuous coordinate estimates.
2. To formulate cephalometric landmark localization as a Gaussian heatmap-regression problem that provides spatially distributed supervision around each annotated landmark.
3. To integrate a lightweight two-level U-Net into the proposed framework and evaluate whether it provides sufficient representational capacity for accurate Gaussian heatmap regression.
4. To evaluate the complete Pekin University (PKU) cephalometric dataset using the primary Holdout 70/30 protocol, with 72 radiographs for model development and 30 radiographs as a fully isolated test set, and to report localization performance in millimeters using image-specific anisotropic conversion.
5. To conduct the complementary K-Fold 70/30 ensemble experiment, in which ten fold-specific models are trained on the development data, their predicted heatmaps are averaged before Soft-Argmax coordinate extraction, and the ensemble is evaluated on the same 30-image external test cohort, and the Nested LOOCV robustness experiment, in which each of the 102 radiographs is excluded once from model development, the training duration is selected by inner validation, a newly initialized U-Net is refitted on the remaining 101 radiographs, and the excluded radiograph is evaluated exactly once.
6. To position the performance of the framework relative to existing methods using Mean Radial Error and Successful Detection Rate on the PKU cephalometric dataset, a benchmark collection of dental X-ray images, available at Figshare (PKU Cephalometric Dataset).

1.7 Research Hypotheses

The proposed methodology is based on three research hypotheses.

Hypothesis 1: Regression strategy.

Gaussian heatmap regression, combined with coordinate extraction through the differentiable Soft-Argmax operator, provides a coherent framework for cephalometric landmark localization [Newell et al., 2016; Payer et al., 2019]. The heatmap representation supplies spatially distributed supervision around each annotated landmark, whereas Soft-Argmax converts the predicted response maps into continuous coordinate estimates. This formulation should support accurate localization in regions with diffuse boundaries or overlapping radiographic structures [Thaler et al., 2021].

Hypothesis 2: Architecture.

The proposed lightweight two-level U-Net provides sufficient representational capacity for accurate cephalometric landmark localization [Ronneberger et al., 2015]. Reducing the network depth decreases parametric and computational complexity while preserving the multiscale contextual information required for precise landmark localization. This architectural simplification is expected to act as an implicit regularizer and thereby improve generalization when training data are limited [Rashmi y Ashok, 2020].

Hypothesis 3: Training and validation strategy.

The primary Holdout 70/30 protocol evaluates the framework on 30 fully isolated external test radiographs, while all model-development decisions, including training-duration selection, remain confined to the 72 development radiographs. The complementary K-Fold 70/30 ensemble experiment, evaluated on the same 30-image external cohort, and the Nested LOOCV experiment, which excludes each radiograph once from all model-development decisions before a single refit-based evaluation, provide additional validation and robustness evidence. Together, these protocols characterize generalization and performance stability under limited data availability while preventing the external test data from influencing any model-development decision.

1.8 Main Contributions

The principal contribution of this thesis is the systematic re-engineering of a U-Net-based methodological framework for automatic cephalometric landmark localization. The framework integrates target representation, coordinate extraction, network design, internal training-duration selection, full development-set refitting, a primary Holdout 70/30 evaluation, complementary validation experiments, and clinically meaningful evaluation within a coherent workflow.

The main contributions are as follows:

1. **An integrated Gaussian heatmap-regression framework with differentiable coordinate extraction.**

The thesis connects dense Gaussian landmark heatmaps with continuous coordinate estimation through the Soft-Argmax operator. The network is optimized through heatmap regression, while Soft-Argmax provides a continuous coordinate representation for evaluation.

2. **Integration of a lightweight U-Net into the localization framework.** This work integrates a two-level U-Net into the re-engineered pipeline. The resulting framework preserves the multiscale contextual information required for anatomical landmark detection while reducing computational complexity, model size, and the risk of overfitting on small medical datasets.

3. **A primary Holdout 70/30 evaluation with explicit refitting.** The principal evaluation reserves 30 radiographs as a fully isolated test set. Model development proceeds on the remaining 72 radiographs: the internal validation subset determines the training duration; a newly initialized U-Net is then refitted from scratch on all 72 development radiographs for the selected number of epoadd, and the external test radiographs are evaluated once all model development decisions have been completed. The localization errors are converted into millimeters using image-specific anisotropic conversion factors.

4. **Complementary K-Fold 70/30 ensemble and Nested LOOCV robustness experiments.** In the K-Fold 70/30 experiment, we train ten instances of the same U-Net on development data and combine their heatmaps before coordinate extraction for ensemble inference on the same 30 images held-out. In Nested LOOCV, we evaluate all radiographs: in each of the 102 outer folds, the inner validation stage determines training duration; a newly initialized U-Net is fitted to all 101 remaining radiographs, and the excluded radiograph is evaluated once by that fold-specific model. These experiments are complementary robustness analyzes rather than alternative primary protocols.

5. **Matched statistical comparison and literature positioning.** The three protocols are compared through paired statistical analyzes on the 30 common external radiographs, combining paired hypothesis tests with effect-size measures and McNemar analyzes of landmark-level classifications. The proposed framework is also positioned relative to representative methods on a public cephalometric benchmark using Mean Radial Error and Successful Detection Rate. This comparison assesses the performance of a rigorously engineered pipeline built around a comparatively simple network architecture.

Overall, this thesis examines whether a carefully engineered localization framework can achieve competitive performance with a simple architecture and a reproducible evaluation workflow. Its methodological emphasis is placed on the interaction between spatial representation, coordinate decoding, training control, physical error conversion, and validation design.

1.9 Manuscript Organization

This thesis is organized as follows:

- **Chapter 1** introduces the clinical and scientific context of cephalometric analysis, defines the research problem, justifies the study, and presents the research objectives, hypotheses, and main contributions.
- **Chapter 2** reviews the state of the art in automatic cephalometric landmark localization. It covers the field's clinical foundations and methodological development, including conventional image-analysis techniques, deep learning approaches, Gaussian heatmap regression, U-Net architectures, differentiable coordinate extraction, validation strategies, clinically relevant evaluation metrics, and the limitations of existing methods.
- **Chapter 3** presents the proposed methodology. It describes the dataset, preprocessing procedures, Gaussian heatmap generation, lightweight U-Net architecture, Soft-Argmax coordinate extraction, the primary Holdout 70/30 protocol, internal training-duration selection, refitting on the 72-image development set, the complementary K-Fold 70/30 ensemble experiment, and Nested LOOCV robustness protocol, ensemble inference, and quantitative and qualitative evaluation.
- **Chapter 4** presents the primary Holdout 70/30 results and the complementary K-Fold ensemble and Nested LOOCV evaluations. This chapter discusses localization performance, cross-protocol comparisons, paired statistical analyzes on the 30 common external radiographs, and the pipeline's consistency across the three validation settings.
- **Chapter 5** summarizes the main findings, reviews the limitations of the research, and outlines directions for future research.

1.10 Chapter Summary

This introductory chapter has established the clinical background, scientific motivation, and research problem addressed in this thesis. Automatic cephalometric landmark localization is not just a coordinate prediction task. It requires accurate spatial prediction, reproducible experimentation, and a clinically meaningful interpretation of localization error.

Although deep learning has advanced significantly in the field, significant challenges remain: methodological rigor, anatomical landmark variability, and the reliable translation of prediction accuracy into clinical measurements. These challenges are particularly acute when the available radiographic dataset is limited.

They motivate an integrated approach in which target representation, coordinate estimation, network architecture, training control, physical error conversion, and evaluation are considered interrelated components. Therefore, the primary evaluation uses the Holdout 70/30 protocol: 30 radiographs are reserved as a fully isolated test set, the internal validation subset determines the training duration within the 72-image development set, a newly initialized model is refitted on all 72 development radiographs, and the external test radiographs are evaluated once. The K-Fold 70/30 ensemble and Nested LOOCV experiments are retained as complementary validation and robustness analyzes to assess the stability of the same localization pipeline under alternative model-development and evaluation designs.

Chapter 2

State of the Art

The previous chapter established the clinical and scientific context of this thesis and defined the research problem and objectives. This chapter, therefore, reviews the current state of the art in automatic cephalometric landmark localization and examines the relevance of this thesis to the literature.

The chapter aims to achieve two goals. First, it traces the development of cephalometric landmark localization from traditional image processing to deep learning approaches. Second, it outlines the methodological challenges that inform the design choices of this thesis: Gaussian heatmap regression, continuous coordinate extraction, millimeter-based evaluation, a primary Holdout 70/30 evaluation, complementary K-Fold 70/30 ensemble and Nested Leave-One-Out Cross-Validation (Nested LOOCV) robustness experiments, and matched statistical comparison on the common external radiographs.

The discussion moves from clinical cephalometry to automated landmark localization and finally to a rigorous evaluation approach. Progress through these steps provides a unified view of the field and demonstrates the necessity of a unified approach to spatial accuracy, reproducibility, and clinical interpretability.

2.1 Lateral Cephalometric Radiography

Lateral cephalometric radiography produces a two-dimensional (2D) projection of the three-dimensional (3D) craniofacial anatomy. Broadbent introduced this imaging modality in 1931, and it remains widely used in orthodontics and dentofacial orthopedics because it provides standardized image acquisition, a relatively low radiation dose, broad availability, and cost-effective clinical assessment [Broadbent, 1931; Izzetti et al., 2026].

A lateral cephalogram projects anatomical structures located at different spatial depths onto a single image plane. As shown in Figure 2.1, this projection can cause bilateral craniofacial structures, including the orbital rims, mandibular borders, and regions surrounding the external auditory meatus, to overlap within the radiograph.

The resulting image represents the craniofacial complex in two dimensions but also introduces anatomical superimposition and geometric distortion.



Figure 2.1: *Lateral radiograph of a human skull, commonly referred to as a lateral cephalogram [Izzetti et al., 2026].*

Lateral cephalometric radiography offers a key clinical advantage because it supports standardized, reproducible examinations of craniofacial morphology. Clinicians use these examinations for orthodontic diagnosis, treatment planning, growth assessment, and longitudinal follow-up [Broadbent, 1931].

However, projecting three-dimensional anatomy onto a two-dimensional image inevitably removes spatial information. Anatomical structures that occupy different depths may overlap in the projection, obscure one another, and increase uncertainty during landmark localization. Magnification, patient-positioning errors, and geometric distortion may further reduce the accuracy and reproducibility of the measurements derived from the radiograph [Izzetti et al., 2026].

Three-dimensional imaging modalities, particularly cone-beam computed tomography (CBCT), provide volumetric information. Nevertheless, their higher cost, greater radiation exposure, and more limited accessibility prevent them from replacing lateral cephalometric radiography in routine practice [Pauwels, 2015; Kaasalainen et al., 2021]. For this reason, most publicly available benchmark datasets and automated landmark-localization studies continue to use lateral cephalograms.

Lateral cephalograms therefore remain a standard imaging modality for routine cephalometric assessment and the principal data source for research on automatic cephalometric landmark localization.

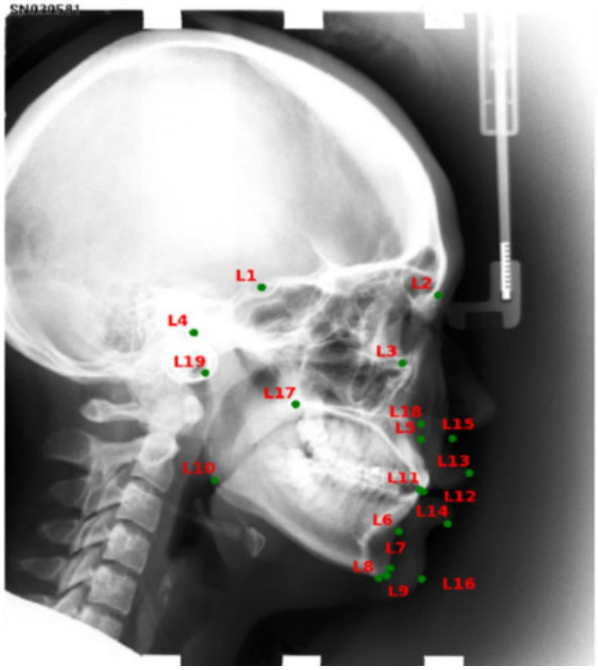
2.2 Cephalometric Analysis in Orthodontics

Cephalometric analysis forms the foundation of orthodontic diagnosis. Clinicians identify anatomical landmarks on standardized radiographs of the craniofacial complex and use their coordinates to calculate linear and angular measurements that describe skeletal, dental, and soft-tissue relationships. These measurements help clinicians quantify maxillomandibular discrepancies, assess craniofacial growth, and plan orthodontic and orthognathic treatments [Juneja et al., 2021; Fernández Sánchez, 2006].

Cephalometry is not just visual interpretation. It follows a structured measurement process with a set of anatomical landmarks that have clinical roles and which are classified by researchers according to their anatomical origin. Some, such as the nasion, point A, and sella, can be identified directly on bony structures, whereas others, such as the gnathion, result from tracing procedures. Researchers also classify landmarks by tissue type, including hard-tissue landmarks on the mandible and maxilla and soft-tissue landmarks on structures such as the lips and nose [Šavc et al., 2022].

This anatomical diversity helps explain why localization difficulty and reproducibility vary across landmarks. Landmarks located on clearly defined bony contours are generally easier to identify than those located in low-contrast regions, areas affected by anatomical overlap, or soft-tissue profiles.

The present thesis uses the 19-landmark annotation protocol associated with the PKU dataset. Figure 2.2 shows their spatial distribution, and these landmarks define the prediction targets used throughout the study.



Serial	Anatomical Landmarks
L1	Sella
L2	Nasion
L3	Orbitale
L4	Porion
L5	Subspinale
L6	Supramentale
L7	Pogonion
L8	Menton
L9	Gnathion
L10	Gonion
L11	Lower incisal incision
L12	Upper incisal incision
L13	Upper lip
L14	Lower lip
L15	Subnasal
L16	Soft tissue pogonion
L17	Posterior nasal spine
L18	Anterior nasal spine
L19	Articulate

Figure 2.2: *Spatial distribution of the 19 cephalometric landmarks considered in this study [Sumon et al., 2025].*

Automatic cephalometric landmark localization must do more than identify isolated points in an image. The predicted positions must be sufficiently accurate and reproducible to preserve the reliability of the linear and angular measurements derived from them [Juneja et al., 2021]. This requirement is particularly important because clinical decisions may depend on differences of only a few millimeters or degrees.

The abbreviations, assigned names, and anatomical definitions of the 19 landmarks used in this study are summarized in Table 2.1.

Table 2.1. *Definition of the 19 cephalometric landmarks considered in this study, including their abbreviations, assigned names, and anatomical descriptions [Juneja et al., 2021].*

No.	Abbrev.	Assigned name	Landmark description
L01	S	Sella	Center of the sella turcica within the sphenoid bone.
L02	N	Nasion	Most anterior point of the frontonasal suture in the midsagittal plane.
L03	O	Orbitale	Lowest point on the inferior orbital rim.
L04	Po	Porion	Most superior point of the external auditory meatus.
L05	A	Point A	Deepest point on the anterior contour of the maxillary alveolar process between the anterior nasal spine and prosthion.
L06	B	Point B	Deepest point on the anterior contour of the mandibular alveolar process between infradentale and pogonion.
L07	Pog	Pogonion	Most anterior point on the bony contour of the chin.
L08	Me	Menton	Lowest point on the mandibular symphysis.
L09	Gn	Gnathion	Most anteroinferior point on the bony contour of the chin.
L10	Go	Gonion	Most posteroinferior point at the angle of the mandible.
L11	Ii	Lower incisor	Incisal edge of the most prominent lower central incisor.
L12	Is	Upper incisor	Incisal edge of the most prominent upper central incisor.
L13	Ul	Upper lip	Most anterior point of the upper lip.
L14	Li	Lower lip	Most anterior point of the lower lip.
L15	Sn	Subnasale	Point at which the lower border of the nasal septum meets the upper lip.
L16	Pos	Soft-tissue pogonion	Most anterior point on the soft-tissue contour of the chin.
L17	PNS	Posterior nasal spine	Most posterior point of the hard palate in the midsagittal plane.
L18	ANS	Anterior nasal spine	Most anterior point of the bony anterior nasal spine.
L19	Ar	Articulare	Intersection of the posterior border of the mandibular ramus and the inferior border of the cranial base.

Together, these landmark definitions establish the anatomical targets used in the automated localization framework and provide the basis for the methodological discussion that follows.

2.3 Cephalometric Analysis by Manual Localization

For many years, trained experts performed manual landmark localization as the standard approach to cephalometric analysis. As illustrated in Figure 2.3, examiners traditionally traced anatomical structures, reference planes, and angular relationships on tracing paper placed over a lateral cephalometric radiograph. They then used these constructions to identify anatomical landmarks and derive the measurements required for diagnosis and treatment planning.

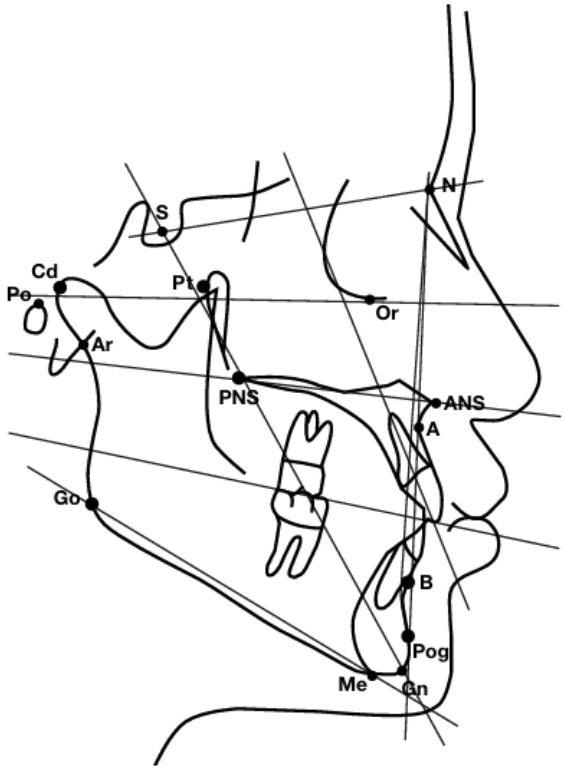


Figure 2.3: *Example of conventional manual localization and tracing of cephalometric landmarks on a lateral radiograph [Bansal et al., 2012].*

Manual cephalometric tracing remains clinically established, but it is time-consuming, requires trained personnel, and depends on the examiner's experience, judgment, and ability to maintain attention throughout the procedure. Examiner fatigue and variations in visual interpretation can further affect the consistency of the resulting annotations [Kim et al., 2020; Yang et al., 2023].

Researchers generally distinguish two principal sources of annotation variability. Intra-observer variability occurs when the same examiner assigns different coordinates to the same radiograph during separate annotation sessions. Inter-observer variability occurs when different examiners assign different coordinates to the same image.

The tracing procedure shown in Figure 2.3 illustrates both the complexity of manual analysis and the

uncertainty associated with landmark annotation. This variability affects not only routine clinical measurements but also the development and evaluation of automated cephalometric systems.

Automatic landmark localization addresses a practical clinical need by reducing tracing time and improving the standardization and reproducibility of cephalometric analysis [Leonardi et al., 2008]. However, the variability of manual annotations also shows that ground-truth coordinates do not represent an error-free anatomical reference. Instead, they constitute expert estimates that may vary across observers and annotation sessions [Baumrind y Frantz, 1971; Durão et al., 2015].

Supervised learning models therefore learn to approximate expert-provided reference annotations rather than an absolute anatomical truth. This consideration supports the use of spatially distributed landmark representations, clinically meaningful error thresholds, landmark-specific analyzes, and rigorous statistical evaluation procedures [Payer et al., 2019; Juneja et al., 2021].

2.4 Classical Landmark Detection Approaches

Before deep learning became widespread, researchers developed automatic cephalometric landmark-localization systems primarily from knowledge-based image-processing methods and conventional machine-learning techniques. These systems used handcrafted image features, edge and contour detection, template matching, Active Shape Models (ASMs), Active Appearance Models (AAMs), and statistical classification or regression methods [Leonardi et al., 2008].

Early approaches helped formalize automatic landmark localization by incorporating explicit anatomical constraints, geometric relationships, and statistical representations of craniofacial shape. Nevertheless, anatomical variability, low image contrast, structural overlap, image noise, and the difficulty of designing handcrafted features that remain robust across radiographs and patient populations often limited their performance [Juneja et al., 2021].

The principal families of automatic landmark-localization methods are summarized in Figure 2.4. The flowchart distinguishes knowledge-based techniques from artificial intelligence-based approaches and further divides the latter into conventional machine learning and deep learning methods.

Knowledge-based techniques locate landmarks by applying predefined anatomical information, geometric constraints, and handcrafted image-processing rules [Neeraja y Anbarasi, 2025]. Examples include edge and pattern detection, contour extraction, template matching, genetic programming, and statistical shape representations such as ASMs and AAMs [Juneja et al., 2021].

Artificial intelligence-based techniques learn discriminative landmark characteristics from annotated data [Neeraja y Anbarasi, 2025]. Conventional machine-learning methods generally combine manually designed image features with classifiers or regression models, including random forests, support vector machines, decision trees, linear discriminant analysis, and principal component analysis [Schwendicke et al., 2020].

Deep learning methods take a different approach by learning hierarchical image representations directly from

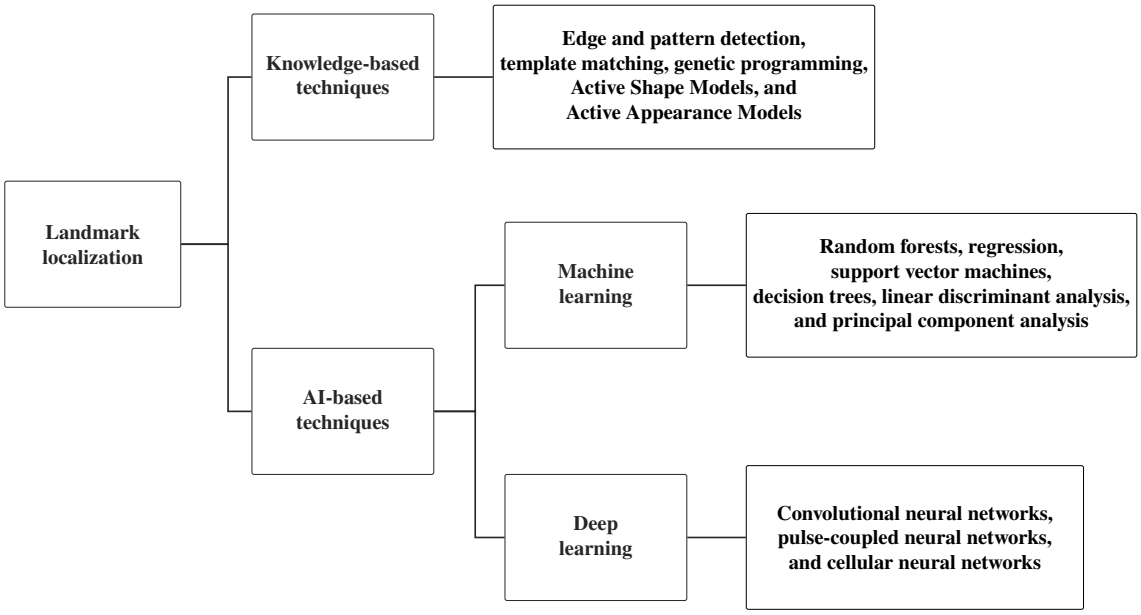


Figure 2.4: *Classification of the principal methodological families used for automatic cephalometric landmark localization, adapted from the categorization reported by Juneja et al. [2021].*

the training data. As a result, they have become a dominant approach in recent cephalometric landmark-localization research [Schwendicke et al., 2021; Neeraja y Anbarasi, 2025]. Most contemporary systems rely on convolutional neural networks, although earlier studies also investigated pulse-coupled neural networks for automated cephalometric landmark detection [Innes et al., 2002].

This transition from knowledge-based techniques to data-driven deep learning reflects the need for models that can achieve greater localization accuracy, adapt to anatomical variability, and reduce dependence on manually engineered features [Izzetti et al., 2026]. The following subsections examine the principal stages of this methodological development in greater detail.

2.4.1 Knowledge-Based Methods and Contour Detection

Knowledge-based and contour-detection techniques represent some of the earliest approaches to automatic cephalometric landmark localization. Researchers designed these methods around predefined anatomical and geometric properties that described the expected shape and approximate position of craniofacial structures.

These systems inferred landmark positions through handcrafted image-processing operations, including edge detection, contour extraction, template matching, geometric constraints, and anatomical heuristics [Leonardi et al., 2008; Juneja et al., 2021].

For example, Liu et al. [2000] proposed an edge-based method to identify 13 landmarks on 10 lateral cephalograms. The system successfully detected only 5 landmarks, illustrating the limitations of relying exclusively on predefined edge patterns and anatomical rules.

Similarly, Grau et al. [2001] introduced a two-stage framework that combined template matching, edge detection, and contour segmentation. The authors reported that approximately 90 % of the evaluated landmarks fell within the clinically acceptable error interval on a dataset of 20 cephalograms.

Knowledge-based techniques offer an important advantage through their interpretability because explicit anatomical and geometric rules govern the localization process. However, these rules limit the adaptability of the methods. Anatomical variability, radiographic noise, low contrast, overlapping structures, positioning errors, and atypical craniofacial morphology can all reduce their performance.

2.4.2 Shape-Based Models: ASM and AAM

Active Shape Models introduced a statistical approach to anatomical landmark localization by representing the variation of landmark configurations observed in annotated training images [Cootes et al., 1995]. During fitting, the model constrains the estimated landmark configuration to remain within the plausible shape space learned from the training data [Leonardi et al., 2008].

Active Appearance Models extended this principle by jointly modeling anatomical shape and image appearance. They combine landmark geometry with local intensity or texture information within a unified statistical framework [Cootes et al., 2001]. This design allows AAM-based methods to exploit both the spatial relationships among landmarks and the visual characteristics of the surrounding anatomy.

Several studies have applied statistical shape models to cephalometric localization. Saad et al. [2006] proposed an automatic landmark-localization method that combined an Active Appearance Model (AAM) with simulated annealing. Using 20 cephalograms for training and 7 for testing, the authors reported an average landmark-localization error of approximately 3.1 mm. In a three-dimensional setting, Montúfar et al. [2018] trained a statistical shape model on cone-beam computed tomography scans and reported a mean localization error of 3.64 mm.

Shape-based methods were generally more robust than methods based solely on local edges or contours because they took global anatomical consistency into account. Their performance still depended on the size, diversity and representativeness of training data used to build the statistical shape space; anatomies which the training set did not well represent could thus be fitted incorrectly.

2.4.3 Random Forests and Classical Machine Learning

Conventional machine learning was a further step towards data-driven landmark localization. Instead of using fixed anatomical rules alone they learned relationships between handcrafted image features and landmarks positions from annotated examples [Leonardi et al., 2008].

Random forests became particularly important because they can model nonlinear relationships, heterogeneous feature sets and train and run inference efficiently on moderately sized datasets.

Among classical machine-learning methods, Lindner et al. [2016] reported one of the strongest performances in the ISBI 2015 cephalometric landmark-detection challenge. Their method achieved a Successful Detection Rate of 74.84 % at the 2 mm threshold on the ISBI 2015 test set, as reported in the challenge literature [Song et al., 2020]..

Using Haar-like features, random forest regression, and a global shape model, Ibragimov et al. [2014] achieved the second-highest Successful Detection Rate of 71.7 % at the same threshold in the same challenge [Song et al., 2020].

Despite these improvements, conventional machine-learning systems remained heavily dependent on handcrafted representations, including texture descriptors, intensity differences, local gradients, Haar-like features, and geometric measurements [Rundo y Militello, 2024]. Their performance consequently depended on the suitability of the selected features and could decline on anatomically heterogeneous or clinically challenging radiographs.

These limitations encouraged the development of deep learning approaches that learn hierarchical feature representations directly from radiographic images, reducing reliance on manual feature engineering.

2.5 Deep Learning for Medical Imaging

Deep learning has transformed medical image analysis by enabling the automatic learning of hierarchical feature representations directly from image data, thereby reducing the need for handcrafted feature engineering. This is useful in medical imaging because anatomical structures have complicated spatial relationships, large inter-patient variability and subtle visual patterns that manually designed descriptors can't capture [Litjens et al., 2017; Shen et al., 2017].

Convolutional Neural Networks (CNNs) are at the heart of many of these advances. CNNs learn increasingly abstract image representations with successive convolutional layers and by capturing hierarchical spatial features CNN-based methods have achieved state-of-the-art performance in medical image classification, segmentation, object detection, feature extraction and anatomical landmark localization [Lindroth et al., 2024; Ma et al., 2024].

Before we talk about CNNs, it is important to review the computational model of deep neural architectures: the artificial neural network (ANN). How individual neurons process information and how they build up layered architectures is what conceptualizes the convolutional models introduced in the following sections.

2.5.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models that are inspired by the organization of biological neurons. Each artificial neuron receives multiple input signals, computes their weighted sum, applies a nonlinear activation function, and transmits the resulting output to subsequent neurons.

Mathematically, the output of a neuron is given according to Eq. (2.1):

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{2.1}$$

where x_i denotes the input signals, w_i denotes the corresponding learnable weights, b represents the bias term, and $f(\cdot)$ denotes the nonlinear activation function. Eq. (2.1) shows that the neuron first combines the input signals through a weighted sum and then applies the activation function to introduce nonlinearity into the model.

Figure 2.5 illustrates the operation of an artificial neuron. The input signals are multiplied by their respective weights, combined with the bias term, passed through the activation function, and transformed into the neuron output.

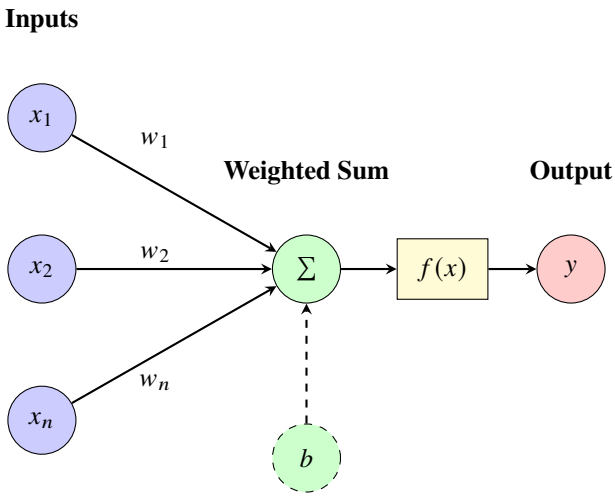


Figure 2.5: *Architecture of an artificial neuron. Input signals x_i are multiplied by their corresponding learnable weights w_i, combined with the bias term b, and transformed by the nonlinear activation function $f(\cdot)$ to produce the output y.*

Although multilayer perceptrons (MLPs) perform well on structured numerical data, their fully connected architecture becomes computationally prohibitive for high-resolution medical images. In addition, flattening an image into a one-dimensional vector discards spatial relationships between neighboring pixels, limiting the suitability of MLPs for image analysis. These limitations motivated the development of Convolutional Neural Networks, which preserve and exploit spatial structure through local connectivity and weight sharing [LeCun et al., 1998; Goodfellow et al., 2016].

2.5.2 Biological Inspiration for CNNs

The conceptual foundations of Convolutional Neural Networks (CNNs) trace back to the pioneering neurophysiological studies of Hubel and Wiesel on the mammalian visual cortex. Their experiments showed that neurons in the primary visual cortex respond selectively to local visual patterns, including oriented edges,

lines, and simple textures. At later stages of visual processing, neurons integrate these elementary responses to represent increasingly complex visual structures [Hubel y Wiesel, 1962].

Inspired by these biological observations, Fukushima proposed the Neocognitron, an early hierarchical convolutional architecture designed to recognize visual patterns while remaining robust to small spatial transformations [Fukushima, 1980]. LeCun and colleagues later developed LeNet-5, which served as the core convolutional network in a complete bank-check recognition system that entered commercial use and processed millions of checks per month [LeCun et al., 1998].

Figure 2.6 illustrates the hierarchical organization of visual processing that inspired the design of CNNs. It depicts successive processing stages, beginning with local luminance contrasts in the retina and progressing through orientation-selective simple cells, position-tolerant complex cells, and shape-selective hypercomplex cells. This progression provides a biological analogy for the way CNNs learn increasingly abstract representations through successive layers.

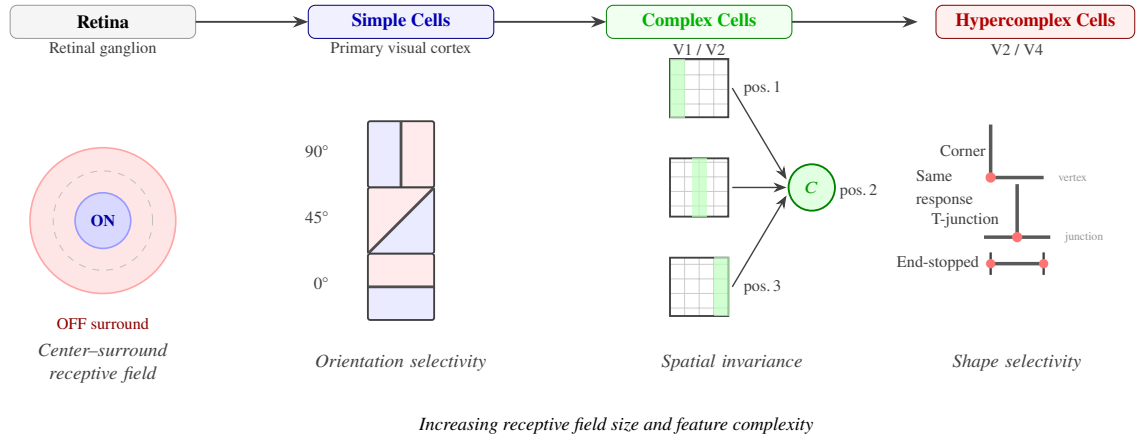


Figure 2.6: *Hierarchical organization of visual processing inspired by the experiments of Hubel and Wiesel [Hubel y Wiesel, 1962]. **Retina:** retinal ganglion cells exhibit center–surround receptive fields with an excitatory ON center (blue) and an inhibitory OFF surround (red), enabling the detection of local luminance contrasts. **Simple cells** in the primary visual cortex respond selectively to oriented edges at specific angles (0°, 45°, and 90°), with distinct ON (blue) and OFF (red) subregions within their receptive fields. **Complex cells** pool inputs across nearby spatial positions, producing responses that are tolerant to variations in the exact location of a stimulus. **Hypercomplex cells** respond to more elaborate features, including corners, T-junctions, and line endings, and may exhibit end-stopping behavior. This hierarchical progression from local contrast detection to increasingly abstract feature representations helped inspire the layered architecture of convolutional neural networks [Fukushima, 1980; LeCun et al., 1998].*

To further emphasize this analogy, Table 2.2 summarizes the correspondence between the biological visual-processing hierarchy described by Hubel and Wiesel and its computational counterparts in CNN architectures. The comparison highlights the roles of local feature extraction, hierarchical abstraction, pooling, and tolerance to spatial variation in both biological and computational vision [Hubel y Wiesel, 1962; Fukushima, 1980; LeCun et al., 1998; Litjens et al., 2017].

Table 2.2. Mapping between the hierarchical visual-processing stages described by Hubel and Wiesel [Hubel y Wiesel, 1962] and their computational counterparts in convolutional neural networks [Fukushima, 1980; LeCun et al., 1998; Litjens et al., 2017].

Biological stage	Function	CNN equivalent	Key property
Retinal ganglion cells	Center–surround contrast detection	First convolutional filters	Local edge sensitivity
Simple cells (V1)	Orientation-selective edge detection	Convolutional kernels (3×3, 5×5)	Linear feature extraction
Complex cells (V1/V2)	Position-tolerant pooling	Pooling layer (max or average)	Spatial tolerance
Hypercomplex cells (V2/V4)	Corner and junction detection	Deep convolutional layers	High-level feature abstraction
Cortical hierarchy	Progressive abstraction of visual information	Stacked convolutional and pooling blocks	End-to-end hierarchical learning

2.5.3 From Fully Connected Networks to CNNs

In a conventional fully connected network, also known as a multilayer perceptron (MLP), every input pixel is connected to every neuron in the first hidden layer. Consequently, an input image with dimensions $w \times h \times d$ provides $w \times h \times d$ input features to each neuron in that layer. For example, a single grayscale image with a resolution of 1920×1080 contains 2,073,600 pixels, resulting in millions of trainable parameters even in the first fully connected layer. Moreover, flattening the image into a one-dimensional vector does not explicitly preserve the spatial relationships between neighboring pixels. This representation therefore prevents the network from exploiting the local spatial structure that is fundamental to image analysis [Litjens et al., 2017].

Figure 2.7 illustrates the operation of a fully connected network applied to a small 4×4 input image. The figure shows how flattening converts the image into a one-dimensional vector and how each input element connects to every neuron in the subsequent layer.

As shown in Figure 2.7, fully connected networks are poorly suited to high-resolution image analysis because they do not explicitly encode local spatial relationships and require many trainable parameters as image resolution increases.

CNNs overcome these limitations with two basic principles: local connectivity and weight sharing. Instead of connecting every element in the input to every neuron, CNNs use small learnable filters (typically 3×3 or 5×5) that slide across the image. At each spatial location a filter computes dot product between its weights and the corresponding local image patch. Since the same filter is used over the whole image, CNNs preserve the spatial organization of the input but have much smaller number of learnable parameters [Fukushima, 1980; LeCun et al., 1989].

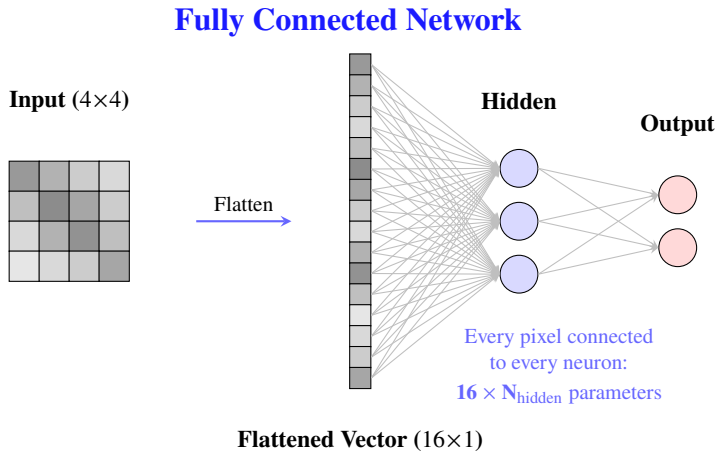


Figure 2.7: *Architecture of a fully connected network applied to a* 4×4 *image. The input is flattened into a* 16×1 *vector, which does not explicitly preserve the image's spatial organization. Each hidden neuron connects to every input element, so the number of parameters increases rapidly with input size.*

Figure 2.8 shows these two principles. Each output value is computed from a local 3×3 receptive field and the same filter coefficients are used at all spatial locations.

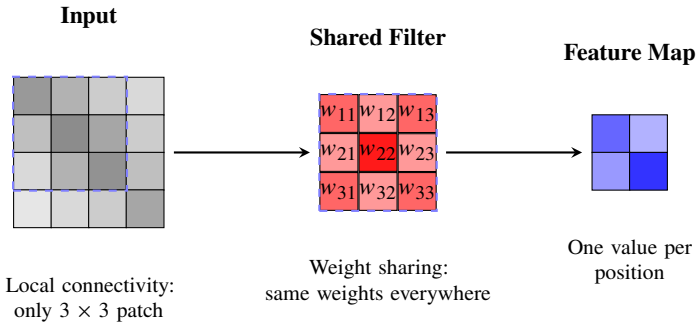


Figure 2.8: *Principles of convolutional layers. A shared* 3×3 *filter slides across the input image. At each position, the element-wise products between the filter coefficients and the corresponding image values are summed to produce one value in the output feature map. The same filter weights are reused across all spatial positions.*

These principles are consistent with the organization of the mammalian visual cortex described by Hubel and Wiesel and later informed the development of the Neocognitron and modern CNN architectures [Hubel y Wiesel, 1962; Fukushima, 1980; LeCun et al., 1989].

The quantitative implications of these architectural differences are summarized in Table 2.3. The comparison assumes a $256 \times 256 \times 3$ input image, 64 hidden neurons in the fully connected layer, and 64 convolutional filters with dimensions of 3×3.

As summarized in Table 2.3, the convolutional layer requires substantially fewer parameters than the fully connected layer while preserving the spatial topology of the input through local connectivity and weight sharing. Weight sharing also provides translation equivariance, meaning that a feature detected by a filter moves consistently

Table 2.3. Structural comparison between a fully connected network and a CNN for a $256 \times 256 \times 3$ *input image with 64 hidden neurons in the fully connected layer or 64 convolutional filters of size* 3×3*. The parameter counts include one bias term per neuron or filter.*

Property	Fully Connected	CNN
Input representation	Flattened vector ($196{,}608 \times 1$)	2D tensor ($256 \times 256 \times 3$)
Spatial structure	Not explicitly encoded	Preserved
Parameters (first layer)	$196{,}608 \times 64 + 64 = 12{,}582{,}976 \approx 12.6\,\text{M}$	$64 \times (3 \times 3 \times 3 + 1) = 1{,}792$
Receptive field	Global (entire image)	Local (3×3 patch)
Translation property	No built-in spatial equivariance	Translation equivariance before pooling

with the corresponding input pattern. Subsequent pooling or aggregation operations can then provide translation tolerance. These properties explain why CNNs have become a foundational architecture for medical image analysis and cephalometric landmark localization.

2.5.4 The Convolution Operation

The fundamental local operation performed by a CNN is two-dimensional discrete cross-correlation rather than convolution in the strict mathematical sense. Given an input image I and a filter K of size $F \times F$, the operation is defined according to Eq. (2.2):

$$S(i,j) = \sum_{u=0}^{F-1} \sum_{v=0}^{F-1} I(i+u, j+v)\, K(u,v), \tag{2.2}$$

Here, S denotes the output feature map, and (i, j) identifies the spatial position of the upper-left corner of the input region covered by the filter. Eq. (2.2) shows that each value in the output feature map results from the element-wise multiplication and summation of the filter coefficients and the corresponding local image patch.

Although the deep learning literature often calls this operation convolution, standard CNN implementations generally compute cross-correlation because they do not reverse the filter before applying it. Since the filter coefficients are learned during training, cross-correlation and strict convolution have the same representational capacity for CNN-based models, provided that the corresponding filter orientation is learned [Litjens et al., 2017].

Figure 2.9 illustrates this operation using a 3×3 Sobel filter applied to a 5×5 input image. The example highlights the concepts of receptive field, filter sliding, stride, and feature-map generation.

The numerical example presented in Figure 2.9 illustrates the computation performed at a single filter position. In practice, two additional hyperparameters determine how the filter is applied across the entire image and therefore control the spatial dimensions of the resulting feature map: padding and stride.

Padding adds zero-valued pixels around the input image borders before applying the filter. Without padding, the operation is called a valid convolution, and the output feature map is smaller than the input. With zero-padding

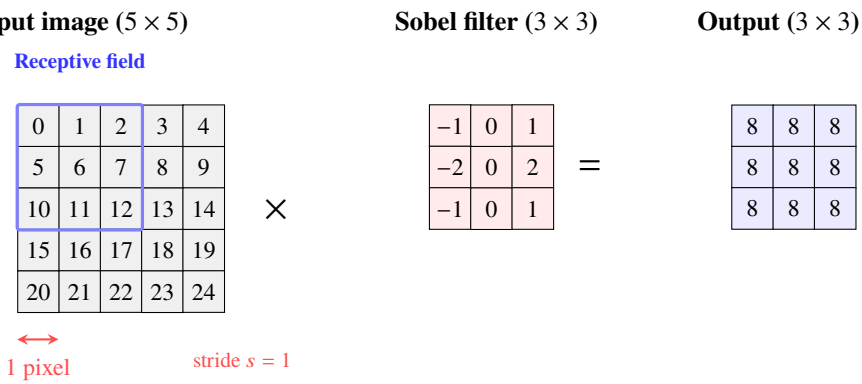


Figure 2.9: *Illustration of a two-dimensional cross-correlation operation, commonly referred to as convolution in deep learning, using a* 3×3 *Sobel filter applied to a* 5×5 *input image. The blue rectangle indicates the receptive field at the first filter position. With a stride of one and no padding, the output dimensions are* $(5 - 3 + 1) \times (5 - 3 + 1) = 3 \times 3$*. For the input values and Sobel filter shown, the operation produces* 8 *at each spatial position.*

of width p, the spatial dimensions can be preserved when the stride equals one, and the padding is selected appropriately [Litjens et al., 2017].

Stride specifies how many pixels the filter moves between successive operations. A stride of one produces the densest feature map, whereas larger strides reduce the spatial resolution and can therefore be used for downsampling. For an input image with dimensions $W \times H$, a filter of size $F \times F$, padding p, and stride s, the output dimensions are given according to Eq. (2.3):

$$W_{\text{out}} = \left\lfloor \frac{W - F + 2p}{s} \right\rfloor + 1, \qquad H_{\text{out}} = \left\lfloor \frac{H - F + 2p}{s} \right\rfloor + 1. \tag{2.3}$$

Here, W_{out} and H_{out} denote the width and height of the output feature map, respectively. The floor operator ensures that only complete filter placements are included. Eq. (2.3) is applied independently along the horizontal and vertical dimensions.

The effect of padding on the spatial dimensions of the output feature map is illustrated in Figure 2.10. The figure compares two common configurations: valid convolution, which applies no padding, and same convolution, which uses zero-padding to preserve the input's spatial dimensions when the stride equals one.

Together, convolution, stride, and padding determine the spatial resolution of the intermediate feature maps throughout a CNN. These operations form the computational basis of hierarchical feature extraction and are

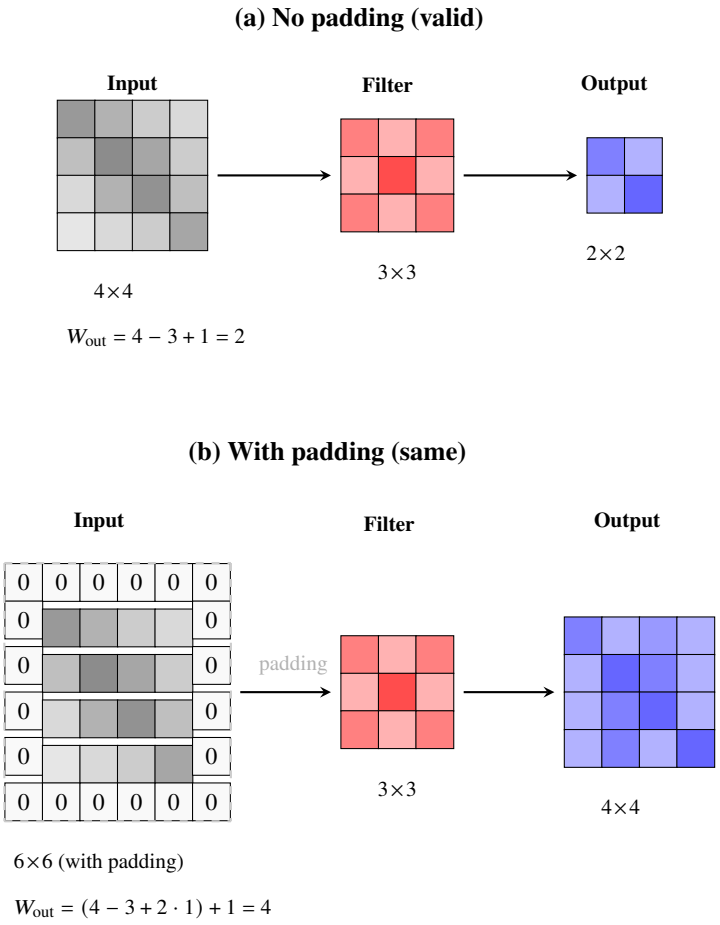


Figure 2.10: *Effect of padding on the spatial dimensions of the output feature map. (a) Valid convolution without padding reduces the output dimensions. (b) Same convolution with zero-padding preserves the spatial dimensions of the input feature map when the stride is equal to one.*

central to encoder-decoder architectures such as U-Net, which is introduced in the following subsection.

2.5.5 Activation Functions and Nonlinear Processing

A nonlinear activation function typically follows each convolutional operation and is applied element-wise to the resulting feature maps. This operation introduces nonlinearity into the network, enabling it to learn complex input–output relationships that purely linear transformations cannot represent. Without nonlinear activations, the composition of multiple convolutional layers would remain equivalent to a single affine transformation, thereby severely limiting the representational capacity of the network [Goodfellow et al., 2016; Litjens et al., 2017].

Among the various activation functions proposed in the literature, the Rectified Linear Unit (ReLU) is defined according to Eq. (2.4):

$$\mathrm{ReLU}(x) = \max(0, x). \tag{2.4}$$

Eq. (2.4) shows that ReLU suppresses negative inputs by mapping them to zero while preserving positive inputs unchanged. ReLU has become a standard choice for hidden layers in many modern CNN architectures because of its computational simplicity, efficient gradient propagation for positive inputs, and favorable optimization properties [Goodfellow et al., 2016]. However, a neuron can become effectively inactive when its inputs remain negative during training. In this case, the ReLU gradient is zero, preventing the neuron from updating and potentially causing the so-called *dying ReLU* problem.

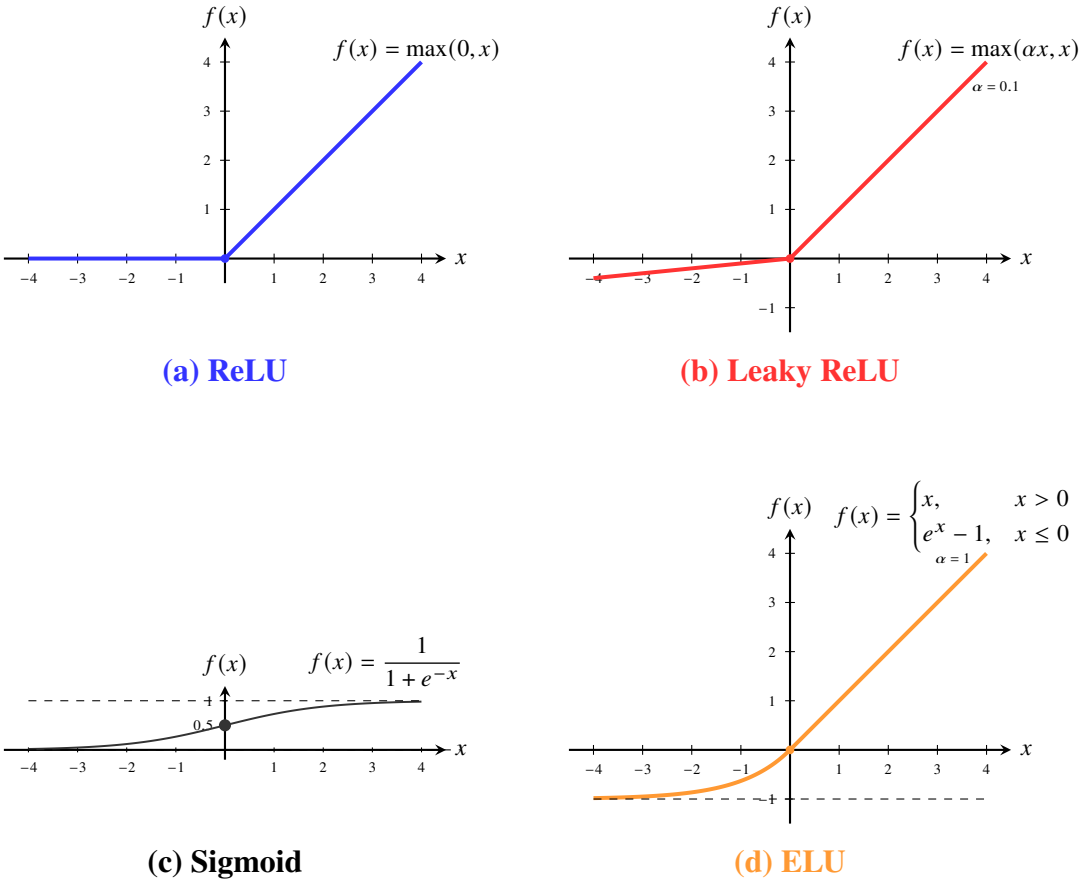
Several alternative activation functions have been proposed to reduce this limitation. Leaky ReLU introduces a small nonzero slope for negative inputs, thereby maintaining a gradient and reducing the risk of permanently inactive neurons. The parameter α controls the slope of the negative branch and is typically assigned a small positive value, such as $\alpha = 0.01$ or $\alpha = 0.1$. The Exponential Linear Unit (ELU) replaces the negative linear branch with a smooth exponential function. This formulation can improve optimization by allowing negative outputs and providing a smooth transition around the origin. In contrast, the Sigmoid activation maps its input continuously to the interval $(0, 1)$ and is therefore primarily used in binary classification output layers rather than in deep hidden layers, where its saturated gradients can cause vanishing-gradient problems [Litjens et al., 2017].

The most commonly used activation functions are illustrated in Figure 2.11. ReLU (Figure 2.11a) suppresses all negative responses while preserving positive activations unchanged. Leaky ReLU (Figure 2.11b) introduces a small negative slope α, preventing inactive neurons. Sigmoid (Figure 2.11c) maps the input continuously into the interval $(0, 1)$, whereas ELU (Figure 2.11d) combines an identity function for positive inputs with a smooth exponential response for negative values, improving optimization stability.

The main characteristics of these activation functions are summarized in Table 2.4. The comparison includes their mathematical definitions, derivatives, output ranges, and typical applications in convolutional neural networks.

For ReLU and Leaky ReLU, and for ELU when $\alpha \neq 1$, the derivative at $x = 0$ is convention-dependent because these functions are not differentiable at that point. Deep learning implementations typically assign a fixed subgradient value or use the derivative convention shown in Table 2.4.

Although several activation functions have been proposed, ReLU remains the standard choice for hidden layers because of its computational efficiency and favorable optimization properties. Alternative functions, such as Leaky ReLU and ELU, are mainly employed when improved gradient propagation is required, whereas Sigmoid is generally reserved for binary classification output layers. More recently, Swish has demonstrated improved performance in very deep architectures by maintaining nonzero gradients over the entire input domain [Ramachandran et al., 2017]. Consequently, the selection of an activation function depends on the specific learning task, network depth, and optimization requirements.



(a) ReLU **(b) Leaky ReLU**

(c) Sigmoid **(d) ELU**

Figure 2.11: Selected activation functions used in convolutional neural networks: (a) ReLU; (b) Leaky ReLU with $\alpha = 0.1$; (c) Sigmoid; and (d) ELU with $\alpha = 1$. These nonlinear transformations differ in how they process negative inputs and influence gradient propagation during network optimization [Litjens et al., 2017].

2.5.6 Pooling and Spatial Downsampling

Pooling is a fundamental operation in Convolutional Neural Networks (CNNs) because it reduces the spatial dimensions of feature maps while retaining the most informative activations. This reduction decreases the computational and memory requirements of subsequent layers and progressively enlarges the network's effective receptive field.

Three main pooling strategies are commonly distinguished [Zeiler y Fergus, 2014]. Max-pooling selects the maximum activation within each local window, typically using a 2×2 window with a stride of 2. Average-pooling computes the arithmetic mean of the activations within the pooling region. Stochastic pooling selects an activation randomly according to a probability distribution derived from the activation magnitudes within the local window.

The behavior of these three operations is illustrated in Figure 2.12. Each method is applied to a 4×4 input matrix using a 2×2 pooling window and a stride of 2. Max-pooling preserves the strongest response within each region, average-pooling summarizes the local activation level, and stochastic pooling samples a single activation

Table 2.4. Comparison of common activation functions, including their mathematical definitions, derivatives, output ranges, and typical applications in convolutional neural networks.

Function	Equation	Derivative	Output range	Typical application
ReLU	$\max(0, x)$	$\begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$	$[0, \infty)$	Hidden layers
Leaky ReLU	$\begin{cases} x, & x > 0 \\ \alpha x, & x \leq 0 \end{cases}$	$\begin{cases} 1, & x > 0 \\ \alpha, & x \leq 0 \end{cases}$	$(-\infty, \infty)$	Hidden layers when reducing the risk of inactive neurons is important
ELU	$\begin{cases} x, & x > 0 \\ \alpha(e^x - 1), & x \leq 0 \end{cases}$	$\begin{cases} 1, & x > 0 \\ f(x) + \alpha, & x \leq 0 \end{cases}$	$(-\alpha, \infty)$	Deep architectures that benefit from smooth negative responses
Sigmoid	$\frac{1}{1 + e^{-x}}$	$f(x)\,(1 - f(x))$	$(0, 1)$	Binary classification output layer
Swish	$x\,\sigma(x)$	$\sigma(x) + x\sigma(x)(1 - \sigma(x))$	Approximately $[-0.278, \infty)$	Very deep architectures, including EfficientNet

Abbreviations: ReLU = Rectified Linear Unit; ELU = Exponential Linear Unit; Swish = self-gated activation function. $\sigma(x)$ denotes the sigmoid function.

based on its relative magnitude.

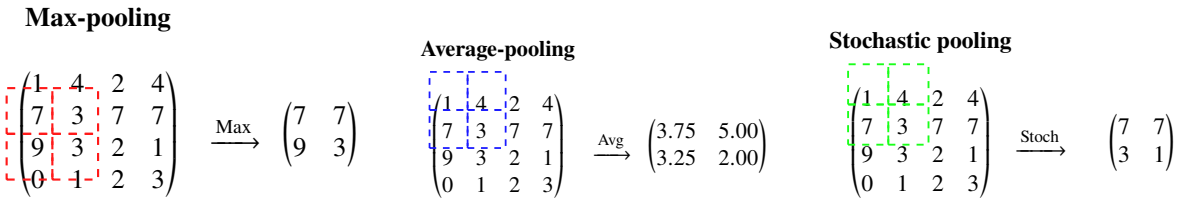


Figure 2.12: Illustration of max-pooling, average-pooling, and stochastic pooling applied to a 4×4 *input matrix using a* 2×2 *pooling window and a stride of 2.*

The principal characteristics of these pooling strategies are summarized in Table 2.5. Max-pooling is appropriate when a strong local response matters more than the precise distribution of activations within the pooling window. Average-pooling provides a smoother summary of local responses and is frequently used in global pooling layers near the output of classification networks. Stochastic pooling introduces randomness by sampling activations according to their normalized magnitudes and may therefore provide a regularizing effect [Zeiler y Fergus, 2014].

In anatomical landmark localization, spatial precision matters because the output must preserve the positions of fine anatomical structures. Max-pooling is still used in encoder-based architectures because it preserves the strongest local responses but reduces feature-map resolution. Pooling is not spatially lossless: it partially discards where activations are located and how they are distributed within each pooling window.

encoder-decoder architectures like U-Net overcome this limitation by linking downsampled semantic representations with high resolution encoder features using skip connections. Skip connections do not recover

Table 2.5. Comparison of common pooling operations, including their mathematical principles, advantages, and limitations.

Pooling type	Operation	Advantages	Limitations
Max-pooling	$f(\mathbf{x}) = \max_i x_i$	Preserves the strongest local activations; computationally efficient; increases tolerance to small spatial variations	May discard weaker but relevant responses; sensitive to isolated high-intensity noise
Average-pooling	$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$	Provides a smooth summary of local activations; reduces local variance; less sensitive to isolated peaks	May attenuate strong discriminative responses and blur fine spatial details
Stochastic pooling	$P(i) = \frac{x_i}{\sum_{j=1}^{n} x_j}$	Introduces regularization through probabilistic sampling; preserves activation diversity	Non-deterministic during sampling; introduces additional variability and requires non-negative activations

all information lost during pooling but introduce fine spatial details back into the decoding path (and therefore improve reconstruction of dense output maps).

Pooling therefore has two complementary functions. First it lowers the computational cost and memory consumption of deeper network layers. Second, it increases the effective receptive field, defined as the region of the original input image that influences a particular activation. In the initial layers, neurons respond to small local regions and primarily encode edges, textures, and simple patterns. After successive convolution and pooling operations, deeper neurons respond to progressively larger image regions and can represent broader anatomical structures and long-range spatial relationships [Zeiler y Fergus, 2014].

This progressive reduction in spatial resolution and the simultaneous expansion of the receptive field constitute a central mechanism of hierarchical feature extraction in CNNs and provide the basis for the contracting path of U-Net.

2.5.7 Hierarchical Feature Learning

The successive application of convolution, nonlinear activation, normalization, and pooling operations produces a hierarchy of feature maps that represent the input image at progressively increasing levels of abstraction. In the earliest layers, the network primarily detects low-level visual primitives such as edges, intensity transitions, and local gradients. As the network deepens, these elementary responses combine into textures, contours, structural patterns, anatomical parts, and higher-level semantic representations [Hubel y Wiesel, 1962; LeCun et al., 1998; Litjens et al., 2017].

Figure 2.13 summarizes this hierarchical progression. The figure illustrates how successive processing stages

transform the original image into increasingly abstract representations, moving from low-level local features to high-level anatomical concepts.

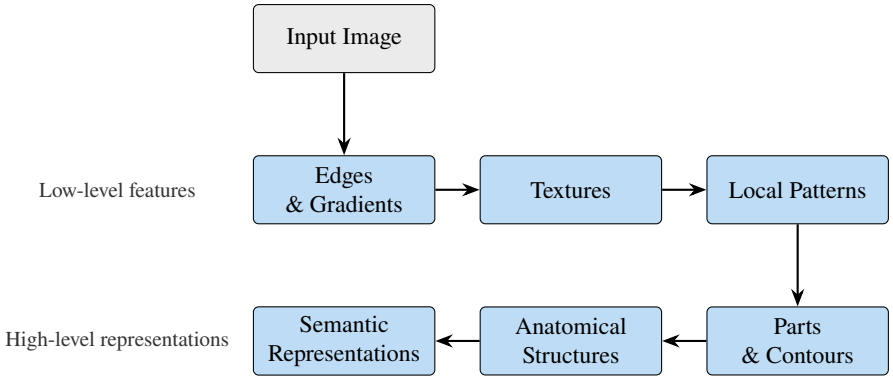


Figure 2.13: *Hierarchical feature learning in convolutional neural networks. Successive processing stages transform the input image from low-level visual primitives, such as edges and gradients, into textures, contours, anatomical structures, and increasingly abstract semantic representations [LeCun et al., 1998; Litjens et al., 2017].*

Computationally, this hierarchy is constructed through repeated convolutional transformations. For an output channel or filter k, the response at spatial position (i, j) is defined according to Eq. (2.5):

$$y_{i,j}^{(k)} = b^{(k)} + \sum_{c} \sum_{u} \sum_{v} w_{u,v,c}^{(k)} \, x_{i+u,j+v,c}, \tag{2.5}$$

Here, $x_{i+u,j+v,c}$ denotes the input activation at spatial position $(i+u, j+v)$ and channel c, $w_{u,v,c}^{(k)}$ represents the learnable kernel coefficient associated with output channel k, $b^{(k)}$ is the corresponding bias term, and $y_{i,j}^{(k)}$ is the resulting output activation. Eq. (2.5) shows how the network combines local spatial information across input channels to generate one activation for the output channel k.

Strictly speaking, Eq. (2.5) corresponds to the cross-correlation operation commonly implemented in deep learning libraries because the kernel is not reversed before multiplication. Nevertheless, the CNN literature conventionally retains the term convolution.

2.5.8 How Convolutional Neural Networks Work

Building on the operations described above, a CNN transforms an input image into progressively richer feature representations through a sequence of structured processing stages. The general workflow is illustrated in Figure 2.14.

The network first processes the input image through one or more convolutional layers. Each layer applies multiple learnable filters across the spatial domain of the input, and each filter produces a feature map that responds to a particular visual pattern. During training, the optimization process adjusts the filter coefficients, enabling the network to learn task-relevant image representations directly from annotated data rather than relying

on manually designed descriptors.

Figure 2.14 summarizes the main processing stages of a CNN for cephalometric image analysis.

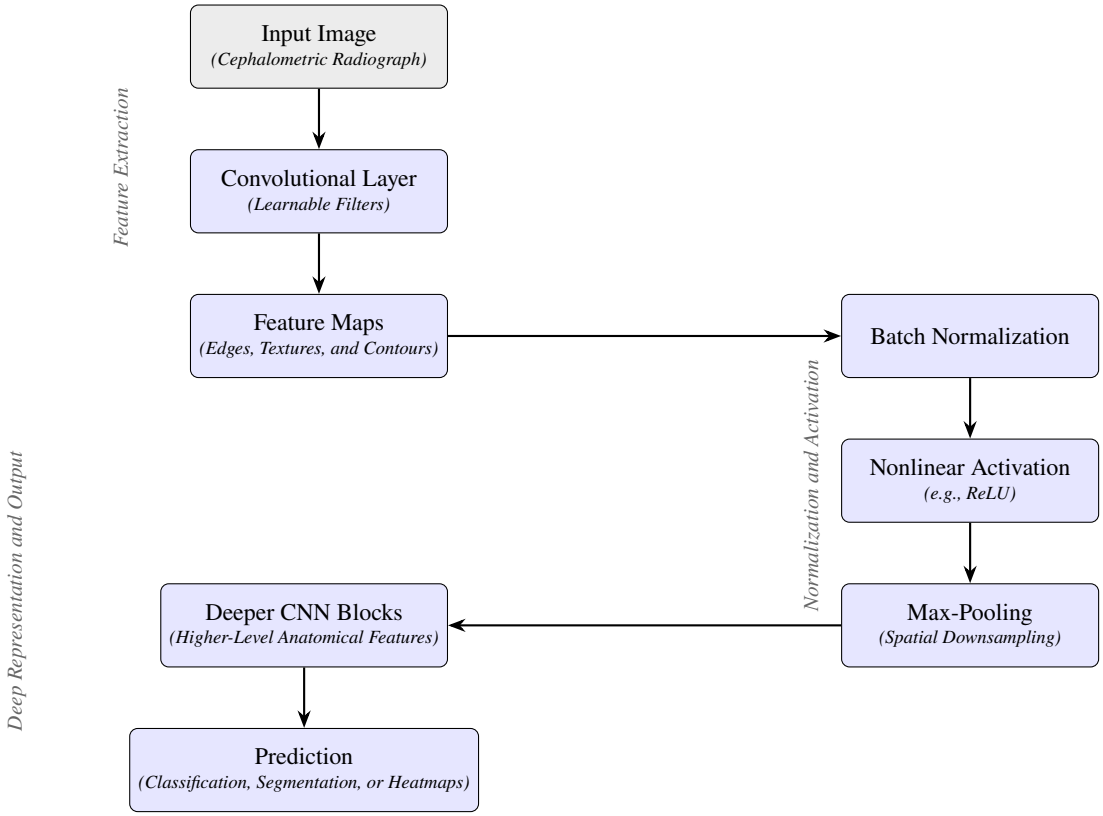


Figure 2.14: *General processing pipeline of a convolutional neural network for cephalometric image analysis. The input radiograph is successively transformed through convolution, normalization, nonlinear activation, and spatial downsampling, producing hierarchical feature representations before the final task-specific prediction is generated [Goodfellow et al., 2016; Litjens et al., 2017].*

Batch normalization standardizes intermediate activations using mini-batch statistics and learnable affine parameters. This operation can improve optimization stability and accelerate convergence by reducing variations in the activation distributions across training iterations. Nonlinear activation functions then introduce the nonlinearity required to represent complex image mappings.

Pooling operations reduce the spatial resolution of the feature maps while retaining informative responses and enlarging the effective receptive field of subsequent layers. The network repeats the sequence of convolution, normalization, activation, and pooling across multiple levels, allowing the learned representations to evolve from simple local image primitives toward broader anatomical structures and task-relevant semantic patterns [Goodfellow et al., 2016; Litjens et al., 2017].

In landmark-localization networks, the final prediction stage may generate one heatmap for each anatomical landmark rather than a conventional classification output. These dense response maps preserve the spatial organization of the localization task and provide the basis for subsequent numerical coordinate extraction.

2.5.9 Advantages and Limitations of CNNs

Convolutional Neural Networks offer several advantages over traditional image-processing methods and handcrafted-feature approaches. As summarized in Table 2.6, these advantages arise primarily from automatic feature learning, parameter sharing, and hierarchical representation learning [Litjens et al., 2017; Shen et al., 2017].

First, CNNs learn discriminative image representations directly from annotated training data, thereby reducing the need for manually designed descriptors. Second, shared convolutional kernels across spatial positions considerably reduce the number of trainable parameters compared with fully connected networks. Third, successive convolutional layers progressively combine local image patterns into higher-level representations, enabling the network to model both fine image details and broader anatomical organization.

Table 2.6 summarizes the main advantages and limitations of conventional CNNs in medical image analysis.

Table 2.6. *Main advantages and limitations of conventional convolutional neural networks for medical image analysis.*

Advantages	Limitations
Automatic feature learning The network learns task-relevant image features directly from annotated training data, reducing dependence on manually engineered descriptors.	**Loss of spatial resolution** Repeated pooling and downsampling operations progressively reduce feature-map resolution, potentially removing fine anatomical details.
Weight sharing The same convolutional kernels are applied across the entire image, reducing trainable parameters and improving computational efficiency.	**Restricted long-range modeling** Local convolutional kernels initially capture information within limited receptive fields, and several layers may be required to model long-range anatomical relationships.
Hierarchical feature representation Successive convolutional layers progressively encode image information from low-level edges and textures to higher-level anatomical structures.	**Challenges in precise localization** Spatial downsampling may reduce localization accuracy, particularly for small or poorly contrasted landmarks requiring high spatial precision.

Despite these advantages, conventional CNN architectures present important limitations for dense anatomical localization. Repeated pooling and downsampling enlarge the receptive field and reduce computational cost, but they also progressively decrease spatial resolution. Consequently, fine anatomical details may be partially lost, compromising the accurate localization of small or poorly defined structures.

These limitations are particularly relevant in cephalometric landmark localization, where errors of only a few millimeters can affect the reliability of subsequent angular and linear measurements. Dense prediction architectures therefore combine contextual feature extraction with mechanisms for recovering high-resolution

spatial information, such as decoder paths, upsampling operations, and skip connections [Litjens et al., 2017].

2.5.10 CNN-Based Landmark Localization

Researchers broadly classify CNN-based landmark-localization methods according to their prediction representation and spatial inference strategy.

The first category comprises direct coordinate-regression methods, in which the network predicts the numerical coordinates of each landmark directly from the input image. As Song et al. [2020] review, early deep learning approaches estimated landmark horizontal and vertical coordinates using dedicated regression networks. Although this formulation is conceptually simple, it compresses the spatial localization problem into a limited set of numerical outputs and provides relatively sparse supervision during training.

The second category includes patch-based methods, which estimate landmark positions from localized image regions around candidate locations [Payer et al., 2016]. By focusing on smaller regions, these methods can capture fine local details. However, their accuracy depends on the initial candidate region and may deteriorate when the extracted patch does not adequately contain the target landmark.

The third category comprises coarse-to-fine strategies. An initial network estimates an approximate landmark position from the complete image, after which one or more refinement stages improve the prediction at finer spatial scales [Payer et al., 2019]. This approach combines global anatomical context with detailed local analysis.

These categories are not mutually exclusive. For example, a coarse-to-fine framework may use patch-based refinement, while either stage may predict coordinates or heatmaps. Localization performance therefore depends not only on the network architecture but also on the selected output representation [Sun et al., 2019].

Consequently, researchers have widely adopted heatmap regression for anatomical landmark localization and human pose estimation. Rather than predicting coordinates directly, the network generates one dense response map for each landmark. The training process commonly supervises these maps with Gaussian target distributions centered on the annotated landmark positions [Newell et al., 2016; Payer et al., 2019].

This formulation preserves the spatial correspondence between the input image and the prediction space while distributing supervision across a neighborhood surrounding each annotated point. It therefore provides richer spatial information than direct coordinate regression and can represent dispersed or multiple responses when the network remains uncertain about the landmark location.

Figure 2.15 illustrates the principle of heatmap-based localization. The strongest response identifies the region most compatible with the landmark position, while the surrounding activation pattern reflects the spatial distribution of the network response. In Figure 2.15, GT denotes the ground-truth landmark location, whereas Pred denotes the location predicted by the network.

Heatmap-based methods generally provide more stable optimization than direct coordinate regression because they preserve the spatial structure of the localization task and provide supervision across the output grid [Sun

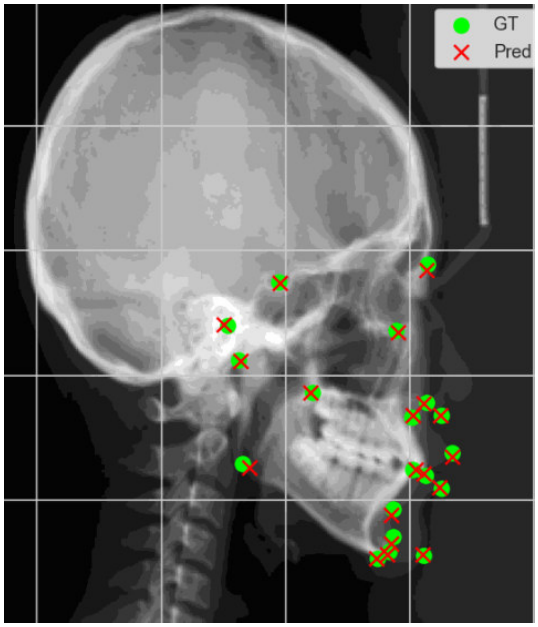


Figure 2.15: *Illustration of heatmap-based landmark localization, in which a spatial response map is generated around the reference anatomical position and its highest-response region indicates the predicted landmark location.*

et al., 2019; Payer et al., 2019]. Their final accuracy, nevertheless, depends on the heatmap resolution, Gaussian width, loss function, and coordinate-decoding method.

2.6 Heatmap-Based Landmark Localization

Heatmap regression has become a widely adopted paradigm for anatomical landmark localization because it often provides greater optimization stability and localization accuracy than direct coordinate regression [Newell et al., 2016].

Instead of predicting landmark coordinates directly, the network generates an independent two-dimensional response map for each landmark. Each pixel represents the network response associated with the possible presence of the landmark at that spatial location. This formulation preserves the spatial correspondence between the input image and the prediction space, provides dense supervision, and retains information about the spatial distribution of the predicted responses [Payer et al., 2019; Juneja et al., 2021].

2.6.1 Coordinate Regression versus Heatmap Regression

In direct coordinate regression, the network estimates the landmark position according to Eq. (2.6):

$$\hat{\mathbf{p}} = (\hat{x}, \hat{y}), \tag{2.6}$$

Here, $\hat{\mathbf{p}}$ denotes the predicted landmark position, and $\hat{x}$ and $\hat{y}$ represent its predicted horizontal and vertical coordinates, respectively. Eq. (2.6) expresses the landmark location as a two-dimensional coordinate pair.

Although straightforward, this formulation compresses high-dimensional spatial information into two numerical values. The prediction space therefore does not explicitly preserve the image's spatial organization, which may complicate optimization [Sun et al., 2019; Payer et al., 2019]. Direct regression also provides limited information about spatial uncertainty and may struggle when several image regions exhibit similar visual characteristics [Newell et al., 2016].

In contrast, heatmap regression predicts a dense spatial response map according to Eq. (2.7):

$$H : \Omega \rightarrow \mathbb{R}, \tag{2.7}$$

Here, H denotes the response map, Ω represents the spatial domain of the output grid, and each value $H(x, y)$ represents the response associated with the proximity of the spatial location (x, y) to the target landmark. Eq. (2.7) defines the heatmap as a spatially distributed representation rather than as a single coordinate pair.

During training, the target heatmap is typically generated using a two-dimensional Gaussian centered on the annotated landmark position. This representation preserves the spatial relationship between the input and prediction spaces while providing dense supervision throughout the output grid [Payer et al., 2019].

Compared with direct coordinate regression, heatmap regression generally yields more stable optimization because the supervision is distributed across neighboring pixels rather than restricted to a single coordinate pair [Sun et al., 2019].

2.6.2 Gaussian Heatmap Representation

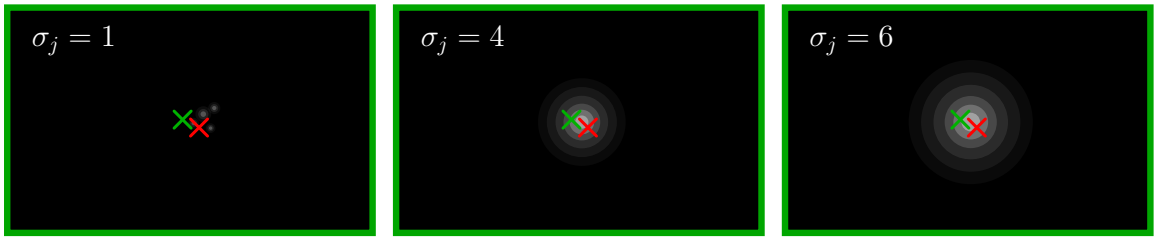
In most heatmap-based methods, each landmark is represented by a two-dimensional Gaussian centered on its annotated position [Yang et al., 2023]. The target heatmap is defined according to Eq. (2.8):

$$H_i(x, y) = \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma_i^2}\right), \tag{2.8}$$

Here, (x_i, y_i) denotes the ground-truth coordinates of landmark i, (x, y) identifies a spatial position in the heatmap, and σ_i controls the spatial spread of the response. Eq. (2.8) assigns the highest response to locations close to the annotated landmark position and progressively decreases the response as the distance from that position increases.

In the continuous formulation, the Gaussian reaches its maximum at the landmark center, such that $H_i(x_i, y_i) = 1$. This property ensures a consistent theoretical peak response regardless of the selected value of σ_i.

The influence of the Gaussian width is illustrated in Figure 2.16. The figure compares target heatmaps generated with different values of σ_i and shows how the parameter controls the supervisory signal's spatial extent.



(a): $\sigma = 1$ *(b):* $\sigma = 4$ *(c):* $\sigma = 6$

Figure 2.16: *Illustration of Gaussian target heatmaps generated with different values of the standard deviation σ. (a) A small value ($\sigma = 1$) produces a highly concentrated response with a narrow supervisory region. (b) An intermediate value ($\sigma = 4$) balances localization precision and optimization stability. (c) A larger value ($\sigma = 6$) yields a broader response, distributing the learning signal over a wider neighborhood, but may reduce localization precision through excessive smoothing.*

The selection of σ therefore defines a trade-off between localization precision and optimization stability. Small values generate sharper targets but provide weaker gradients away from the landmark center, whereas larger values extend the supervisory signal over a broader region. Gaussian width is consequently an important design parameter in heatmap-based landmark-localization systems [Yang et al., 2023].

2.7 U-Net Architectures for Segmentation and Landmark Localization

The transition from direct coordinate regression to heatmap-based prediction raises the question of which architecture is best suited to dense pixel-wise prediction. Among the principal encoder-decoder architectures developed for this purpose, Fully Convolutional Networks (FCNs), SegNet, and U-Net have played central roles.

Fully Convolutional Networks, introduced by Long et al. [2015], adapted image-classification networks for dense prediction by replacing fully connected layers with convolutional layers. This modification enabled the networks to produce pixel-wise outputs for images of arbitrary size.

SegNet [Badrinarayanan et al., 2017] subsequently introduced a symmetric encoder-decoder architecture in which the decoder reconstructs feature maps using the pooling indices retained from the encoder. This strategy reduces trainable parameters while preserving some of the spatial information lost during downsampling. Nevertheless, both FCN and SegNet present limitations when the network must preserve fine spatial detail and high-level semantic information simultaneously.

U-Net [Ronneberger et al., 2015] addressed this challenge through systematic skip connections between corresponding encoder and decoder levels. These connections combine high-resolution spatial information with

deep semantic representations, making U-Net particularly suitable for biomedical images characterized by low contrast, anatomical variability, and complex spatial organization.

U-Net's success in biomedical image segmentation led researchers to adopt the architecture for anatomical landmark localization. In this setting, U-Net predicts one dense response map for each landmark rather than estimating coordinates directly, thereby formulating localization as a dense prediction task.

2.7.1 Original U-Net Architecture

Originally proposed by Ronneberger et al. [2015] for biomedical image segmentation, U-Net is a fully convolutional encoder-decoder architecture designed for dense pixel-wise prediction. Its characteristic U-shaped topology comprises a contracting path, or encoder, that extracts hierarchical feature representations, and an expanding path, or decoder, that restores spatial resolution. Skip connections link the two paths, as illustrated in Figure 2.17.

Figure 2.17 presents the original U-Net architecture and highlights the interaction between the contracting path, the expanding path, and the corresponding skip connections.

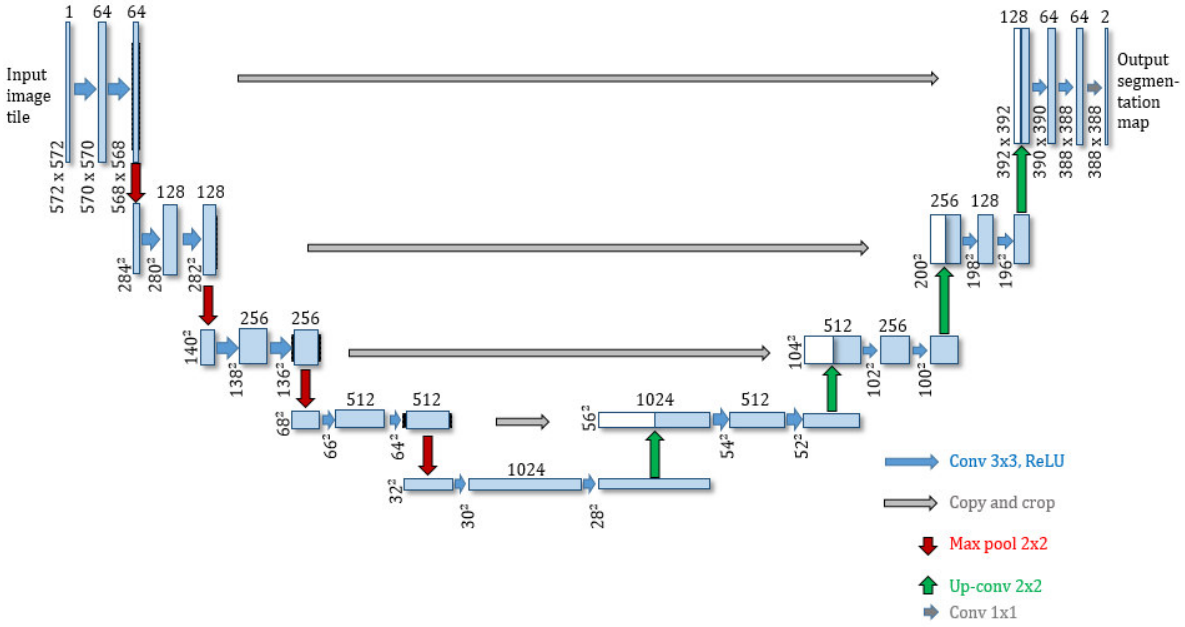


Figure 2.17: Original U-Net architecture proposed by Ronneberger et al. [2015]. The encoder progressively reduces the spatial resolution while increasing the number of feature channels. The decoder restores the original resolution through transposed convolutions and skip connections, producing a dense output map.

Unlike classification networks that terminate with fully connected layers, U-Net remains fully convolutional throughout its main processing path. A final 1×1 convolution projects the decoder features onto the required output channels while preserving the spatial correspondence between the input image and the output representation [Long et al., 2015; Ronneberger et al., 2015].

In the original formulation, these output channels represent semantic segmentation classes. For landmark

localization, researchers adapt the architecture by assigning one output channel to each anatomical landmark. Each channel predicts a Gaussian heatmap rather than a segmentation mask, thereby converting landmark localization into a dense prediction problem [Newell et al., 2016; Wei et al., 2016; Khan et al., 2025].

Skip connections are the core of U-Net. High resolution features are passed directly from shallow encoder layers to decoder levels (where they are fused with deeper semantic representations) [Ronneberger et al., 2015]. Fine anatomical information is preserved but broader context information is retained.

The encoder downsamples spatial resolution (convolution and pooling) while increasing the number of feature channels. This enlarges receptive field and produces increasingly abstract representations. The decoder then restores spatial resolution by combining upsampled features with corresponding encoder outputs — information lost during downsampling is recovered.

This balance between global anatomy and local spatial precision is crucial for cephalometric landmark localization. Local appearance alone may not be sufficient because different craniofacial structures can have similar radiographic patterns or exclusively global representations may omit details necessary to distinguish neighboring landmarks.

U-Net thus provides an effective backbone for heatmap-based cephalometric landmark localization that jointly leverages contextual and high resolution information. The following subsections discuss its principal components: the contracting path, expanding path and skip connections.

2.7.2 Contracting and Expanding Paths

U-Net belongs to the family of fully convolutional encoder-decoder architectures designed for dense image prediction. These architectures generate an output for every pixel rather than for the entire image [Long et al., 2015; Ronneberger et al., 2015]. Unlike conventional classification CNNs, encoder-decoder architectures learn hierarchical representations while reconstructing high-resolution output maps, making them well suited to biomedical segmentation and anatomical landmark localization [Gu et al., 2018].

2.7.2.1 Contracting Path (Encoder)

The contracting path extracts increasingly abstract feature representations while progressively reducing the spatial resolution. In the original U-Net, each encoder block contains two successive 3×3 convolutional layers, each followed by a Rectified Linear Unit (ReLU), and a 2×2 max-pooling operation with a stride of two. The original architecture uses valid padding, so the convolutional and pooling operations reduce the spatial dimensions [Ronneberger et al., 2015].

As the feature maps are downsampled, the receptive field expands, enabling the network to integrate information from increasingly large anatomical regions [Litjens et al., 2017]. At the same time, the number of channels doubles after each pooling stage, following the progression $64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$. This

design increases the representational capacity as the spatial resolution decreases [Ronneberger et al., 2015].

Consequently, the network progressively transforms low-level features, such as edges, gradients, and textures, into higher-level semantic representations of anatomical structures [Zeiler y Fergus, 2014; Goodfellow et al., 2016].

2.7.2.2 Bottleneck

The bottleneck is the deepest part of the network and contains the most abstract representation of the input at the lowest spatial resolution [Ronneberger et al., 2015]. Its size depends on the size of the input and valid padding strategy used throughout the encoder. It combines broad contextual information and gives the global anatomical description from which decoder starts reconstruction.

2.7.2.3 Expanding Path (Decoder)

The expanding path restores the spatial resolution gradually. In each decoder stage a 2×2 transposed convolution (also known as an up-convolution) is applied (doubles spatial dimensions while reducing number of channels) [Ronneberger et al., 2015].

Then the network upsamples the feature maps with two successive 3×3 convolutional layers followed by ReLU activations and recovers increasingly detailed spatial information [Gu et al., 2018]. In original implementation dropout is applied near bottleneck to avoid overfitting when training data are not abundant [Ronneberger et al., 2015].

A final 1×1 convolution maps the decoder features onto the required output channels [Ronneberger et al., 2015]. In landmark-localization applications each channel represents a two-dimensional Gaussian heatmap associated with one anatomical landmark. This design allows the network to predict dense spatial responses rather than numerical coordinates directly [Newell et al., 2016; Khan et al., 2025].

Through this encoder-decoder process, U-Net transforms the input radiograph into one heatmap for each anatomical landmark. Accurate heatmap prediction, however, requires the network to preserve fine spatial information, which motivates the use of skip connections.

2.7.3 Skip Connections and Preservation of Spatial Information

A defining feature of U-Net is its systematic use of long-range encoder-decoder skip connections [Ronneberger et al., 2015]. These connections transfer high-resolution encoder feature maps directly to the corresponding decoder stages, thereby compensating for information lost during downsampling.

Let $f_e^{(l)}$ denote the encoder feature map at level l, and let $f_d^{(l+1)}$ denote the decoder feature map from the next deeper level. The skip connection is expressed according to Eq. (2.9):

$$f_d^{(l)} = \mathrm{Up}\left(f_d^{(l+1)}\right) \oplus f_e^{(l)}, \tag{2.9}$$

Here, $\mathrm{Up}(\cdot)$ denotes the upsampling operator and $\oplus$ represents channel-wise concatenation. Eq. (2.9) shows how the decoder combines the upsampled representation from the deeper level with the corresponding high-resolution feature map from the encoder. The concatenated tensor is then processed by convolutional layers that integrate information from both pathways.

For cephalometric landmark localization, deep encoder layers provide semantic information about the anatomical structure associated with each landmark, whereas shallow layers preserve the high-resolution spatial details required for precise localization [Ronneberger et al., 2015; Litjens et al., 2017].

Without skip connections, the decoder would reconstruct fine anatomical details solely from low-resolution bottleneck features, which could produce blurred heatmaps and reduce localization accuracy. By combining detailed spatial information from shallow layers with contextual information from deeper layers, skip connections enable U-Net to preserve both anatomical context and localization precision [Ronneberger et al., 2015].

2.8 Recent Developments in Deep Learning for Cephalometric Landmark Localization

Recent advances in deep learning have substantially improved automatic cephalometric landmark localization. Early convolutional neural networks (CNNs) have evolved into multiscale, attention-based, and hybrid CNN–Transformer architectures that jointly exploit local image features and long-range anatomical relationships. These developments have improved localization accuracy but also introduced additional challenges related to computational complexity, generalization, and reproducibility. Representative methods are summarized in Table 2.7, and the studies that are not included in the table are discussed in the following subsections.

2.8.1 Evolution of Deep Learning Architectures

The evolution of cephalometric landmark localization reflects a gradual transition from conventional CNNs to architectures that integrate multiscale information and anatomical reasoning.

Song et al. [2020] proposed a two-stage framework based on ResNet50 and evaluated it on the ISBI 2015 benchmark. The first stage performs global localization on the entire radiograph, while the second stage refines the predictions within automatically extracted regions of interest (ROIs). Although this coarse-to-fine strategy improves localization accuracy, its performance depends strongly on ROI extraction and does not explicitly model anatomical relationships.

Yang et al. [2023] proposed Ceph-Net, an attention-based stacked regression network evaluated on 400 lateral cephalograms. The architecture uses Dual Spatial Attention Modules and Multi-path Convolution Modules to

iteratively refine landmark heatmaps while exploiting multiscale contextual information. Although the attention mechanisms improve robustness, the stacked architecture substantially increases computational cost.

Takahashi et al. [2023] combined HRNetV2-based heatmap regression with multilayer perceptron (MLP) coordinate refinement. The MLP explicitly models spatial relationships among landmarks after heatmap prediction, thereby addressing a limitation of conventional heatmap regression. However, the evaluation used a private dataset of facial profile photographs, which restricts direct comparison with standard cephalometric benchmarks.

2.8.2 Emerging Methodological Trends

Beyond architectural innovations, several methodological trends characterize the current state of the art.

Heatmap regression as the dominant output representation.

Gaussian heatmap regression has become a standard representation for landmark localization. By predicting dense response maps instead of coordinates, this approach preserves spatial information, provides richer supervision, and supports heatmap-based optimization [Newell et al., 2016; Payer et al., 2019; Sun et al., 2019]. The dispersion of a response map may describe spatial ambiguity, but researchers should not interpret it as calibrated predictive uncertainty without a dedicated uncertainty-estimation procedure.

Multiscale feature extraction.

Modern architectures increasingly combine global anatomical context with fine local details through encoder-decoder networks, high-resolution representations, and feature pyramids. U-Net [Ronneberger et al., 2015] and HRNet [Wu et al., 2023] illustrate the importance of maintaining high-resolution features for accurate landmark localization.

Explicit modeling of anatomical relationships.

Recent methods incorporate attention mechanisms, graph reasoning, Transformer blocks, or coordinate-refinement modules to model landmark dependencies. These approaches address the independence assumption underlying conventional heatmap regression [Jiang et al., 2022; Yang et al., 2023; Takahashi et al., 2023].

Hybrid CNN–Transformer architectures.

Combining CNNs for local feature extraction with Transformer self-attention for global contextual modeling has become a promising research direction. Architectures such as CephalFormer [Jiang et al., 2022] and Domain-Adaptive Transformer [Zhu et al., 2022] improve the representation of long-range anatomical relationships, although they also increase computational complexity.

Differentiable coordinate extraction.

Researchers have increasingly investigated differentiable methods for converting heatmaps into landmark coordinates. Soft-Argmax establishes a continuous mapping between dense response maps and numerical

coordinates and can support coordinate-level optimization when included in the training objective [Nibali et al., 2018; Payer et al., 2019]. It also provides continuous sub-pixel estimates without discrete Argmax post-processing. Coordinate decoding therefore remains an important but comparatively underexplored component of heatmap-based localization frameworks.

Cross-dataset generalization.

An increasing number of studies evaluate their methods across multiple datasets, emphasizing the importance of domain generalization. Domain adaptation and multi-domain training strategies can improve robustness across imaging conditions, although their effectiveness often depends on dataset characteristics and hyperparameter selection [Zhu et al., 2022; Sumon et al., 2025].

2.8.3 Performance Comparison

Table 2.7 summarizes representative deep learning approaches for automatic cephalometric landmark localization, their evaluation datasets, their reported localization performance, and their principal contributions. Performance values are reported as published by the original studies, which use different evaluation datasets and protocols; the dataset-specific values are used for the comparison with the proposed method in Section 4.5.4.

The studies listed in Table 2.7 illustrate the progression from conventional CNN-based approaches toward architectures that combine coarse-to-fine refinement, high-resolution feature extraction, attention mechanisms, and explicit anatomical reasoning. The results reported also show that localization accuracy as well as generalization capacity of each method should be taken into account.

Three main points can be drawn from Table 2.7. First, reported performance varies substantially across benchmarks. Methods evaluated on ISBI 2015 (19 landmarks) generally report an MRE between 1.08 and 1.64 mm and an SDR@2 mm between 70.58 and 88.03 %, whereas methods evaluated on more recent or more heterogeneous benchmarks reach different operating points: the dual-encoder model of Dai et al. [2024] achieves an MRE of 0.94 mm and an SDR@2 mm of 90.68 % on ISBI 2023, evaluating 19 of its 29 landmarks, while the method of Wu et al. [2023] achieves an MRE of 1.52 mm and an SDR@2 mm of 75.72 % on CL-Detection2023 with 38 landmarks. The effect of the benchmark is particularly visible when a single method is evaluated on two datasets: the cascaded CNN of Zeng et al. [2021] reports a pooled MRE of 1.46 mm and an SDR@2 mm of 76.82 % on ISBI 2015, but an MRE of 2.02 ± 1.89 mm and an SDR@2 mm of only 64.81 % on the PKU dataset [Zeng et al., 2021]. Because the studies use different datasets, test partitions, and implementation settings, these values should be interpreted as reported results rather than as a strictly controlled ranking.

Second, several of the higher-performing methods combine local feature extraction with mechanisms for modeling global anatomical context, such as attention modules, Transformer blocks, or high-resolution representations. This pattern suggests that local appearance and global anatomical relationships provide complementary information for accurate landmark localization.

Third, reported performance gains generally come with increased architectural complexity. Hybrid

Table 2.7. Representative deep learning approaches for automatic cephalometric landmark localization. Performance values are reported according to the evaluation protocols of the original studies and are therefore not necessarily directly comparable across datasets or test partitions. For studies reporting Test1 and Test2 separately, values are given as Test1 / Test2.

Study	Architecture	Dataset	MRE [mm]	SDR@2 [%]	Main contribution
Zeng et al. [2021]	Cascaded three-stage CNN	ISBI 2015 (Test1 / Test2)	1.34 / 1.64	81.37 / 70.58	Coarse-to-fine landmark localization using global face alignment, simultaneous landmark prediction, and landmark-specific high-resolution refinement.
Ronneberger et al. [2015]	U-Net	Biomedical images	–	–	Foundational encoder-decoder architecture with skip connections for dense spatial prediction.
Rashmi et al. [2022]	U-Net + heatmap regression	ISBI 2015	1.17	85.36	Heatmap-based landmark regression using a U-Net-based convolutional framework.
Jiang et al. [2022]	CephalFormer (CNN–Transformer)	Cephalometric datasets (2D/3D)	–	–	Transformer-based modeling of global structural relationships between visual concepts and anatomical landmarks.
Zhu et al. [2022]	Domain-Adaptive Transformer (DATR)	ISBI 2015 (multi-domain training)	1.47	78.07	Domain-adaptive Transformer modeling long-range dependencies for landmark detection across multiple anatomical domains.
Ao y Wu [2023]	FARNet	ISBI 2015 (Test1 / Test2)	1.12 / 1.42	88.03 / 77.00	Multi-scale feature aggregation with high-resolution heatmap refinement and Exponential Weighted Center loss.
Dai et al. [2024]	Dual-encoder regression (Transformer encoder)	ISBI 2023 (19 of 29 landmarks)	0.94	90.68	End-to-end differentiable coarse-to-fine localization combining a feature extractor, a reference encoder, and a fine-tuning encoder.
Wu et al. [2023]	HRNet + super-resolution head	CL-Detection2023 (38 landmarks)	1.52	75.72	High-resolution pose-estimation framework with a lightweight super-resolution head reducing quantization bias.
Sumon et al. [2025]	Self-CephaloNet (HRNetV2 + Self-ONN)	ISBI 2015 (Test1 / Test2)	1.08 / 1.44	87.82 / 76.68	Two-stage high-resolution landmark localization incorporating self-operational neural networks and heatmap-based prediction.

CNN-Transformer models, attention mechanisms, and multistage refinement can improve localization accuracy, but they also increase computational cost and implementation complexity. This trade-off remains important when considering clinical deployment.

2.8.4 Remaining Challenges

Despite substantial progress, several challenges remain throughout the cephalometric landmark-localization pipeline.

Heatmap representation.

Although Gaussian heatmap regression has become the standard output representation, researchers have not yet fully established how key design choices, including Gaussian width, target normalization, and output resolution, affect localization performance.

Architectural complexity and clinical deployment.

State-of-the-art methods increasingly rely on hybrid architectures that combine CNNs, attention mechanisms, Transformers, and multiple refinement stages. While these designs can improve accuracy, they also increase computational cost, inference time, and implementation complexity, which may limit their suitability for routine clinical practice.

Validation protocol rigor.

Validation strategies remain heterogeneous across studies. Many methods rely on a single train–test split or limited cross-validation, making the reported performance dependent on a particular data partition and complicating comparisons across studies. K-fold cross-validation and nested validation can reduce this dependence, particularly when the annotated dataset is small.

Framework integration and methodological consistency.

Most studies emphasize architectural innovations while paying comparatively less attention to preprocessing, coordinate extraction, evaluation methodology, and validation. Consequently, relatively few studies propose an integrated framework in which the complete processing pipeline is designed and evaluated consistently.

2.8.5 Motivation for the Proposed Methodology

The literature reviewed in this chapter shows that recent research has focused primarily on architectural innovation, whereas the methodological consistency of the complete localization process has received comparatively less attention. In particular, researchers have rarely examined the interactions among heatmap representation, coordinate extraction, validation strategy, and quantitative evaluation within a unified framework.

Rather than proposing a new network architecture, this thesis re-engineers the complete localization framework by coordinating well-established components within a reproducible workflow. The methodology combines Gaussian heatmap regression, Soft-Argmax coordinate decoding, a lightweight U-Net architecture, controlled preprocessing, anisotropic coordinate reprojection, training-duration selection, and clinically meaningful evaluation in millimeters.

Despite the advances summarized in Table 2.7, several challenges remain and directly motivate the proposed methodology.

The first challenge concerns landmark heterogeneity. Landmarks in low-contrast regions or affected by anatomical superposition remain harder to localize than landmarks on well-defined bony structures [Payer et al., 2019; Izzetti et al., 2026]. Representing each landmark as a Gaussian heatmap provides dense spatial supervision around its annotated position.

The second challenge concerns the clinical interpretation of localization accuracy. Although most studies report errors in millimeters, relatively few describe the complete transformation pipeline, including image resizing, coordinate reprojection, and pixel-to-millimeter conversion [Juneja et al., 2021]. This thesis explicitly documents every transformation step to ensure that all reported metrics remain clinically interpretable.

The third challenge concerns experimental validation. Many published methods rely on a single data split or limited validation, making their performance estimates dependent on a particular data partition. This thesis therefore adopts a three-level validation hierarchy with the Holdout 70/30 protocol as the primary evaluation. The primary protocol reserves 30 radiographs as a fully isolated external test set while model development proceeds on the remaining 72 radiographs, with an internal validation subset determining the training duration and a newly initialized U-Net refitted from scratch on all 72 development radiographs before the external test radiographs are evaluated once.

Two complementary analyzes strengthen this primary evaluation. The K-Fold 70/30 ensemble experiment trains ten fold-specific instances of the same U-Net on the development data, combines their predicted heatmaps before Soft-Argmax decoding, and evaluates the resulting ensemble on the same 30-image external test cohort, yielding $30 \times 19 = 570$ external landmark predictions. The Nested LOOCV experiment excludes each of the 102 radiographs once from all model-development decisions, selects the training duration by inner validation without using the excluded radiograph, refits a newly initialized U-Net on the remaining 101 radiographs, and evaluates the excluded radiograph exactly once, yielding $102 \times 19 = 1,938$ out-of-sample landmark predictions. Paired statistical analyzes on the 30 common external radiographs then compare the three protocols using matched observations.

Overall, this methodology addresses these challenges by coherently integrating established techniques rather than introducing architectural novelty. A lightweight U-Net combines local spatial precision with contextual information [Ronneberger et al., 2015]; Gaussian heatmap regression provides dense supervision; and Soft-Argmax yields continuous sub-pixel coordinates [Nibali et al., 2018]. The framework is evaluated primarily with the Holdout 70/30 protocol and supported by the complementary K-Fold 70/30 ensemble and Nested LOOCV experiments, using metrics expressed in millimeters.

2.9 Differentiable Coordinate Extraction

Although deep learning architectures have improved heatmap-prediction accuracy, clinicians cannot use heatmaps directly for clinical cephalometric analysis [Lachinov et al., 2020]. Orthodontic measurements require numerical landmark coordinates from which distances, angles, and other cephalometric variables can be computed. Each predicted heatmap must therefore be decoded into a pair of continuous coordinates.

When coordinate-based losses are used, differentiable decoding allows gradients to propagate through the coordinate-estimation stage during optimization.

2.9.1 Existing Coordinate Extraction Strategies

Researchers use several strategies to convert predicted heatmaps into landmark coordinates.

- **Argmax.** The simplest approach selects the pixel with the highest heatmap response. Although computationally efficient, this method restricts predictions to integer pixel positions and therefore introduces quantization error. Moreover, argmax is non-differentiable and cannot propagate gradients through the coordinate-decoding stage during end-to-end optimization [Payer et al., 2019; Sun et al., 2019].
- **Soft-Argmax (integral regression).** Soft-Argmax normalizes the heatmap into a spatial probability distribution and computes the landmark coordinate as the expected position over all pixels [Nibali et al., 2018]. This operation is differentiable and produces continuous sub-pixel estimates, making it suitable for high-precision medical image analysis.
- **Coordinate refinement.** Refinement methods first estimate an initial landmark position and then improve it using an additional regression network or an iterative optimization stage [Payer et al., 2019]. Although these approaches can potentially improve accuracy, they generally increase architectural complexity, computational cost, and training time.

2.9.2 Mathematical Formulation of Soft-Argmax

Soft-Argmax provides a practical compromise between mathematical simplicity, differentiability, and localization precision.

Given a predicted heatmap $H(x, y)$, the method first converts the heatmap into a normalized spatial probability distribution using a temperature-scaled softmax. It then computes the landmark coordinates as the expected spatial positions according to Eq. (2.10):

$$p(x,y) = \frac{\exp(\beta H(x,y))}{\sum_{u,v} \exp(\beta H(u,v))}, \qquad \hat{x} = \sum_{x,y} x\, p(x,y), \qquad \hat{y} = \sum_{x,y} y\, p(x,y), \tag{2.10}$$

Here, $p(x, y)$ denotes the normalized spatial probability at position (x, y), β controls the sharpness of the distribution, and $(\hat{x}, \hat{y})$ denotes the predicted landmark position. The denominator sums the exponentiated responses over all spatial positions in the heatmap. The coordinate estimates correspond to the expected horizontal and vertical positions under this distribution.

Large values of β concentrate the probability mass around the heatmap maximum and increasingly approximate hard argmax. Smaller values produce smoother distributions that may be less sensitive to noise but can reduce localization precision. The parameter β therefore balances a sharply localized response with a spatially distributed one.

Because this computation is differentiable, Soft-Argmax can support coordinate-level optimization while providing continuous sub-pixel estimates [Nibali et al., 2018]. In this thesis, however, the U-Net is optimized

using heatmap MSE, and Soft-Argmax decodes the predicted heatmaps for coordinate-based evaluation. The methodology chapter details the implementation.

2.10 Image Preprocessing and Loss Function

Landmark-localization performance depends strongly on a consistent preprocessing pipeline [Ronneberger et al., 2015]. Resizing radiographs to a fixed resolution, such as 256×256 pixels, standardizes network inputs, facilitates mini-batch training, and reduces computational cost. Because resizing changes the image coordinate system, the same geometric transformation is applied to the landmark annotations. If $(W_{\text{orig}}, H_{\text{orig}})$ and $(W_{\text{new}}, H_{\text{new}})$ denote the original and resized image dimensions, respectively, the transformed coordinates are given according to Eq. (2.11):

$$x_{\text{new}} = x_{\text{orig}} \frac{W_{\text{new}}}{W_{\text{orig}}}, \qquad y_{\text{new}} = y_{\text{orig}} \frac{H_{\text{new}}}{H_{\text{orig}}}, \tag{2.11}$$

Here, $(x_{\text{orig}}, y_{\text{orig}})$ denotes the landmark position in the original image and $(x_{\text{new}}, y_{\text{new}})$ denotes the corresponding position in the resized image. Eq. (2.11) preserves the correspondence between the resized image and its landmark annotations [Payer et al., 2019].

For clinical interpretation, localization errors should be expressed in millimeters rather than pixels. Because resizing may produce different horizontal and vertical scale factors, the conversion must be performed separately along both axes according to Eq. (2.12):

$$\Delta x_{\text{mm}} = \Delta x_{\text{new}} \frac{W_{\text{orig}}}{W_{\text{new}}} r_x, \qquad \Delta y_{\text{mm}} = \Delta y_{\text{new}} \frac{H_{\text{orig}}}{H_{\text{new}}} r_y, \qquad d_{\text{mm}} = \sqrt{\Delta x_{\text{mm}}^2 + \Delta y_{\text{mm}}^2}, \tag{2.12}$$

Here, Δx_{new} and Δy_{new} denote the coordinate errors in the resized image, r_x and r_y denote the original physical pixel spacings along the horizontal and vertical axes, respectively, and d_{mm} denotes the resulting Euclidean localization error in millimeters [Izzetti et al., 2026]. Eq. (2.12) preserves the original image geometry and supports the use of clinically meaningful distance thresholds.

Mean Squared Error (MSE) is commonly used to optimize heatmap-regression models [Payer et al., 2019; Ronneberger et al., 2015]. Given a predicted heatmap H_n^{pred} and its corresponding Gaussian target H_n^{gt}, the loss is defined according to Eq. (2.13):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(H_n^{\text{pred}}(i,j) - H_n^{\text{gt}}(i,j) \right)^2, \tag{2.13}$$

Here, N denotes the number of landmarks, and H and W denote the height and width of each heatmap, respectively. Eq. (2.13) computes the average squared difference between the predicted and target heatmaps across

all landmarks and spatial positions. Minimizing this objective encourages the predicted responses to approximate the Gaussian targets while providing smooth gradients for network optimization.

2.11 Evaluation Metrics and Validation Protocol

Researchers commonly evaluate cephalometric landmark-localization performance using Euclidean distance, Mean Radial Error (MRE), and Successful Detection Rate (SDR) [Juneja et al., 2021; Izzetti et al., 2026]. For clinical evaluation, the predicted and reference coordinates should be expressed in millimeters before calculating these metrics.

The Mean Radial Error is defined according to Eq. (2.14):

$$\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}, \tag{2.14}$$

Here, $(\hat{x}_i, \hat{y}_i)$ and (x_i, y_i) denote the predicted and reference coordinates of landmark i, respectively, and N denotes the total number of evaluated landmarks. Eq. (2.14) computes the average Euclidean distance between the predicted and reference positions. When the coordinates are expressed in millimeters, the resulting MRE is also reported in millimeters.

The Successful Detection Rate measures the percentage of landmarks localized within a predefined threshold τ according to Eq. (2.15):

$$\mathrm{SDR}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(d_i \le \tau) \times 100, \tag{2.15}$$

Here, d_i denotes the Euclidean localization error for landmark i, τ denotes the selected distance threshold, and $\mathbf{1}(\cdot)$ is the indicator function, which equals one when the condition is satisfied and zero otherwise. Eq. (2.15) therefore reports the percentage of landmarks whose localization error does not exceed τ.

Researchers commonly set τ to 2.0, 2.5, 3.0, or 4.0 mm. The 2.0 mm threshold is generally considered the most demanding and is widely used as the primary benchmark for comparing cephalometric localization methods [Sumon et al., 2025].

The validation protocol is equally important because it determines the reliability of the reported performance [Juneja et al., 2021]. This consideration is particularly relevant in medical imaging, where annotated datasets are often limited. A hold-out protocol uses one fixed training, validation, and test partition and therefore provides a simple reference whose performance estimate remains conditional on that particular split.

K-fold cross-validation divides the development data into K folds and successively uses each fold for validation while using the remaining folds for training. In the complementary K-Fold 70/30 experiment adopted in this

thesis, ten instances of the same U-Net architecture are trained using different development partitions. The instances differ in their learned parameters, not in their architecture. Their predicted heatmaps are averaged during ensemble inference on a fixed, fully isolated 30-image external test cohort, the same cohort used by the primary Holdout 70/30 protocol, yielding $30 \times 19 = 570$ external landmark predictions.

The Nested LOOCV experiment provides the additional validation and robustness analysis. The outer loop comprises 102 iterations, and each iteration reserves one radiograph as an external test sample. The remaining 101 radiographs constitute the development set for that outer fold. An inner validation stage selects only the training duration, or best epoch, based on the inner validation loss. The excluded radiograph does not influence training, inner validation, training-duration selection, early stopping, or scheduler operation.

After the inner validation stage selects the training duration, the procedure initializes a new U-Net and refits it on all 101 radiographs in the development set for exactly the selected number of epoadd. The refitted model then evaluates the excluded radiograph once, producing one out-of-sample prediction for each outer fold and $102 \times 19 = 1,938$ landmark-level predictions overall.

2.12 Chapter Conclusion

This chapter reviewed the main scientific developments in automatic cephalometric landmark localization, from conventional image-analysis techniques to modern deep learning methods based on dense heatmap prediction. The literature identifies Gaussian heatmap regression, encoder-decoder architectures, and continuous coordinate extraction as central components of accurate anatomical landmark-localization systems.

The review also shows that reliable performance depends on more than network architecture. Researchers must integrate preprocessing, heatmap representation, coordinate decoding, validation strategy, and evaluation in clinically meaningful units consistently to ensure reproducible and interpretable results.

These findings provide the scientific basis for the methodology presented in the following chapter. The proposed methodology integrates Gaussian heatmap regression, a lightweight U-Net, Soft-Argmax coordinate decoding, anisotropic evaluation in physical units, and the Holdout 70/30 protocol as the primary evaluation. The complementary K-Fold 70/30 ensemble experiment and the Nested LOOCV robustness analysis provide additional evidence of performance consistency under alternative model-development and evaluation designs.

Chapter 3

General Methodology

3.1 Introduction

This chapter presents the methodological framework developed for automatic cephalometric landmark localization in lateral radiographs. The contribution of this work is not the proposal of a new U-Net architecture. Rather, it lies in the systematic integration of established components into a coherent, reproducible, and clinically interpretable framework.

The framework combines deterministic image preprocessing, Gaussian heatmap regression, continuous coordinate decoding through Soft-Argmax, training-duration selection based on internal validation, refitting on the complete development set, and evaluation in physical units. The primary evaluation protocol is a fixed Holdout 70/30 design, in which 72 radiographs are used for model development and 30 radiographs are reserved as a final test set. The same 30-image test cohort is also used for the complementary K-Fold 70/30 evaluation. Nested Leave-One-Out Cross-Validation (LOOCV) provides an additional complementary assessment of robustness across the complete dataset.

The method formulates cephalometric landmark localization as a dense heatmap-regression problem. The network predicts one spatial response map for each anatomical landmark, and a differentiable coordinate-decoding operation converts the response maps into continuous landmark coordinates. This representation preserves the spatial nature of the task while avoiding the integer-grid restriction of classical Argmax.

The framework comprises three principal stages:

1. Preprocessing the radiographs and transforming the landmark annotations into the resized image coordinate system.
2. Generating Gaussian target heatmaps and predicting one landmark-specific heatmap per anatomical

landmark with a lightweight U-Net.

3. Converting the predicted heatmaps into continuous landmark coordinates using Soft-Argmax and expressing the resulting localization errors in millimeters.

The validation protocols evaluate the complete framework after these processing stages. They are evaluation designs rather than additional localization operations. In the primary Holdout 70/30 protocol, 72 radiographs are used for development and 30 radiographs are reserved for final testing. The development set is internally divided into training and validation subsets for training-duration selection. After the best training duration has been determined from the internal validation loss, a newly initialized network is refitted from scratch on all 72 development radiographs and evaluated once on the fixed 30-image test set.

The complementary K-Fold 70/30 protocol uses the same 30-image test cohort. The 72 development radiographs are divided into ten folds of seven or eight radiographs, and one model is trained for each fold-specific development partition. During inference, the predicted heatmaps from the ten models are averaged element-wise before Soft-Argmax decoding. The resulting predictions are then evaluated on the same isolated test cohort.

The complementary Nested LOOCV protocol provides a robustness analysis over all 102 radiographs. In each outer iteration, one radiograph is reserved as an external test sample, while the remaining 101 radiographs form the outer-development set. An internal training-validation partition is used only to determine the training duration. The network is subsequently reinitialized and refitted on all 101 outer-development radiographs for the selected number of epoadd. The excluded radiograph is then evaluated once, without being used in training, internal validation, scheduler control, early stopping, or training-duration selection.

3.2 Heatmap-Based Landmark Localization Framework

The network belongs to the family of skip-connected encoder-decoder architectures. U-Net uses a contracting path to learn increasingly abstract representations and an expanding path to reconstruct high-resolution spatial information. Skip connections transfer features from the encoder to corresponding decoder stages, thereby supporting precise localization while retaining contextual information [Ronneberger et al., 2015].

The network is not used for semantic segmentation. The objective is to localize 19 anatomical landmarks. The output therefore contains one continuous two-dimensional heatmap for each landmark, with each output channel representing one anatomical structure.

Compared with direct coordinate regression, Gaussian heatmap regression preserves the spatial structure of the localization problem by distributing supervision over the image grid and around each annotated landmark rather than penalizing only a final coordinate vector. The predicted heatmaps are then converted into continuous coordinates through Soft-Argmax, a differentiable alternative to classical Argmax that is conceptually related to the Differentiable Spatial to Numerical Transform [Nibali et al., 2018]. Soft-Argmax does not introduce additional trainable parameters in the present implementation. It provides the coordinate-decoding operation applied after

heatmap prediction, whereas the U-Net remains a heatmap-regression model.

Accordingly, the methodological contribution lies in the controlled integration of data preprocessing, Gaussian target generation, lightweight U-Net regression, continuous coordinate decoding, image-specific physical conversion, and a coherent evaluation design consisting of the primary Holdout 70/30 protocol and the complementary K-Fold 70/30 and Nested LOOCV protocols. The contribution is therefore methodological and experimental rather than architectural.

3.3 Computational Environment

The experiments were conducted on a Dell G15 5530 laptop equipped with a 13th-generation Intel Core i7-13650HX processor operating at 2.60 GHz, 16 GB of DDR5 memory operating at 4800 MT/s, and a 954 GB solid-state drive formatted with the NTFS file system. Deep-learning computations were accelerated with an NVIDIA GeForce RTX 4060 Laptop GPU containing 8 GB of dedicated video memory. Figure 3.1 summarizes the hardware and software organization of this experimental environment.

The GPU executed convolutional operations, tensor computations, and model-training steps, while the CPU and system memory handled data loading, preprocessing, experiment organization, and result export, as illustrated in Figure 3.1. The implementation used Python 3.10 under Windows 11 Home Single Language, version 25H2. The principal software libraries and CUDA configuration are summarized in Table 3.1.

Before each experiment, an initialization procedure verified the availability of the selected CUDA device, the accessibility of the PKU dataset, and the correspondence between radiographs and their landmark annotations. It also checked the native image dimensions, the spatial resolution, the resized dimensions, and the axis-specific millimeter-per-pixel conversion factors.

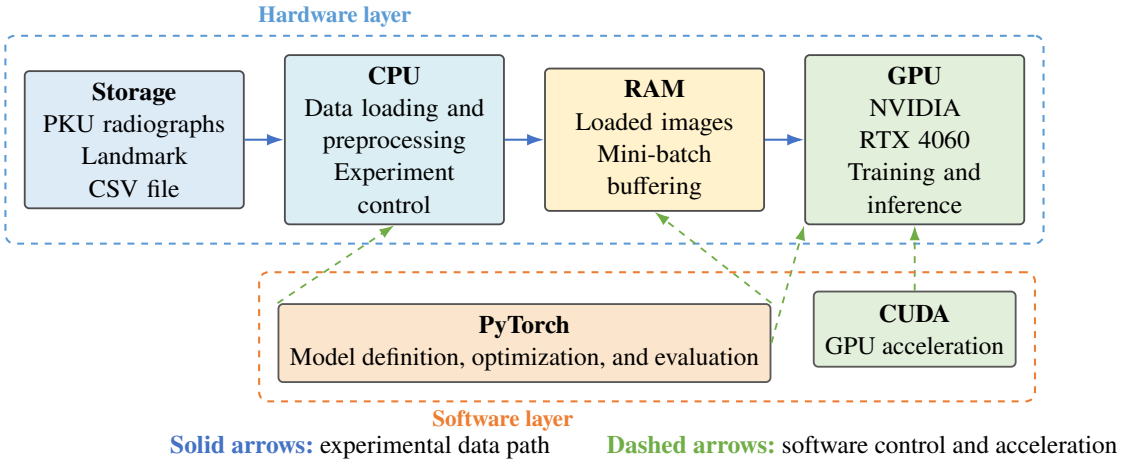


Figure 3.1: Hardware and software architecture of the experimental environment.

Table 3.1. Software dependencies and CUDA configuration used in the implementation.

Category	Component	Version or value
CUDA configuration	CUDA-enabled devices	1
	CUDA device index	0
	PyTorch build	2.5.1+cu121
	CUDA version	12.1
Deep learning	torch	2.5.1
	torchvision	Compatible with PyTorch 2.5.1
Data processing	numpy	1.24.0
	pandas	2.0.0
Image processing	Pillow	10.0.0
	imageio	2.31.0
Visualization	matplotlib	3.7.0
	seaborn	0.12.0

3.4 Methodological Workflow

Figure 3.2 presents the complete workflow, from data preparation and landmark representation to validation, performance evaluation, and result export. Each lateral cephalogram is loaded with its corresponding landmark annotations. The image is resized and converted to a tensor, while the landmark coordinates are remapped from native image space to resized image space. A Gaussian target heatmap is then generated for each landmark.

The primary evaluation uses a fixed Holdout 70/30 protocol. Seventy-two radiographs form the development set, and 30 radiographs form the final test set. The development set is internally divided into training and validation subsets. The internal validation loss is used to determine the training time and then a new initialized network is refitted from scratch on all 72 development radiographs and once on the fixed test set.

We test the same heatmap-regression framework using a K-Fold 70/30 alternative. The 72 development radiographs are split into ten folds of seven or eight radiographs each. Each fold trains one instance of the U-Net (i.e., we have ten trained interations). During inference, the predicted heatmaps tensors are averaged element-wise across all ten models before Soft-Argmax decoding. Predictions are tested on the same 30 image test set used by the main Holdout 70/30 protocol.

Nested LOOCV provides a complementary robustness assessment using all 102 radiographs. For each of the $N = 102$ outer iterations, one radiograph is isolated for external testing. The remaining 101 radiographs are used only for internal model development. An internal training-validation partition determines the training duration. The network is then reinitialized and refitted on all 101 outer-development radiographs for the selected number of epoadd. The refitted model is applied once to the external test radiograph.

During internal selection, gradient updates use only the internal-training subset. The internal-validation

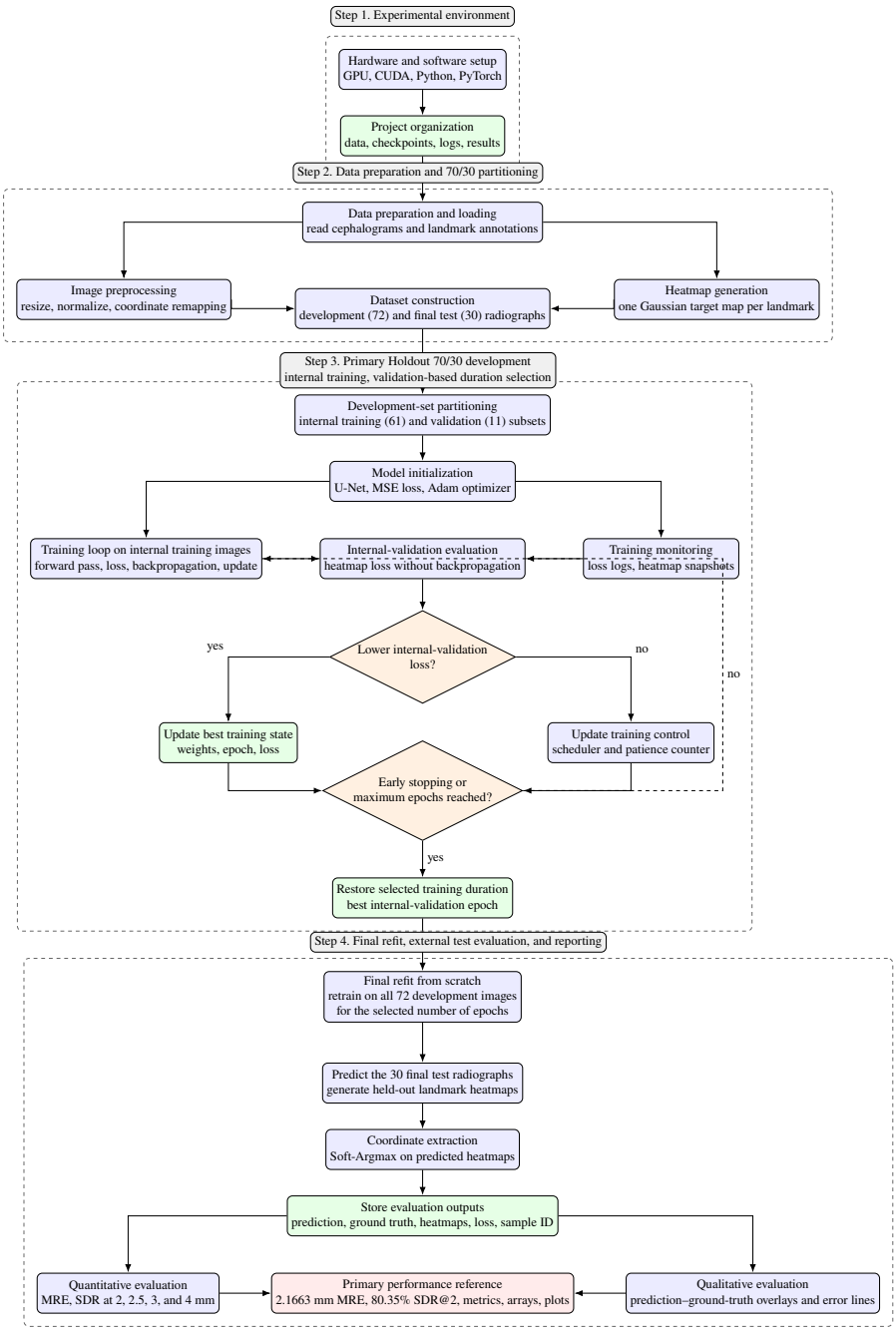


Figure 3.2: *Overall workflow of the cephalometric landmark-localization framework. The Holdout 70/30 protocol provides the primary evaluation, whereas K-Fold 70/30 and Nested LOOCV provide complementary assessments. In the Holdout 70/30 protocol, 72 radiographs are used for development and 30 radiographs are reserved for final testing. The K-Fold 70/30 protocol evaluates the same 30-image test cohort after heatmap-level ensemble inference. In Nested LOOCV, each external radiograph is excluded from training, internal validation, scheduler control, early stopping, and training-duration selection.*

heatmap loss determines the training duration and controls early stopping, but it is never backpropagated. The learning-rate scheduler follows an epoch-dependent cosine schedule and does not receive either the

internal-validation loss or the external-test loss. After training-duration selection, a newly initialized network is trained on the complete development set for exactly the selected number of epoadd. The external test images are then processed in evaluation mode, and their predicted heatmaps are converted into continuous coordinates with Soft-Argmax.

For the primary Holdout 70/30 protocol, the final refit uses all 72 development radiographs and is evaluated once on the 30-image test set. For the K-Fold 70/30 protocol, the ten fold-specific predictions are combined at the heatmap level before coordinate decoding. For Nested LOOCV, the same procedure is repeated across all 102 outer iterations, with one excluded radiograph evaluated after each refit.

The implementation stores predictions, transformed reference coordinates, heatmaps, losses, sample identifiers, and spatial-conversion metadata generated during each experiment. The resulting outputs are aggregated into protocol-specific datasets. Quantitative evaluation is centered on the Mean Radial Error (MRE) and the Successful Detection Rate (SDR) at thresholds of 2.0, 2.5, 3.0, and 4.0 mm. These primary measures are complemented by image-level and landmark-level analyzes, including the median localization error, the standard deviation of radial errors, the distribution of individual errors, landmark-specific MRE and variability, and the Root Mean Squared Error (RMSE) as a complementary measure of sensitivity to large deviations.

3.5 Dataset Overview

This study used the Peking University cephalometric radiograph dataset described by Zeng et al. [2021]. The dataset was developed at the Peking University School and Hospital of Stomatology, where the radiographs were acquired with a Planmeca ProMax 3D cephalometric X-ray system. The dataset is commonly described with a reference native geometry of 2089×1937 pixels and an original spatial resolution of 0.125 mm per native pixel [Khalid et al., 2025]. This 2089×1937 geometry is a reference description of the dataset and is not imposed as a fixed dimension on every file. The implementation reads the native width and height separately for each radiograph and uses those actual dimensions for coordinate transformation and physical-unit conversion.

The dataset contains 102 lateral cephalometric radiographs annotated by two experts. The data comprise an image directory containing radiographs in `.bmp` format and a comma-separated values file containing the coordinates of 19 anatomical landmarks for each radiograph. Each annotation row is associated with one image and follows

$$\mathbf{a}_j = [\text{Image}_j, x_{j,1}, y_{j,1}, x_{j,2}, y_{j,2}, \ldots, x_{j,19}, y_{j,19}], \tag{3.1}$$

where $(x_{j,k}, y_{j,k})$ denotes the reference position of landmark k in radiograph j. The reference positions correspond to centroid coordinates derived from the annotations supplied by the two experts.

3.6 Data Preprocessing

The preprocessing stage comprises image resizing, intensity scaling, landmark-coordinate transformation, Gaussian heatmap generation, and verification of image-annotation correspondence. These operations preserve the correspondence between the anatomical landmarks and the radiographs after transformation from native image space to network input space.

The preprocessing operations are deterministic and are implemented within the dataset-loading procedure. No dataset-level intensity statistics are estimated from the complete dataset. Intensity scaling uses individual image values and maps them to the interval $[0, 1]$.

3.6.1 Image Resizing and Intensity Scaling

Each lateral cephalometric radiograph is resized from its native dimensions to 256×256 pixels using bilinear interpolation. This standardizes network input dimensions and reduces memory and computational requirements. The native dimensions remain available for image-specific coordinate and physical-unit conversion.

Although the radiographs are grayscale, each image is converted to a three-channel representation before conversion to a PyTorch tensor. The three channels contain identical grayscale information and are compatible with the `in_channels=3` configuration of the U-Net. After tensor conversion, image intensities are scaled to $[0, 1]$:

$$\mathbf{I}_j \in \mathbb{R}^{C \times H_{\text{new}} \times W_{\text{new}}} = \mathbb{R}^{3 \times 256 \times 256}, \tag{3.2}$$

where $C = 3$ denotes the number of replicated channels and $H_{\text{new}} = W_{\text{new}} = 256$. Because the native radiographs are rectangular, direct resizing to a square representation produces different horizontal and vertical scale factors.

3.6.2 Landmark Coordinate Transformation

Let $W_{\text{orig}}^{(j)}$ and $H_{\text{orig}}^{(j)}$ denote the native width and height of radiograph j. Let $W_{\text{new}} = 256$ and $H_{\text{new}} = 256$. The axis-specific coordinate-scaling factors are

$$S_x^{(j)} = \frac{W_{\text{new}}}{W_{\text{orig}}^{(j)}}, \qquad S_y^{(j)} = \frac{H_{\text{new}}}{H_{\text{orig}}^{(j)}}, \tag{3.3}$$

For the reference geometry of 2089×1937 pixels,

$$S_x = \frac{256}{2089} \approx 0.1225, \qquad S_y = \frac{256}{1937} \approx 0.1322. \tag{3.4}$$

The implementation nevertheless uses the actual native width and height of every loaded radiograph. For landmark k in radiograph j,

$$\begin{aligned} x_{j,k,\text{new}} &= x_{j,k,\text{orig}} S_x^{(j)}, \\ y_{j,k,\text{new}} &= y_{j,k,\text{orig}} S_y^{(j)}. \end{aligned} \tag{3.5}$$

The transformed coordinates generate the Gaussian target heatmaps and define the reference positions for Soft-Argmax evaluation. The implementation verifies

$$0 \le x_{j,k,\text{new}} < W_{\text{new}}, \qquad 0 \le y_{j,k,\text{new}} < H_{\text{new}}. \tag{3.6}$$

The image identifier is checked against the corresponding annotation row before training or evaluation.

3.6.3 Anisotropic Conversion from Resized Pixels to Millimeters

Let the original physical pixel spacing be

$$r_{\text{orig}} = 0.125 \text{ mm/native pixel}. \tag{3.7}$$

The physical distances represented by one resized pixel along the two axes are

$$M_x^{(j)} = r_{\text{orig}} \frac{W_{\text{orig}}^{(j)}}{W_{\text{new}}}, \qquad M_y^{(j)} = r_{\text{orig}} \frac{H_{\text{orig}}^{(j)}}{H_{\text{new}}}. \tag{3.8}$$

For the reference geometry, these values are

$$\begin{aligned} M_x &= 0.125 \frac{2089}{256} = 1.0200195 \approx 1.0200 \text{ mm/resized pixel}, \\ M_y &= 0.125 \frac{1937}{256} = 0.9458008 \approx 0.9458 \text{ mm/resized pixel}. \end{aligned} \tag{3.9}$$

Let $(\hat{x}_{j,k}, \hat{y}_{j,k})$ denote the prediction in resized image space and $(x_{j,k,\text{new}}, y_{j,k,\text{new}})$ the transformed reference coordinate. The coordinate differences are

$$\Delta x_{j,k} = \hat{x}_{j,k} - x_{j,k,\text{new}}, \qquad \Delta y_{j,k} = \hat{y}_{j,k} - y_{j,k,\text{new}}. \tag{3.10}$$

The axis-specific errors in millimeters are

$$\Delta x_{j,k,\mathrm{mm}} = \Delta x_{j,k} M_x^{(j)}, \qquad \Delta y_{j,k,\mathrm{mm}} = \Delta y_{j,k} M_y^{(j)}. \tag{3.11}$$

The physical radial error is

$$e_{j,k} = \sqrt{\left(\Delta x_{j,k,\mathrm{mm}}\right)^2 + \left(\Delta y_{j,k,\mathrm{mm}}\right)^2}. \tag{3.12}$$

Equivalently,

$$e_{j,k} = r_{\mathrm{orig}} \sqrt{\left[\Delta x_{j,k} \frac{W_{\mathrm{orig}}^{(j)}}{W_{\mathrm{new}}}\right]^2 + \left[\Delta y_{j,k} \frac{H_{\mathrm{orig}}^{(j)}}{H_{\mathrm{new}}}\right]^2}. \tag{3.13}$$

No averaged or isotropic millimeter-per-pixel factor is used in the final evaluation. The conversion expresses the measured discrepancy in physical units; it does not restore anatomical detail lost during resizing.

3.6.4 Landmark Heatmap Generation

The network predicts one two-dimensional response map for each landmark. Let $K = 19$ and let

$$\mathbf{c}_{j,k} = \left(x_{j,k,\mathrm{new}}, y_{j,k,\mathrm{new}}\right) \tag{3.14}$$

be the transformed landmark center. The target heatmap at pixel position (u, v) is

$$\mathcal{H}_{j,k}(u, v) = \exp\left[-\frac{\left(u - x_{j,k,\mathrm{new}}\right)^2 + \left(v - y_{j,k,\mathrm{new}}\right)^2}{2\sigma^2}\right], \tag{3.15}$$

where $u \in \{0, \ldots, W_{\mathrm{new}} - 1\}$ and $v \in \{0, \ldots, H_{\mathrm{new}} - 1\}$ denote the horizontal and vertical pixel coordinates, respectively.

The implementation uses $\sigma = 4$ resized-image pixels. The Gaussian maps are not normalized to sum to one. They are dense scalar supervision maps with values in $(0, 1]$.

The standard deviation σ controls the spatial extent of the supervisory signal. Gaussian heatmaps provide spatially distributed supervision around the annotated position [Thaler et al., 2021]. The 19 maps are stacked as

$$\mathbf{H}_j \in \mathbb{R}^{K \times H_{\mathrm{new}} \times W_{\mathrm{new}}} = \mathbb{R}^{19 \times 256 \times 256}. \tag{3.16}$$

For a mini-batch of size B,

$$\mathbf{H}_B \in \mathbb{R}^{B\times K\times H_{\text{new}}\times W_{\text{new}}} = \mathbb{R}^{B\times 19\times 256\times 256}. \tag{3.17}$$

3.7 Architecture of the Proposed Model

The localization system uses a lightweight two level U-Net. The contracting path captures increasingly abstract anatomical information while the skip connections preserve high resolution spatial information. The small size of the architecture balances representational capacity, computational cost and limited annotated dataset size practically. Architecture is not presented as new U-Net variant.

3.7.1 Detailed Network Structure

The proposed network follows the U-Net paradigm introduced by Ronneberger et al. [2015], whose defining elements, a contracting path for hierarchical feature extraction, an expanding path for spatial reconstruction, and skip connections that carry fine positional information between the two, are precisely the properties that a task of this kind requires. The adaptation to cephalometric landmark localization consists of expressing each landmark as a dense Gaussian response rather than as a point prediction, which is why the network outputs 19 heatmap channels.

The network receives a $3 \times 256 \times 256$ tensor and contains two encoder blocks, one bottleneck, two decoder blocks, and a final output projection. All 3×3 convolutions use one-pixel padding. Spatial resolution changes only through 2×2 max-pooling and 2×2 transposed-convolution operations.

The first encoder block contains two Conv–BN–ReLU sequences, maps 3 channels to 64 channels, and retains a $64 \times 256 \times 256$ feature map for the first skip connection. The first pooling operation produces a $64 \times 128 \times 128$ feature map.

The second encoder block contains two Conv–BN–ReLU sequences, maps 64 channels to 128 channels, and retains a $128 \times 128 \times 128$ feature map for the second skip connection. The second pooling operation produces a $128 \times 64 \times 64$ feature map.

The bottleneck contains two Conv–BN–ReLU sequences and maps 128 channels to 256 channels at 64×64 spatial resolution. The first decoder uses a 2×2 transposed convolution to obtain $128 \times 128 \times 128$ features, concatenates them with the output of Encoder Block 2, and applies two Conv–BN–ReLU sequences.

The second decoder uses a 2×2 transposed convolution to obtain $64 \times 256 \times 256$ features, concatenates them with the output of Encoder Block 1, and applies a Conv–BN–ReLU sequence. A final 1×1 convolution maps 64 channels to 19 landmark heatmaps. No final Sigmoid or Softmax activation is used because the model does direct continuous heatmap regression.

Heatmap-based U-Net approaches have been used before to this task. Zhong et al. [2019], for example, have a global and local stage: the first regresses coarse heatmaps over the whole image and the second refines them on small patches guided by an attention mechanism; the whole problem is formulated as pixel classification. The present design is deliberately simpler: one encoder-decoder network trained on the full image with MSE loss (mean squared error), coordinates recovered continuously from heatmap responses by Soft-Argmax. The comparison with Zhong et al. [2019] is therefore context not lineage: their results show that heatmap learning for cephalometric localization is feasible but they do not prescribe the architecture here.

The channel progression (64 –> 128 and 256 on the contracting side and back to 128 and 64 on the expanding side, see Table 3.2), follows the doubling convention of the original U-Net [Ronneberger et al., 2015] but stops two levels earlier. This truncation is motivated by the task at hand: Localizing cephalometric landmarks reads both the global craniofacial configuration and the fine local detail around each point so two pooling stages are enough to give the network the former while skip connections and shallow decoder keep the latter intact for Soft-Argmax decoding. Going deeper would add little of the former and blur gradually the latter while multiplying parameters that 102 annotated radiographs cannot reliably support; whether this is so in practice is just the kind of question depth and channel-count ablations planned for Section 5.6.5 are meant to answer. For now both depth and channel counts should be taken as fixed design settings chosen for the dataset at hand and the 256×256 input resolution (Figure 3.3), not optima found by an architecture search. Nor shallow design driven by gradient-flow considerations; such effects would require dedicated gradients analyzes (not performed here).

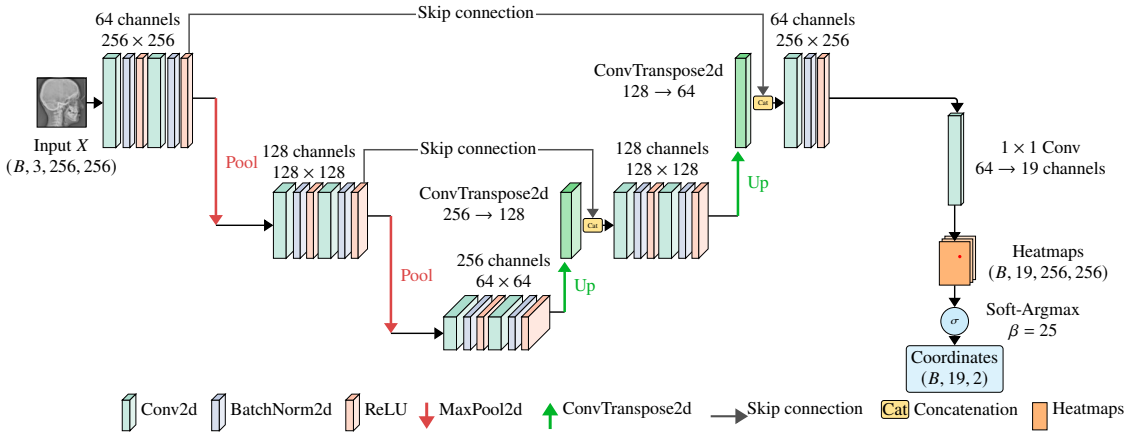


Figure 3.3: *Two-level U-Net architecture used for cephalometric landmark localization. The network receives a $3 \times 256 \times 256$ input tensor, reconstructs high-resolution features through skip connections, and produces 19 landmark heatmaps. Spatial Soft-Argmax with $\beta = 25$ converts the predicted heatmaps into continuous landmark coordinates.*

3.7.2 Output Dimensions and Basic Operations

For one radiograph, the network produces a set of $K = 19$ predicted landmark heatmaps:

Table 3.2. Composition and output dimensions of the functional blocks in the proposed U-Net.

Block	Composition	Output size
Input	RGB-compatible radiograph tensor	$3 \times 256 \times 256$
Encoder 1	Conv 3×3 ($3 \to 64$) and Conv 3×3 ($64 \to 64$), each followed by BN and ReLU	$64 \times 256 \times 256$
Pool 1	MaxPool 2×2, stride 2	$64 \times 128 \times 128$
Encoder 2	Conv 3×3 ($64 \to 128$) and Conv 3×3 ($128 \to 128$), each followed by BN and ReLU	$128 \times 128 \times 128$
Pool 2	MaxPool 2×2, stride 2	$128 \times 64 \times 64$
Bottleneck	Conv 3×3 ($128 \to 256$) and Conv 3×3 ($256 \to 256$), each followed by BN and ReLU	$256 \times 64 \times 64$
Decoder 1	ConvTranspose 2×2 ($256 \to 128$), concatenation with Encoder 2, and two Conv–BN–ReLU sequences	$128 \times 128 \times 128$
Decoder 2	ConvTranspose 2×2 ($128 \to 64$), concatenation with Encoder 1, and Conv–BN–ReLU	$64 \times 256 \times 256$
Output	Conv 1×1 ($64 \to 19$) without a final activation	$19 \times 256 \times 256$
Soft-Argmax	Spatial Softmax with $\beta = 25$ followed by coordinate expectation	19×2

$$\widehat{\mathbf{H}}_j \in \mathbb{R}^{K \times H_{\text{new}} \times W_{\text{new}}} = \mathbb{R}^{19 \times 256 \times 256}. \tag{3.18}$$

For a mini-batch of size B, the predicted heatmaps are represented as

$$\widehat{\mathbf{H}}_B \in \mathbb{R}^{B \times K \times H_{\text{new}} \times W_{\text{new}}} = \mathbb{R}^{B \times 19 \times 256 \times 256}. \tag{3.19}$$

A convolutional block can be written in simplified form as

$$\mathbf{F}_{l+1} = \text{ReLU}\left(\text{BN}\left(\mathbf{W}_l * \mathbf{F}_l + \mathbf{b}_l\right)\right), \tag{3.20}$$

where $\mathbf{F}_l$ is the input feature tensor at layer l, $\mathbf{W}_l$ and $\mathbf{b}_l$ are the convolution kernel and bias parameters, and $*$ denotes convolution. The rectified linear unit is defined by

$$\text{ReLU}(z) = \max(0, z). \tag{3.21}$$

For each channel, max pooling over a 2×2 neighborhood with stride two is written as

$$\mathbf{F}_{\text{pool}}(i, j) = \max_{(u,v)\in\Omega} \mathbf{F}(2i + u, 2j + v), \qquad \Omega = \{0, 1\} \times \{0, 1\}. \tag{3.22}$$

This operation reduces the spatial resolution by a factor of two while retaining the largest activation in each local neighborhood. Decoder upsampling uses learned transposed convolutions:

$$\mathbf{F}_{\text{up}} = \text{ConvTranspose}\left(\mathbf{F}; \mathbf{W}_{\text{up}}\right), \tag{3.23}$$

where $\mathbf{W}_{\text{up}}$ denotes the learned transposed-convolution kernels. Skip connections concatenate the upsampled decoder features with the corresponding encoder features:

$$\mathbf{F}_{\text{concat}} = \text{Concat}\left(\mathbf{F}_{\text{up}}, \mathbf{F}_{\text{encoder}}\right). \tag{3.24}$$

The concatenation is performed along the channel dimension. Because the encoder and decoder feature maps have matching spatial dimensions at each skip connection, the operation preserves the spatial resolution while combining contextual and high-resolution information.

3.8 Heatmap Decoding to Coordinates

3.8.1 Classical Argmax

A conventional heatmap decoder selects the pixel with the maximum response:

$$(\hat{x}, \hat{y}) = \arg\max_{(x,y)} \widehat{\mathbf{H}}(x, y). \tag{3.25}$$

Classical Argmax is non-differentiable and returns an integer grid location. It therefore does not provide direct sub-pixel coordinates and uses only the location of the strongest response. These limitations motivate a differentiable continuous-coordinate decoder.

3.8.2 Soft-Argmax

Let $\widehat{\mathbf{H}}_{j,k}$ denote the predicted heatmap for landmark k in radiograph j. The heatmap is converted into a spatial probability distribution using a temperature-control parameter $\beta > 0$:

$$P_{j,k}(x,y) = \frac{\exp\left(\beta \widehat{\mathbf{H}}_{j,k}(x,y)\right)}{\sum_{v=0}^{H_{\text{new}}-1} \sum_{u=0}^{W_{\text{new}}-1} \exp\left(\beta \widehat{\mathbf{H}}_{j,k}(u,v)\right)}. \tag{3.26}$$

The resulting distribution is non-negative and normalized over the spatial domain:

$$P_{j,k}(x,y) \geq 0, \qquad \sum_{y=0}^{H_{\text{new}}-1} \sum_{x=0}^{W_{\text{new}}-1} P_{j,k}(x,y) = 1. \tag{3.27}$$

The continuous horizontal coordinate is obtained as the spatial expectation of the horizontal coordinate:

$$\hat{x}_{j,k} = \sum_{y=0}^{H_{\text{new}}-1} \sum_{x=0}^{W_{\text{new}}-1} x P_{j,k}(x,y). \tag{3.28}$$

Similarly, the continuous vertical coordinate is obtained as

$$\hat{y}_{j,k} = \sum_{y=0}^{H_{\text{new}}-1} \sum_{x=0}^{W_{\text{new}}-1} y P_{j,k}(x,y). \tag{3.29}$$

The predicted coordinate for landmark k in radiograph j is therefore

$$\widehat{\mathbf{c}}_{j,k} = \left(\hat{x}_{j,k}, \hat{y}_{j,k}\right) \in \mathbb{R}^2. \tag{3.30}$$

For all 19 landmarks, the coordinate output is represented by

$$\widehat{\mathbf{C}}_j \in \mathbb{R}^{19\times 2}. \tag{3.31}$$

Increasing β concentrates the spatial probability distribution around the largest heatmap responses, whereas decreasing β produces a flatter distribution. In the present implementation, $\beta = 25$ is used consistently for all experiments. Because no systematic ablation study of β was conducted, this value is treated as a fixed implementation setting rather than as an empirically demonstrated optimum [Nibali et al., 2018].

The Soft-Argmax operation is differentiable with respect to the heatmap values. It consequently permits the predicted coordinates to be obtained from the heatmaps without introducing the discrete, non-differentiable selection associated with classical Argmax. The resulting coordinates remain expressed in the resized-image coordinate system and are converted into physical millimeters during the evaluation stage using the anisotropic pixel-spacing factors defined in Section 3.6.3.

3.9 Learning Strategy

The three experiments share the same heatmap-regression framework but serve different methodological purposes. The Holdout 70/30 protocol provides the primary evaluation, whereas K-Fold 70/30 and Nested LOOCV provide complementary evidence under alternative data-partitioning strategies. The three protocols use the same lightweight two-level U-Net and are not intended to compare different architectures.

3.9.1 Primary Holdout 70/30 Experiment

The primary Holdout 70/30 experiment uses 70% of the radiographs for model development and reserves the remaining 30% as a final external test set. For the PKU dataset, this corresponds to 72 development radiographs and 30 final-test radiographs.

The 72 development radiographs are further divided reproducibly into an internal-training subset and an internal-validation subset using a validation ratio of 15%. This yields 61 internal-training radiographs and 11 internal-validation radiographs. The internal-training subset is used for parameter optimization, whereas the internal-validation subset is used for model-development decisions, including training-duration selection and checkpoint selection. The 30 final-test radiographs remain completely isolated from parameter optimization, validation, training-duration selection, checkpoint selection, and refitting.

After the training duration has been selected using the internal-validation subset, the network is reinitialized and refitted from scratch on all 72 development radiographs for the selected number of epoadd. The final model is then evaluated once on the isolated 30-radiograph test set. The resulting performance is interpreted as conditional on this fixed Holdout 70/30 partition.

3.9.2 Complementary K-Fold 70/30 Experiment

The complementary K-Fold 70/30 experiment uses the same 72 development radiographs and the same 30-radiograph final-test cohort as the primary Holdout 70/30 experiment. The 72 development radiographs are divided reproducibly into ten folds of seven or eight radiographs.

During each fold iteration, nine folds are used for optimization and one fold is used for internal validation. The internal-validation loss is used to select the training duration and the corresponding model checkpoint. The final 30-radiograph test set is not used during fold-specific optimization, validation, training-duration selection, or checkpoint selection.

This procedure produces ten trained instances of the same two-level U-Net. During external inference, each instance predicts 19 heatmaps for every radiograph in the isolated 30-radiograph test cohort. The ten heatmap tensors are averaged element-wise before coordinate extraction:

$$\overline{\mathbf{H}}_j = \frac{1}{F} \sum_{f=1}^{F} \widehat{\mathbf{H}}_j^{(f)}, \qquad F = 10. \tag{3.32}$$

Soft-Argmax is then applied once to the averaged heatmaps. Thus, the ensemble operation is performed at the heatmap level rather than by averaging the coordinates predicted independently by the ten models. This procedure belongs only to the complementary K-Fold 70/30 experiment.

3.9.3 Complementary Nested LOOCV Partitioning

Nested LOOCV is used as a complementary robustness assessment. The complete PKU dataset contains $N = 102$ radiographs. The procedure consists of 102 outer iterations, with one radiograph held out for external evaluation during each iteration.

For outer iteration i, the dataset is partitioned as

$$\mathcal{D} = \mathcal{D}_{\text{dev}}^{(i)} \cup \mathcal{D}_{\text{test}}^{(i)}, \qquad \mathcal{D}_{\text{dev}}^{(i)} \cap \mathcal{D}_{\text{test}}^{(i)} = \varnothing. \tag{3.33}$$

The subset sizes are

$$\left|\mathcal{D}_{\text{dev}}^{(i)}\right| = N - 1 = 101, \qquad \left|\mathcal{D}_{\text{test}}^{(i)}\right| = 1. \tag{3.34}$$

The held-out radiograph is excluded from training, internal validation, scheduler decisions, early stopping, training-duration selection, refitting, and all other model-development operations. The 101 outer-development radiographs are divided reproducibly into internal-training and internal-validation subsets using a validation ratio of 10%. This yields 90 internal-training radiographs and 11 internal-validation radiographs.

The internal-validation subset is used only to select the training duration and to control early stopping during the selection stage. After the training duration has been selected, the network is reinitialized and refitted from scratch on all 101 outer-development radiographs for exactly the selected number of epoadd. The held-out radiograph is then evaluated once using the refitted model.

3.9.4 Nested LOOCV Training Procedure

Each outer iteration contains an internal selection stage followed by a refitting stage. During the selection stage, a new U-Net, an Adam optimizer, a mean-squared-error criterion, and an epoch-dependent learning-rate scheduler are initialized. The internal-training radiographs are presented in shuffled mini-batches. The internal-validation radiographs are evaluated in model-evaluation mode without gradient computation.

For a mini-batch of size B, the predicted heatmap tensor is

$$\widehat{\mathbf{H}}_B \in \mathbb{R}^{B\times 19\times 256\times 256}. \tag{3.35}$$

At the end of each selection epoch, the model is evaluated only on the internal-validation subset. The validation loss updates the early-stopping counter and determines the selected training duration. It is never backpropagated. The scheduler depends only on the epoch index and does not use internal-validation loss or external-test loss.

After the training duration has been selected, a second, newly initialized U-Net is trained on all 101 outer-development radiographs for exactly the selected number of epoadd. No validation-based early stopping is performed during refitting. The external radiograph is then processed once in evaluation mode under `torch.no_grad()`.

3.9.5 Training Hyperparameters

The initial Adam learning rate is

$$\eta_0 = 3 \times 10^{-4}. \tag{3.36}$$

The final configuration uses `CosineAnnealingLR` with a minimum learning rate of 10^{-6}, a maximum of 250 epoadd, a mini-batch size of eight, early-stopping patience of 20 epoadd during internal selection, and heatmap MSE as the training-duration selection criterion. The input resolution is 256×256, each radiograph is represented using three replicated channels, the Gaussian standard deviation is $\sigma = 4$ pixels, and Soft-Argmax uses $\beta = 25$.

The random seed is

$$\text{SEED} = 42. \tag{3.37}$$

The main configuration is summarized in Table 3.3.

3.9.6 Heatmap Regression Loss

Let $\widehat{\mathbf{H}}_{b,k} \in \mathbb{R}^{H\times W}$ denote the predicted heatmap and $\mathbf{H}^*_{b,k}$ the corresponding target heatmap. The mini-batch heatmap loss is

$$\mathcal{L}_{\text{MSE}} = \frac{1}{BKHW} \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{y=0}^{H-1} \sum_{x=0}^{W-1} \left[\widehat{\mathbf{H}}_{b,k}(x,y) - \mathbf{H}^*_{b,k}(x,y) \right]^2. \tag{3.38}$$

This expression corresponds to the default mean reduction implemented by `nn.MSELoss`. The loss is averaged over all samples, landmark channels, and spatial positions.

Table 3.3. Main hyperparameters shared by the primary Holdout 70/30 protocol and the complementary K-Fold 70/30 and Nested LOOCV experiments.

Hyperparameter	Value
Validation protocols	Holdout 70/30 (primary), K-Fold 70/30, and Nested LOOCV with internal validation
External test radiographs	30 (Holdout 70/30 and K-Fold 70/30), 1 per iteration (Nested LOOCV)
Development radiographs	72 (Holdout 70/30 and K-Fold 70/30), 101 per iteration (Nested LOOCV)
Landmark-level external predictions	570 (Holdout 70/30 and K-Fold 70/30), 1,938 (Nested LOOCV)
Internal-validation ratio (Holdout 70/30)	15%, 11 radiographs
Internal-validation ratio (Nested LOOCV)	10%, 11 radiographs per iteration
Internal-training radiographs	90 per Nested LOOCV iteration
Maximum epoadd	250
Batch size	8
Input size	256×256
Input channels	3
Number of landmarks	19
Heatmap standard deviation σ	4 pixels
Initial learning rate	3×10^{-4}
Optimizer	Adam
Heatmap loss	Mean Squared Error
Scheduler	CosineAnnealingLR
Minimum learning rate	10^{-6}
Early-stopping patience	20 epoadd
Training-duration selection criterion	Minimum internal-validation heatmap MSE
Refit duration	Selected number of epoadd
Soft-Argmax parameter β	25
Random seed	42
Physical-error conversion	Image-specific anisotropic conversion

3.9.7 Training-Duration Selection and Refitting

Let T_i denote the set of epoadd completed during the selection stage of outer iteration i. The selected training duration is

$$t_i^* = \arg\min_{t \in T_i} \mathcal{L}_{\text{ival}}^{(i)}(t). \tag{3.39}$$

Whenever the internal-validation loss improves, the early-stopping counter is reset. Otherwise, the counter increases by one. Training ends when the counter reaches 20 or when 250 epoadd have been completed. A new model is then refitted on all 101 outer-development radiographs for t_i^* epoadd. The refitted state, rather than the selection-stage state, generates the external prediction.

One refitted checkpoint is saved for each completed outer iteration, yielding 102 outer-iteration-specific checkpoints. Predictions, reference coordinates, heatmaps, losses, sample identifiers, and physical-conversion

metadata are stored for subsequent analysis.

3.10 Aggregation of Predictions across Protocols

The primary Holdout 70/30 experiment produces one refitted model whose 30 external test predictions form the primary dataset-level estimate. The complementary K-Fold 70/30 experiment produces ten fold-specific models whose averaged heatmap ensemble is decoded once on the same 30 external test radiographs. Nested LOOCV produces 102 outer-iteration-specific trained instances of the same U-Net. Let

$$\widehat{\mathbf{C}}^{(i)} \in \mathbb{R}^{19\times 2} \tag{3.40}$$

represent the prediction for the external radiograph in outer iteration i. The complete prediction set is

$$\widehat{C}_{\mathrm{LOOCV}} = \left\{\widehat{\mathbf{C}}^{(1)}, \widehat{\mathbf{C}}^{(2)}, \ldots, \widehat{\mathbf{C}}^{(102)}\right\}. \tag{3.41}$$

Each radiograph contributes exactly one external prediction to the LOOCV evaluation. Aggregation therefore collects one prediction per radiograph rather than averaging several predictions for the same radiograph. The radiograph held out during outer iteration i is excluded from parameter optimization, internal validation, scheduler control, early stopping, training-duration selection, checkpoint selection, refitting, and training-monitoring visualization for that iteration.

3.11 Quantitative Evaluation Protocol

The three evaluation protocols are interpreted according to their methodological roles. The Holdout 70/30 protocol provides the primary evaluation on 30 final-test radiographs. The K-Fold 70/30 experiment provides complementary evidence on the same 30-radiograph test cohort, whereas Nested LOOCV provides a complementary robustness assessment across all 102 radiographs.

For the primary Holdout 70/30 and complementary K-Fold 70/30 experiments, the final-test cohort contains 30 radiographs and therefore contributes

$$30 \times 19 = 570 \tag{3.42}$$

landmark-level errors. Nested LOOCV evaluates all 102 radiographs and therefore contributes

$$102 \times 19 = 1{,}938 \tag{3.43}$$

landmark-level errors.

Unless stated otherwise, the metrics use the axis-specific millimeter errors defined in Eq. (3.12).

3.11.1 Mean Radial Error

The Mean Radial Error is defined as

$$\mathrm{MRE} = \frac{1}{NK} \sum_{j=1}^{N} \sum_{k=1}^{K} e_{j,k}. \tag{3.44}$$

It summarizes the average physical distance between predicted and reference landmark positions across the evaluated radiographs and landmarks [Schwendicke et al., 2021]. For Nested LOOCV, $N = 102$ and $K = 19$. For the Holdout 70/30 and K-Fold 70/30 evaluations, $N = 30$ and $K = 19$.

3.11.2 Successful Detection Rate

The Successful Detection Rate at threshold τ is

$$\mathrm{SDR}_{\tau} = \frac{1}{NK} \sum_{j=1}^{N} \sum_{k=1}^{K} \mathbf{1}\left(e_{j,k} \leq \tau\right) \times 100. \tag{3.45}$$

The predefined thresholds are

$$\tau \in \{2.0, 2.5, 3.0, 4.0\} \text{ mm}. \tag{3.46}$$

These thresholds are reported as evaluation criteria and are not treated as universal clinical rules without supporting clinical evidence [Sumon et al., 2025]. The 4.0 mm threshold is more permissive than the lower thresholds and therefore corresponds to a less stringent detection criterion.

3.11.3 Root Mean Squared Error

The Root Mean Squared Error is computed from the individual radial localization errors:

$$\mathrm{RMSE} = \sqrt{\frac{1}{NK} \sum_{j=1}^{N} \sum_{k=1}^{K} e_{j,k}^{2}}. \tag{3.47}$$

RMSE assigns greater weight to large radial deviations and therefore complements the MRE by characterizing

the influence of unusually large localization errors.

3.11.4 Standard Deviation of Localization Error

Let

$$\bar{e} = \frac{1}{NK} \sum_{j=1}^{N} \sum_{k=1}^{K} e_{j,k} \tag{3.48}$$

be the mean radial error. The sample standard deviation is

$$\mathrm{SD}_e = \sqrt{\frac{1}{NK-1} \sum_{j=1}^{N} \sum_{k=1}^{K} \left(e_{j,k} - \bar{e}\right)^2}. \tag{3.49}$$

This statistic summarizes the dispersion of individual radial errors around their mean.

3.11.5 Per-Image and Per-Landmark Evaluation

Global metrics summarize the available landmark-level predictions but do not reveal variation across radiographs or anatomical structures. The per-image mean radial error is

$$\mathrm{MRE}_j = \frac{1}{K} \sum_{k=1}^{K} e_{j,k}, \tag{3.50}$$

and the per-landmark mean error is

$$\mathrm{MRE}_k = \frac{1}{N} \sum_{j=1}^{N} e_{j,k}. \tag{3.51}$$

The corresponding per-landmark standard deviation is

$$\mathrm{SD}_k = \sqrt{\frac{1}{N-1} \sum_{j=1}^{N} \left(e_{j,k} - \mathrm{MRE}_k\right)^2}. \tag{3.52}$$

These analyzes identify radiographs and anatomical structures associated with larger or more variable localization errors. Possible contributing factors, such as anatomical variability, low contrast, overlapping structures, or annotation ambiguity, are considered only when supported by qualitative evidence.

3.11.6 Statistical Reporting of Evaluation Results

For each evaluation protocol, the image-level MRE distribution is summarized using the mean, standard deviation, median, quartiles, minimum, maximum, and, where appropriate, a 95% confidence interval for the mean. These summaries describe both central tendency and between-image variability and are particularly relevant when a small number of difficult radiographs may produce large deviations.

Comparisons between protocols are conditioned on the identity of the evaluated radiographs. When the same Sample IDs are available under two protocols, their image-level MRE values are treated as matched observations. For the Holdout 70/30 and K-Fold 70/30 comparison, the analysis is restricted to the 30 common test radiographs.

Matched protocol results may be compared using the paired Student's t-test and the Wilcoxon signed-rank test as complementary parametric and non-parametric procedures. The paired effect is additionally characterized using Cohen's d_z and Cliff's δ.

When the evaluated radiographs are not matched, protocol-level differences are treated as comparisons between independent samples. Welch's t-test and the Mann–Whitney U test may then be used as complementary parametric and non-parametric procedures, respectively, when the available sample sizes support such analyzes.

Statistical significance is not interpreted in isolation. The direction and magnitude of the observed difference, confidence intervals, effect sizes, sample identity, and validation design are considered jointly. In particular, a non-significant result is not interpreted as proof of equivalence between protocols.

3.12 Qualitative Analysis and Visualization

3.12.1 Overlay Visualization of Predictions

Qualitative assessment superimposes predicted landmarks on the corresponding radiographs. Each overlay contains the resized radiograph, the reference landmark positions, the predicted positions generated by the corresponding model, and line segments connecting each prediction to its reference position. These segments show the direction and approximate magnitude of the localization discrepancy.

For Nested LOOCV, the prediction shown for a radiograph is generated by the model trained during the outer iteration in which that radiograph was held out. For the Holdout 70/30 and K-Fold 70/30 experiments, overlays are generated using the corresponding final-test predictions.

3.12.2 Visual Monitoring of Training Dynamics

Training and internal-validation loss curves are inspected during the training-duration selection stage to assess convergence behavior and fold-dependent optimization dynamics. These curves are generated exclusively from

development data and do not include outer-test observations. The resulting visualizations are diagnostic rather than performance metrics.

For Nested LOOCV, the internal-validation curves are used only during the selection stage. The held-out outer-test radiograph is not used to generate, interpret, or select these curves.

3.12.3 Graphical Performance Analysis

Boxplots summarize the distribution of landmark-specific errors, whereas histograms summarize the complete set of individual radial errors. SDR is plotted across the four predefined localization thresholds. Training and internal-validation loss curves document model-development and training-duration selection behavior.

Prediction–reference overlays and localization error vectors provide qualitative inspection of representative cases and difficult localization failures. Together, these visualizations complement the numerical metrics and help distinguish overall performance from image-specific and landmark-specific behavior.

3.13 Organization of the Experimental Set-up

The primary implementation follows the sequence shown in Algorithm 1. The algorithm separates preprocessing, internal model selection, refitting, external prediction, aggregation, and final evaluation. The complementary Nested LOOCV protocol follows the same internal-selection and refitting logic within each of its 102 outer iterations, as shown in Algorithm 2.

Algorithm 1 Primary Holdout 70/30 evaluation protocol

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{102}$, fixed external test set $\mathcal{D}_{\text{test}}$ of 30 radiographs, development set $\mathcal{D}_{\text{dev}}$ of 72 radiographs, and the defined preprocessing, model, and training criteria
Ensure: One out-of-sample prediction for each external test radiograph and the corresponding evaluation metrics
1: Preprocess radiographs and resize them to 256×256 pixels.
2: Transform landmark coordinates and generate Gaussian target heatmaps.
3: Partition $\mathcal{D}_{\text{dev}}$ into internal-training and internal-validation subsets.
4: Train a newly initialized model using only the internal-training subset.
5: Select the training duration using internal-validation loss and early stopping.
6: Initialize a second new model instance.
7: Refit the new model on all 72 development radiographs for the selected number of epoadd.
8: **for** each external test radiograph x_j in $\mathcal{D}_{\text{test}}$ **do**
9: $\quad \hat{y}_j \leftarrow f_\theta(x_j)$
10: $\quad$ Store $\hat{y}_j$ and the corresponding reference data and metadata.
11: **end for**
12: Aggregate the 30 external predictions.
13: Compute the performance metrics and generate the visualizations.

The external test radiographs enter the workflow only after internal training-duration selection and refitting. They are not used for training, validation, scheduler control, early stopping, epoch selection, or checkpoint

Algorithm 2 Complementary Nested LOOCV evaluation protocol

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{102}$ and the defined preprocessing, model, and training criteria
Ensure: One out-of-sample prediction for each radiograph and the corresponding evaluation metrics

```
Preprocess radiographs and resize them to 256 × 256 pixels.
Transform landmark coordinates and generate Gaussian target heatmaps.
for i = 1, ..., 102 do
    D_test ← {(x_i, y_i)}
    D_dev ← D \ D_test
    Partition D_dev into internal-training and internal-validation subsets.
    Train a newly initialized model using only the internal-training subset.
    Select the training duration using internal-validation loss and early stopping.
    Initialize a second new model instance.
    Refit the new model on all samples in D_dev for the selected number of epoadd.
    ŷ_i ← f_θi(x_i)
    Store ŷ_i and the corresponding reference data and metadata.
end for
Aggregate the 102 external predictions.
Compute the performance metrics and generate the visualizations.
```

selection. This preserves the separation between model development and external evaluation.

The same organizational principle is applied to the three retained experimental protocols. The primary protocol is the fixed Holdout 70/30 design, in which 72 radiographs are used for development and 30 radiographs are reserved for final testing. The complementary K-Fold 70/30 protocol uses the same 72-image development cohort and the same isolated 30-image test cohort. Nested LOOCV provides a complementary robustness assessment by evaluating each of the 102 radiographs once as an outer held-out sample.

3.13.1 Reproducibility Considerations

The random seed is fixed before each experiment to promote reproducible data ordering and network initialization when permitted by the computational environment. All experimental protocols use the same preprocessing procedure, architecture, optimization strategy, and hyperparameter configuration. The batch size is fixed at 8 throughout the experiments.

For each protocol, predicted coordinates, reference coordinates, heatmaps, external-test losses, image identifiers, per-image errors, and per-landmark errors are stored automatically. For the Nested LOOCV protocol, these records are generated separately for each outer iteration and are subsequently aggregated across the 102 held-out radiographs. For the Holdout 70/30 and K-Fold 70/30 protocols, the final-test predictions are associated with the same 30-image test cohort.

These records support the reported metrics and visualizations without requiring the complete training procedure to be repeated. Stored metadata also permits verification that every radiograph was evaluated according to the intended protocol and that the physical conversion factors correspond to the native dimensions of the evaluated image. In particular, the image-specific horizontal and vertical conversion factors are retained together

with the corresponding predictions and reference coordinates.

3.14 Scientific Rationale of the Methodology

The methodology responds to the characteristics of cephalometric landmark localization and to the limited size of the annotated PKU dataset. Reliable evaluation depends not only on network architecture, but also on consistent preprocessing, stable optimization, strict separation of model selection from external evaluation and transparent calculation of physical errors.

The lightweight U-Net combines contextual feature learning with high resolution spatial reconstruction. Skip connections preserve fine anatomical information, and contracting paths capture broader relationships among radiographic structures Ronneberger et al. [2015]. Gaussian heatmap regression models each landmark as a spatially distributed target and densely supervises around its annotated position. Soft-Argmax transforms the response maps to continuous coordinates but avoids the integer-grid restriction of Argmax Nibali et al. [2018].

The main evaluation scheme is the fixed Holdout 70/30 design. It leaves 30 radiographs as a final test cohort that remains unexposed to model development, while the remaining 72 radiographs are divided into internal-training and internal-validation subsets for training duration selection. Once the training duration has been selected, a newly initialized model is refitted from scratch on all 72 development radiographs and evaluated on the 30 isolated test radiographs.

The K-Fold 70/30 protocol gives complementary evidence when training partitions are repeated over the same development cohort. Ten fold specific models are trained on the 72 development radiographs; each model predicts the isolated 30 image test cohort (i.e., takes a single prediction from its own fold), and the heatmaps are averaged elementwise before Soft-Argmax is applied once (this is different from averaging final coordinates because aggregation is done in heatmap space, not coordinates).

Nested LOOCV provides a complementary robustness assessment on the whole set of 102 radiographs. Its defining feature is the strict separation between internal model development and outer evaluation. For each outer iteration, the excluded radiograph is not used for training, internal validation, scheduler control, early stopping, training-duration selection, checkpoint selection, refitting, or monitoring. The training duration is selected using only the internal-validation loss. A newly initialized model is then refitted on all 101 outer-development radiographs for the number of epoadd chosen, and once the excluded radiograph is evaluated.

The three protocols therefore have different but complementary purposes: Holdout 70/30 protocol is the primary fixed cohort evaluation; K-Fold 70/30 evaluates the stability of predictions under repeated development partitions while keeping the same final test cohort; Nested LOOCV evaluates out-of-sample behavior across all radiographs by repeated outer held out evaluations. They are not interchangeable because they have different sampling structures, training procedures, and evaluation populations.

Overall, the scientific contribution is in integrating established components into a reproducible framework for cephalometric landmark localization. We combine a light U-Net, Gaussian heatmap regression with continuous

Soft-Argmax decoding, image specific anisotropic millimeter evaluation, main Holdout 70/30 protocol, and KFold 70/30 and Nested LOOCV experiments without claiming architectural novelty.

3.15 Chapter Conclusion

This chapter presented the methodological framework used for automatic cephalometric landmark localization. It described the computational environment, the PKU dataset, deterministic preprocessing, coordinate transformation, anisotropic physical conversion, Gaussian heatmap generation, the lightweight two-level U-Net, Soft-Argmax decoding, the learning strategy, the primary Holdout 70/30 protocol, the complementary K-Fold 70/30 protocol, the complementary Nested LOOCV protocol, and the quantitative and qualitative evaluation procedures.

The primary Holdout 70/30 protocol uses 72 radiographs for development and 30 radiographs for final testing. The K-Fold 70/30 protocol uses the same development and final-test cohorts while training ten fold-specific models and averaging their predicted heatmaps before coordinate decoding. The Nested LOOCV protocol evaluates all 102 radiographs through 102 outer iterations. In each iteration, the held-out radiograph remains completely isolated from model development, and a newly initialized model is refitted on the remaining 101 radiographs for the training duration selected through internal validation.

The final evaluation uses physical millimeter errors calculated with image-specific horizontal and vertical conversion factors. Performance is summarized primarily through mean radial error (MRE) and successful detection rate (SDR) at 2.0, 2.5, 3.0, and 4.0 mm. These measures are complemented by median error, standard deviation, radial RMSE, per-image error distributions, and landmark-specific errors. Statistical comparisons are performed using appropriately paired image-level errors when the same radiographs are evaluated by different protocols.

Qualitative evaluation is based on prediction-reference overlays, localization error vectors, and training-monitoring plots. For Nested LOOCV, the overlays are predictions of the model corresponding to the outer iteration. This design lays the methodological groundwork for experimental results and their interpretation in the following chapter.

Chapter 4

Results and Discussion

4.1 Introduction

This chapter presents and discusses the experimental results obtained with the proposed cephalometric landmark-localization framework on the PKU Cephalometric Dataset. In accordance with the methodology described in Chapter 3, the analysis considers the Holdout 70/30, K-Fold 70/30, and Nested Leave-One-Out Cross-Validation (Nested LOOCV) evaluation protocols, with the Holdout 70/30 design providing the primary evaluation. The results are first presented at the global, image, and landmark levels, followed by cross-protocol comparisons, paired statistical analyzes, error analysis, comparison with the state of the art, and integrated methodological interpretation.

All localization errors are expressed in millimeters using the image-specific anisotropic conversion factors defined in Chapter 3, and the Mean Radial Error (MRE) and Successful Detection Rate (SDR) follow the definitions given in Chapter 3; these definitions are therefore not repeated here. Particular attention is given to performance variability across evaluation protocols, folds, radiographs, and anatomical landmarks.

4.2 Experimental Configuration

The experiments use the PKU Cephalometric Dataset, comprising 102 lateral cephalometric radiographs annotated with 19 anatomical landmarks. The Holdout 70/30 protocol constitutes the primary evaluation and uses 72 radiographs for model development (61 internal training, 11 internal validation, selected epoch 250) and 30 radiographs as a final held-out test set evaluated once. The K-Fold 70/30 ensemble trains ten fold-specific models on the same development data and evaluates their heatmap-averaged predictions on the same 30-image test cohort, enabling paired comparisons. The Nested LOOCV protocol excludes each of the 102 radiographs once

from model development and evaluates it exactly once after refitting on the remaining 101 radiographs, yielding $102 \times 19 = 1{,}938$ landmark-level predictions.

Localization errors are computed in millimeters through anisotropic, per-image conversion, defined according to Eq. (4.1):

$$M_x^{(j)} = 0.125 \, \frac{W_{\text{orig}}^{(j)}}{256}, \qquad M_y^{(j)} = 0.125 \, \frac{H_{\text{orig}}^{(j)}}{256}, \tag{4.1}$$

where $W_{\text{orig}}^{(j)}$ and $H_{\text{orig}}^{(j)}$ denote the native dimensions of radiograph j. For the reference geometry 2089×1937 pixels, these factors are approximately $M_x \approx 1.0200$ and $M_y \approx 0.9458$ mm/pixel; they are provided only as reference values since the actual factors are retrieved separately for each radiograph. No fixed average spacing is used. The localization network is the lightweight two-level U-Net with Gaussian heatmap targets and differentiable Soft-Argmax decoding ($\beta = 25$) defined in Chapter 3. Table 4.1 summarizes the main experimental parameters and the three validation protocols.

Table 4.1. *Summary of the experimental configuration and validation protocols.*

Parameter	Value
Dataset	PKU Cephalometric Dataset
Total images / landmarks	102 / 19
Image preprocessing	Bilinear resizing to 256×256 pixels
Native spatial resolution	0.125 mm/pixel
Conversion mode	Anisotropic, per-image ($M_x \approx 1.0200$, $M_y \approx 0.9458$ mm/resized pixel for the reference geometry)
Network architecture	Two-level U-Net, Gaussian heatmaps
Coordinate extraction	Soft-Argmax, $\beta = 25$
Batch size / maximum epoadd	8 / 250
Primary protocol	Holdout 70/30: 72 development (61 train, 11 validation), 30 test
Additional validation	K-Fold 70/30 ensemble on the same 30 test radiographs
Robustness protocol	Nested LOOCV, 102 outer folds, 1,938 landmark-level predictions

4.3 Primary Results: Holdout 70/30

4.3.1 Experimental Configuration

The primary experiment used the PKU cephalometric dataset containing 102 lateral cephalograms. The dataset was divided into 72 development images and 30 final test images. The development subset was further divided

into 61 internal training images and 11 internal validation images. The final test set remained isolated from model training, validation, checkpoint selection, and refitting.

The best training epoch selected during the internal development procedure was epoch 250. After checkpoint selection, the final model was refitted from scratch using all 72 development images and was evaluated once on the 30-image final test set.

Table 4.2 summarizes the configuration of the primary protocol.

Table 4.2. *Experimental configuration of the primary Holdout 70/30 protocol.*

Item	Value
Total dataset	102 images
Development subset	72 images
Final test subset	30 images
Internal training subset	61 images
Internal validation subset	11 images
Selected best epoch	250
Final refit	From scratch on 72 development images
Final test usage	One final evaluation
Physical conversion	Per-image anisotropic X/Y
Native dimensions	Preserved individually per image
Mean X/Y spacing	Not used

4.3.2 Global Performance

The final Holdout 70/30 evaluation produced a mean test heatmap loss of 0.00011049. The mean radial error was 2.1663 mm, with an image-level standard deviation of 3.8597 mm. The successful detection rates were 80.35% at 2.0 mm, 87.37% at 2.5 mm, 91.40% at 3.0 mm, and 95.79% at 4.0 mm.

Table 4.3 reports the definitive global results of the primary protocol. These values constitute the principal performance reference of this study.

Table 4.3. *Global performance on the 30-image Holdout 70/30 final test set.*

Metric	Value
Number of test images	30
Number of landmark predictions	570
Mean test heatmap loss	0.00011049
Mean radial error (MRE)	2.166340 mm
Standard deviation of image-level MRE	3.859689 mm
SDR@2 mm	80.350877%
SDR@2.5 mm	87.368421%
SDR@3 mm	91.403509%
SDR@4 mm	95.789474%

4.3.3 Image-Level Results

The final test set contained the following 30 image identifiers: {3, 9, 14, 15, 18, 24, 25, 32, 35, 36, 37, 39, 40, 43, 45, 48, 49, 52, 59, 63, 64, 70, 85, 88, 89, 90, 91, 95, 98, 99}.

Table 4.4 reports the image-level MRE, final test heatmap loss, and SDR values obtained for each test image. These image-level MRE values are the quantities used for the paired statistical comparisons reported in Section 4.5.

Table 4.4. *Image-level results for the Holdout 70/30 final test set.*

ID	MRE [mm]	Loss	SDR@2	SDR@2.5	SDR@3	SDR@4
3	1.541119	0.00011353	78.95%	89.47%	94.74%	94.74%
9	1.337113	0.00010264	84.21%	84.21%	89.47%	94.74%
14	1.481768	0.00011114	78.95%	84.21%	89.47%	100.00%
15	2.709012	0.00013452	78.95%	89.47%	89.47%	94.74%
18	22.514344	0.00026592	57.89%	57.89%	63.16%	68.42%
24	1.599024	0.00014187	68.42%	89.47%	94.74%	94.74%
25	1.359307	0.00010794	84.21%	94.74%	94.74%	100.00%
32	1.493436	0.00010904	73.68%	73.68%	89.47%	100.00%
35	1.167955	0.00007163	84.21%	94.74%	100.00%	100.00%
36	1.258562	0.00010026	84.21%	94.74%	100.00%	100.00%
37	1.548483	0.00012350	78.95%	84.21%	84.21%	94.74%
39	1.273263	0.00008909	94.74%	94.74%	94.74%	100.00%
40	1.200943	0.00008869	73.68%	89.47%	89.47%	100.00%
43	1.364673	0.00010236	84.21%	89.47%	89.47%	89.47%
45	2.248880	0.00012922	78.95%	89.47%	89.47%	89.47%
48	1.174740	0.00007682	89.47%	89.47%	94.74%	100.00%
49	1.360714	0.00010658	78.95%	78.95%	94.74%	94.74%
52	1.309833	0.00008682	89.47%	89.47%	100.00%	100.00%
59	1.350564	0.00009837	89.47%	89.47%	89.47%	94.74%
63	1.132458	0.00008096	84.21%	94.74%	94.74%	100.00%
64	1.574088	0.00012439	68.42%	78.95%	89.47%	94.74%
70	1.177795	0.00009630	89.47%	89.47%	89.47%	100.00%
85	1.180253	0.00009267	89.47%	89.47%	94.74%	100.00%
88	2.241306	0.00014112	73.68%	89.47%	89.47%	89.47%
89	1.275705	0.00008866	78.95%	89.47%	89.47%	100.00%
90	1.265356	0.00009292	78.95%	94.74%	94.74%	94.74%
91	1.419350	0.00010946	84.21%	84.21%	94.74%	94.74%
95	1.536341	0.00011460	73.68%	84.21%	89.47%	94.74%
98	1.195050	0.00009558	78.95%	89.47%	94.74%	100.00%
99	1.698750	0.00011814	78.95%	89.47%	89.47%	94.74%

The image-level MRE values in Table 4.4 range from 1.13 mm to 22.51 mm. Twenty-nine of the 30 test radiographs had an image-level MRE below 2.71 mm, whereas image ID 18 exhibited a markedly larger error of 22.51 mm. Consequently, the image-level distribution is strongly right-skewed, with the extreme error of ID 18 contributing substantially to the observed dispersion.

4.3.4 Landmark-Level Results

The final test set generated $30 \times 19 = 570$ landmark-level radial errors. Table 4.5 presents the MRE, standard deviation, and SDR values for each of the 19 landmarks.

Table 4.5. Landmark-level results for the Holdout 70/30 final test set.

Landmark	MRE [mm]	SD [mm]	SDR@2	SDR@2.5	SDR@3	SDR@4
L01	1.703701	3.073314	83.33%	86.67%	93.33%	93.33%
L02	2.569072	6.416730	90.00%	90.00%	90.00%	93.33%
L03	1.648309	0.981030	60.00%	83.33%	90.00%	100.00%
L04	2.113429	1.306109	60.00%	66.67%	83.33%	90.00%
L05	1.150598	0.584453	90.00%	100.00%	100.00%	100.00%
L06	1.522364	1.006935	76.67%	83.33%	96.67%	96.67%
L07	1.360140	0.859319	86.67%	93.33%	93.33%	96.67%
L08	0.708944	0.374452	100.00%	100.00%	100.00%	100.00%
L09	0.874397	0.433061	96.67%	100.00%	100.00%	100.00%
L10	3.360521	3.714300	40.00%	50.00%	60.00%	80.00%
L11	1.079907	0.608334	93.33%	100.00%	100.00%	100.00%
L12	0.950859	0.563346	90.00%	100.00%	100.00%	100.00%
L13	3.300687	13.350860	96.67%	96.67%	96.67%	96.67%
L14	4.256133	19.128511	96.67%	96.67%	96.67%	96.67%
L15	3.432805	13.902500	93.33%	96.67%	96.67%	96.67%
L16	6.033069	20.753777	63.33%	80.00%	83.33%	93.33%
L17	2.175341	1.271347	56.67%	63.33%	76.67%	93.33%
L18	1.609962	1.027400	70.00%	83.33%	86.67%	96.67%
L19	1.310213	1.191951	83.33%	90.00%	93.33%	96.67%

The landmark-level MRE values in Table 4.5 range from 0.71 mm for L08 to 6.03 mm for L16. Three landmarks achieved an MRE below 1 mm (L08, L09, and L12), five landmarks achieved an MRE below 1.5 mm (L08, L09, L11, L12, and L05), and eleven landmarks achieved an MRE below 2 mm. In contrast, L16, L14, L15, L10, and L13 exhibited the largest average errors. The large standard deviations associated with L13, L14, L15, and L16 indicate substantial variability across radiographs. For L13, L14, and L15, the high SDR@2 values suggest that their large standard deviations are mainly caused by occasional extreme deviations. In contrast, L16 combines a large standard deviation with a lower SDR@2, indicating that its errors are both more frequent and more variable.

4.3.5 Graphical Analysis

The graphical analysis is organized by function. The global behavior of the localization errors is documented by the image-level MRE distribution across the 30 final test radiographs (Figure 4.1), by the distribution of the 570 landmark-level radial errors (Figure 4.2), and by the SDR as a function of the error threshold (Figure 4.3).

The per-landmark behavior is summarized by the boxplot of the radial errors across the 19 landmarks (Figure 4.4), with complementary visualizations provided by the violin plots (Figure 4.5), the landmark-level error heatmap (Figure 4.6), and the radar chart of the landmark-specific performance (Figure 4.7).

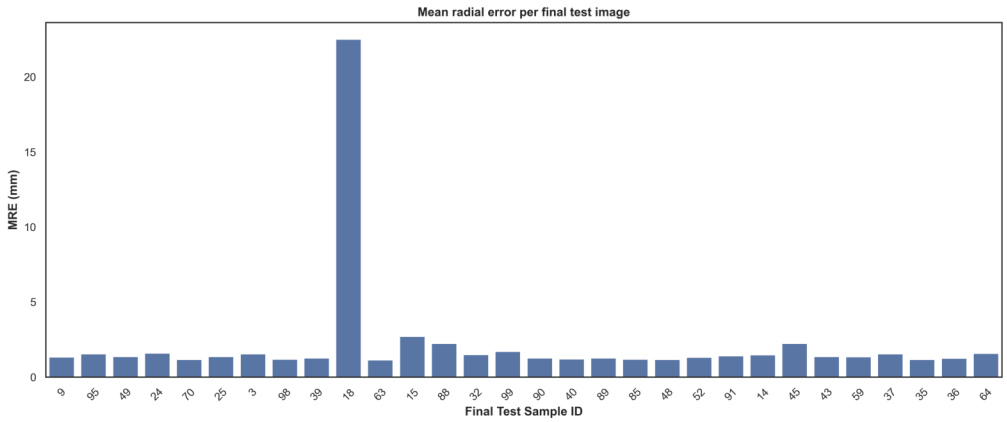


Figure 4.1: Image-level MRE values for the Holdout 70/30 final test set.

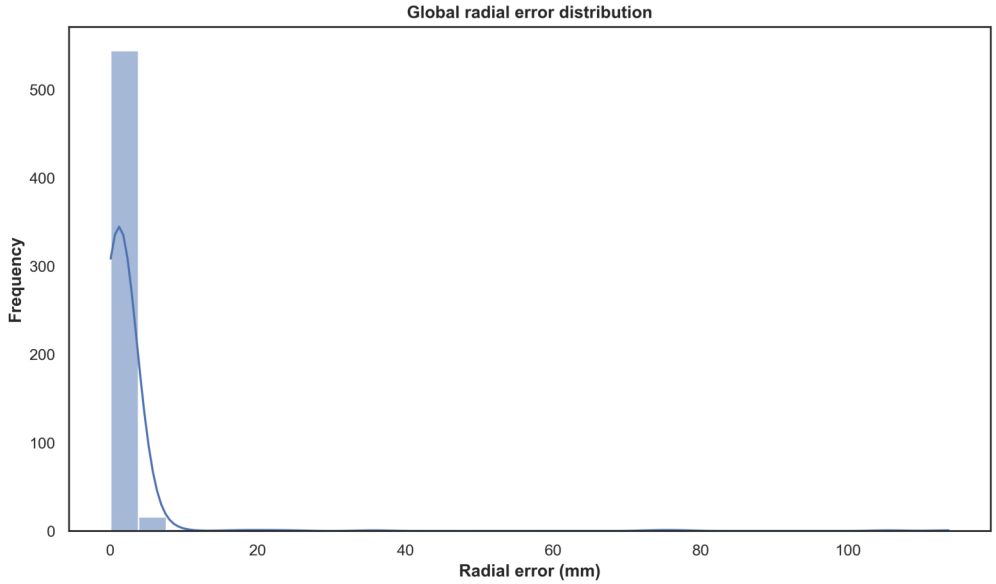


Figure 4.2: Distribution of the landmark-level radial errors in the Holdout 70/30 protocol.

The relation between predictions and references is shown by the ground-truth versus predicted coordinate scatter (Figure 4.8), and the convergence of the final model refitted on the 72 development images is documented by its training curve (Figure 4.9).

4.3.6 Interpretation

The primary Holdout 70/30 protocol achieved a mean radial error of 2.1663 mm and an SDR@2 of 80.35% on the 30-image final test set. Because the final test radiographs remained completely isolated from training, validation, checkpoint selection, and refitting, these values constitute the principal out-of-sample performance estimate of the proposed pipeline.

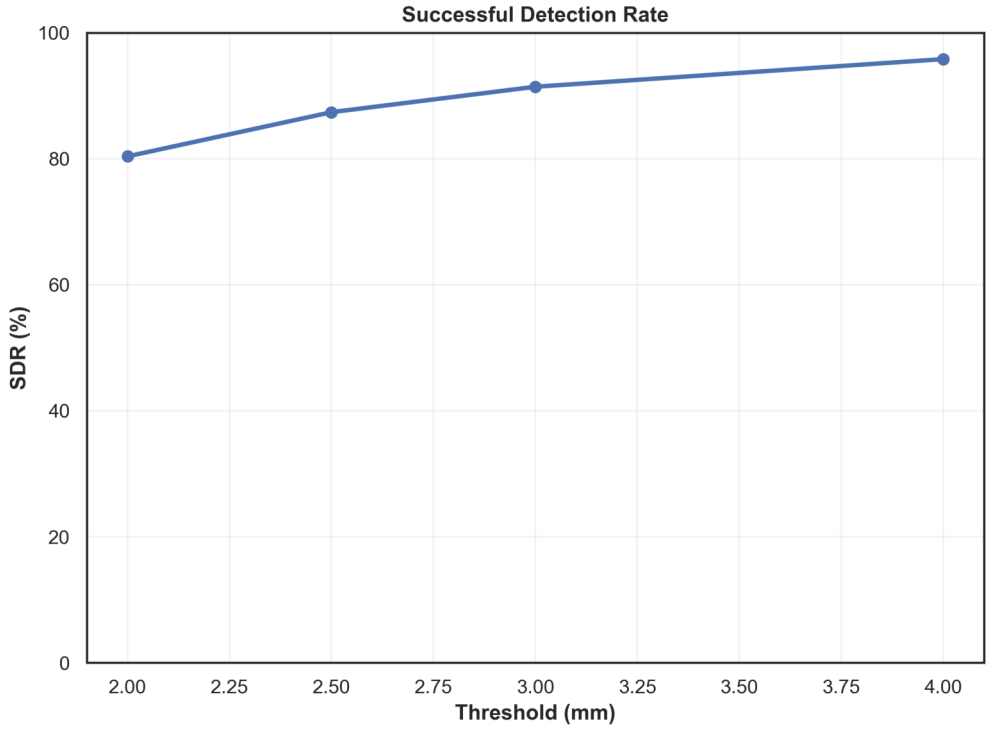


Figure 4.3: Successful Detection Rate as a function of the error threshold in the Holdout 70/30 protocol.

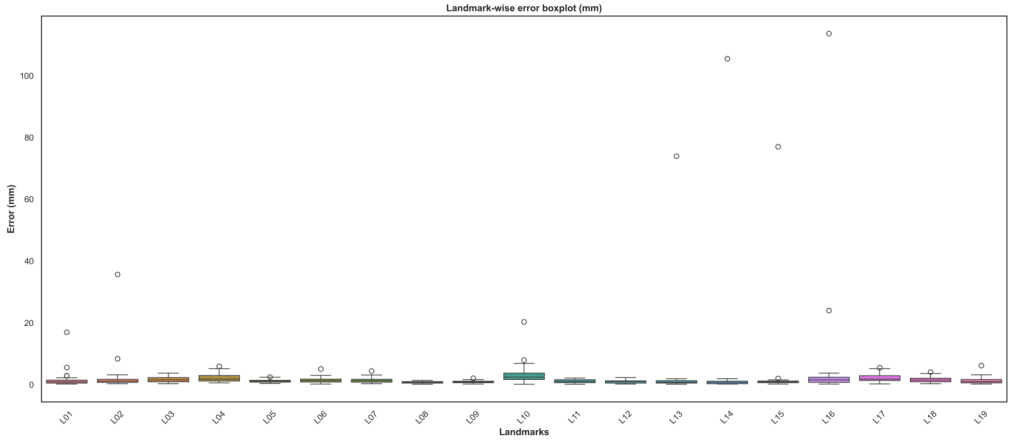


Figure 4.4: Distribution of radial errors across the 19 landmarks in the Holdout 70/30 protocol.

The interpretation of the per-landmark patterns, including the distinction between landmarks whose errors are occasional and extreme (L13, L14, L15) and those whose errors are frequent and variable (L16, L10), is developed in Section 4.5.3.

These results should be interpreted together with the landmark-level and image-level distributions rather than through the global mean alone. The K-Fold 70/30 and Nested LOOCV experiments reported in the following

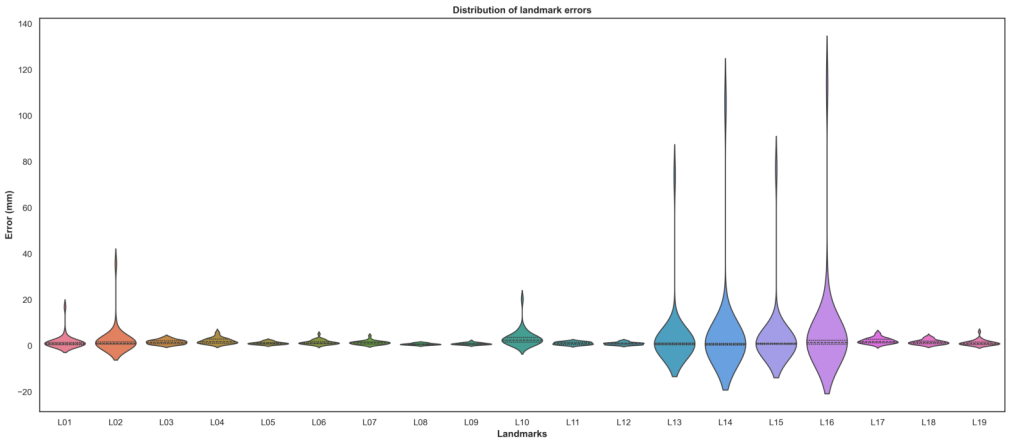


Figure 4.5: Violin plots of the landmark-level radial errors in the Holdout 70/30 protocol.

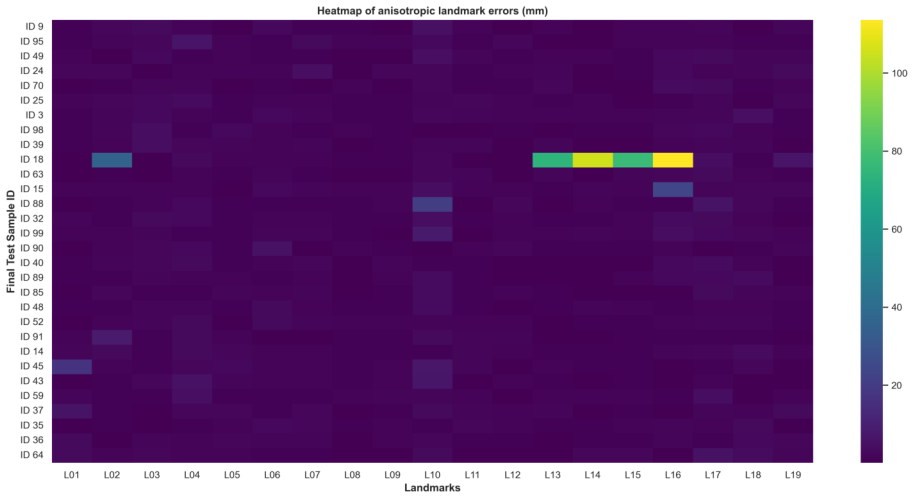


Figure 4.6: Heatmap of the landmark-level radial errors in the Holdout 70/30 protocol.

sections are interpreted as complementary validation and robustness analyzes of the same pipeline rather than as independent test-set evaluations.

4.3.6.1 Qualitative Analysis of Representative Cases

Tables 4.6 to 4.9 present the per-landmark localization results for four representative test images from the primary Holdout 70/30 protocol. The tables include the ground-truth and predicted coordinates, localization errors in the resized image space and in millimeters, as well as image-level performance metrics. The corresponding qualitative overlays are shown in Figure 4.10.

The four selected images illustrate different levels of localization performance. Images ID 63 and ID 90 represent relatively accurate localization cases, whereas ID 45 illustrates a more heterogeneous prediction pattern with a larger individual landmark deviation. Image ID 18 represents an extreme case with a substantially larger

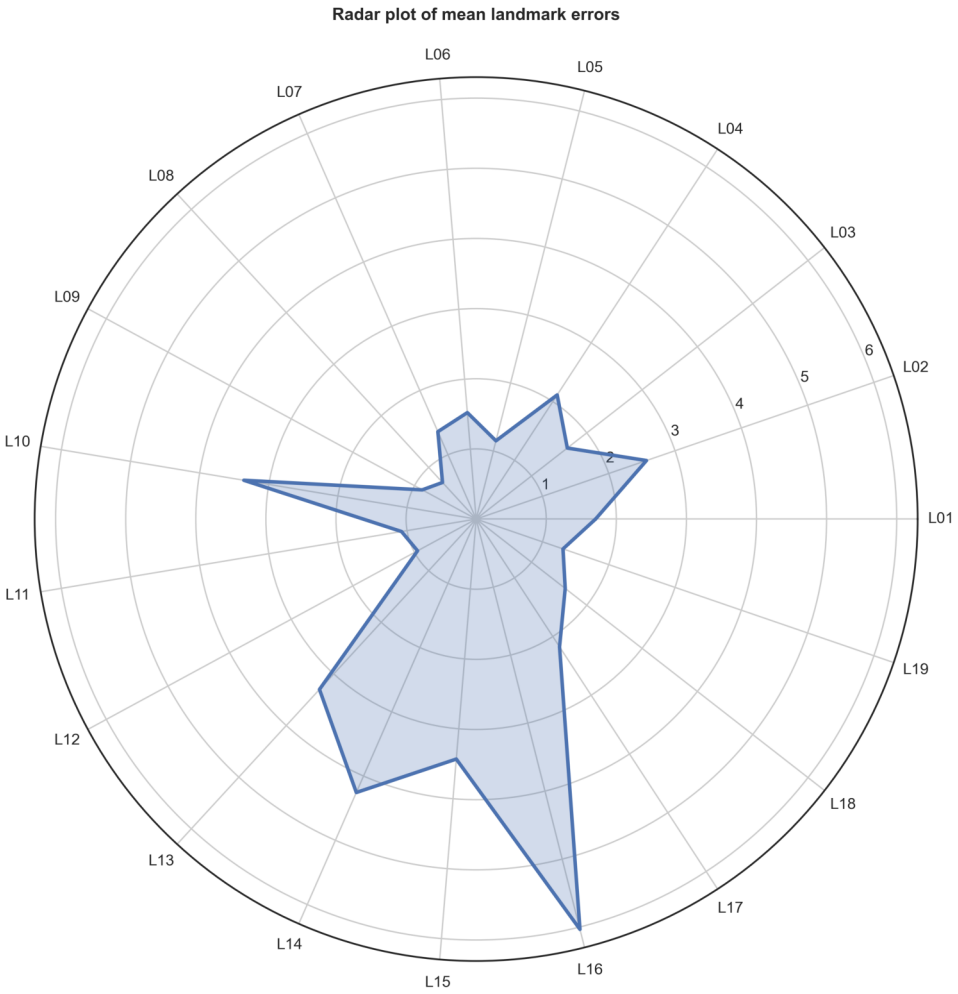


Figure 4.7: *Landmark-specific performance represented as a radar chart in the Holdout 70/30 protocol.*

image-level MRE and a pronounced maximum landmark error. Taken together these examples show variability between radiographs even when using the same trained model and evaluation protocol.

Figure 4.10(a) and Figure 4.10(b), corresponding to ID 63 and ID 90, respectively show the most consistent localization behavior of the four examples. Image ID 63 obtained an image-level MRE of about 1.1325mm while ID90 had MRE of about 1.2654 mm. Their low MRE values mean that most of the 19 landmarks were localized close to their reference positions.

Figure 4.10(c) shows ID 45 with more heterogeneous error distribution. This image achieved an MRE of approximately 2.2489 mm. Although its average error remained moderate, the presence of one or more larger landmark deviations increased the image-level dispersion and contributed to the difference between the mean error and the errors observed for most individual landmarks.

The most pronounced case is shown in Figure 4.10(d), corresponding to ID 18. This image reached an

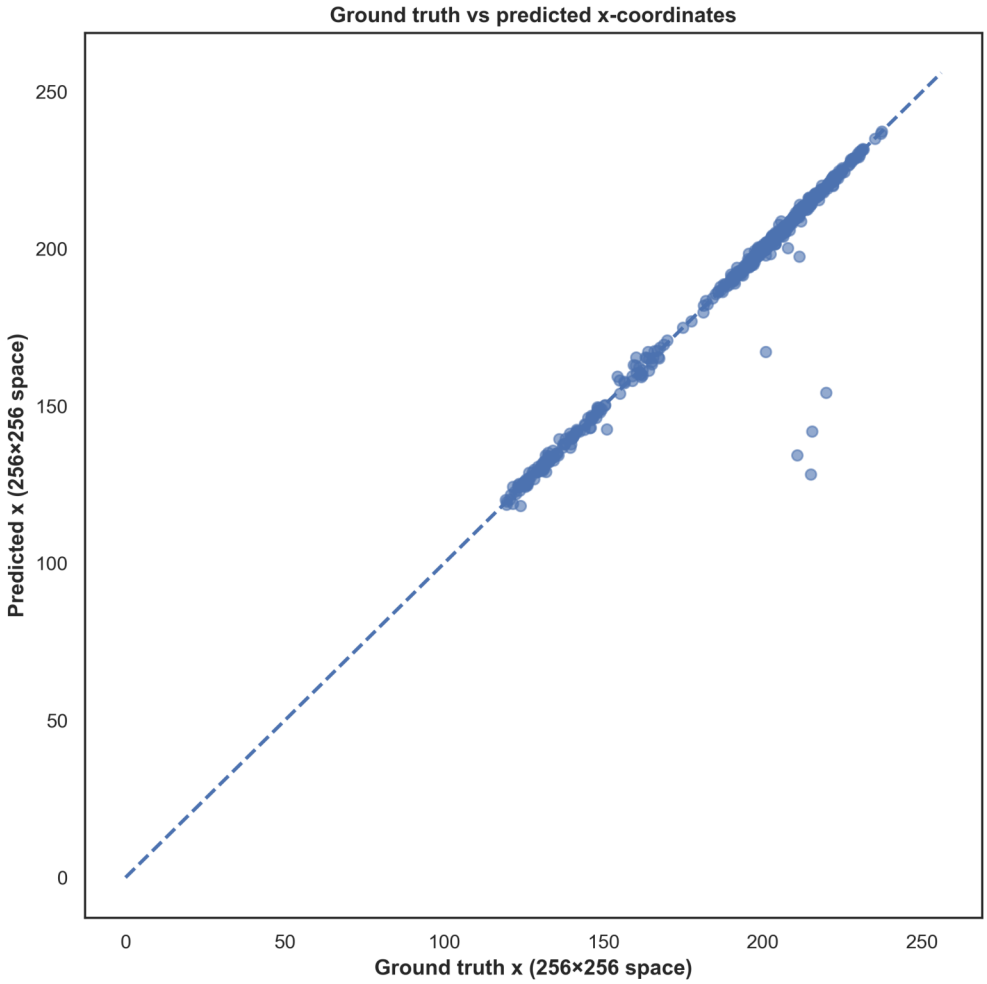


Figure 4.8: Ground-truth versus predicted landmark coordinates in the Holdout 70/30 protocol.

image-level MRE of approximately 22.5143 mm and a maximum individual landmark error of approximately 113.6368 mm. This substantial deviation has a strong influence on the image-level summary statistics and illustrates how a small number of large localization errors can affect the mean error of an individual radiograph.

These examples demonstrate that image-level MRE should not be interpreted in isolation. The per-landmark results and qualitative overlays provide additional information by distinguishing generally consistent localization from cases affected by isolated or severe landmark failures. The extreme case represented by ID 18 should therefore be interpreted as an individual difficult case rather than as a direct description of the predominant localization behavior across the complete test set.

Table 4.6 summarizes the results for image ID 63, shown in Figure 4.10(a). This image achieved an image-level MRE of approximately 1.1325 mm. The corresponding landmark-level results show that the predicted coordinates remained close to the ground-truth coordinates for most of the 19 landmarks. The maximum error was

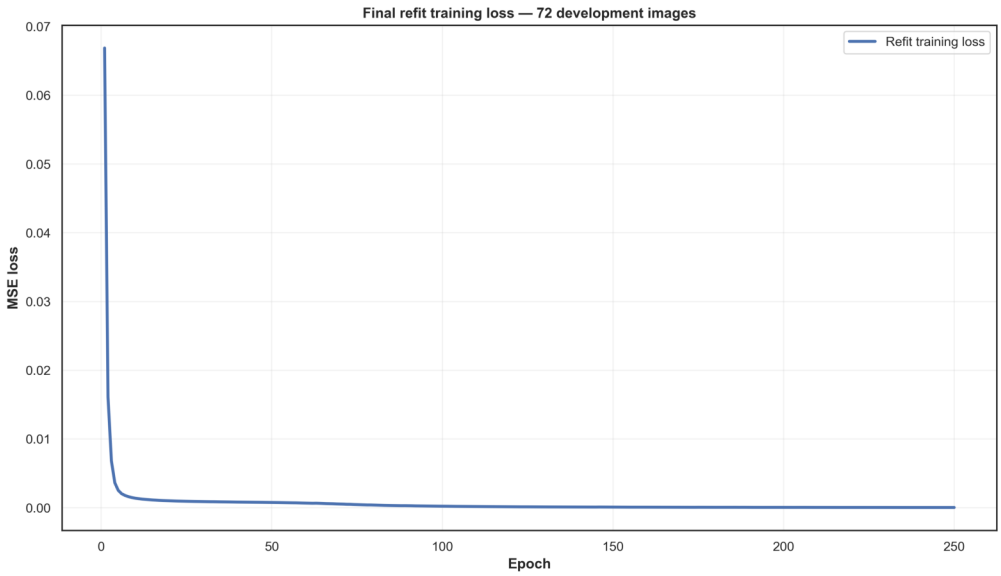


Figure 4.9: *Training curve of the final model refitted on the 72 development images.*

approximately 3.0922 mm, indicating that no very large individual deviation was observed in this representative case. The relatively low MRE and the limited maximum error suggest a generally consistent localization pattern across the image.

Table 4.7 presents the results for image ID 90, corresponding to Figure 4.10(b). The image achieved an image-level MRE of approximately 1.2654 mm. Its maximum landmark error was approximately 5.0771 mm, and its SDR@4 was 94.74%, indicating that only one of the 19 landmarks exceeded the 4 mm threshold. The difference between the image-level MRE and the maximum error shows that the larger deviation was restricted to a small portion of the landmark set rather than being representative of all 19 landmarks.

Table 4.8 reports the results for image ID 45, shown in Figure 4.10(c). This image obtained an image-level MRE of approximately 2.2489 mm. Compared with IDs 63 and 90, the larger MRE indicates a less homogeneous localization pattern. Nevertheless, the error distribution should be interpreted at the landmark level because the image-level mean may be influenced by a limited number of relatively large deviations. The maximum error for this image was approximately 16.9317 mm. Its SDR@4 was 89.47%, indicating that two of the 19 landmarks exceeded the 4 mm threshold. These relatively large deviations increased the image-level MRE and dispersion.

Table 4.9 presents the results for image ID 18, corresponding to Figure 4.10(d). This image represents the most difficult case among the four selected examples. Its image-level MRE reached approximately 22.5143 mm. For ID 18, six of the 19 landmarks exceeded the 4 mm threshold, including several very large deviations. The maximum individual error was approximately 113.6368 mm. Therefore, the image-level MRE was strongly affected by a small subset of severe landmark failures rather than by a uniform degradation across all landmarks.

This case illustrates the importance of reporting both average and landmark-level errors. The MRE summarizes the mean radial error across the 19 landmarks, whereas the maximum error reveals whether the image contains

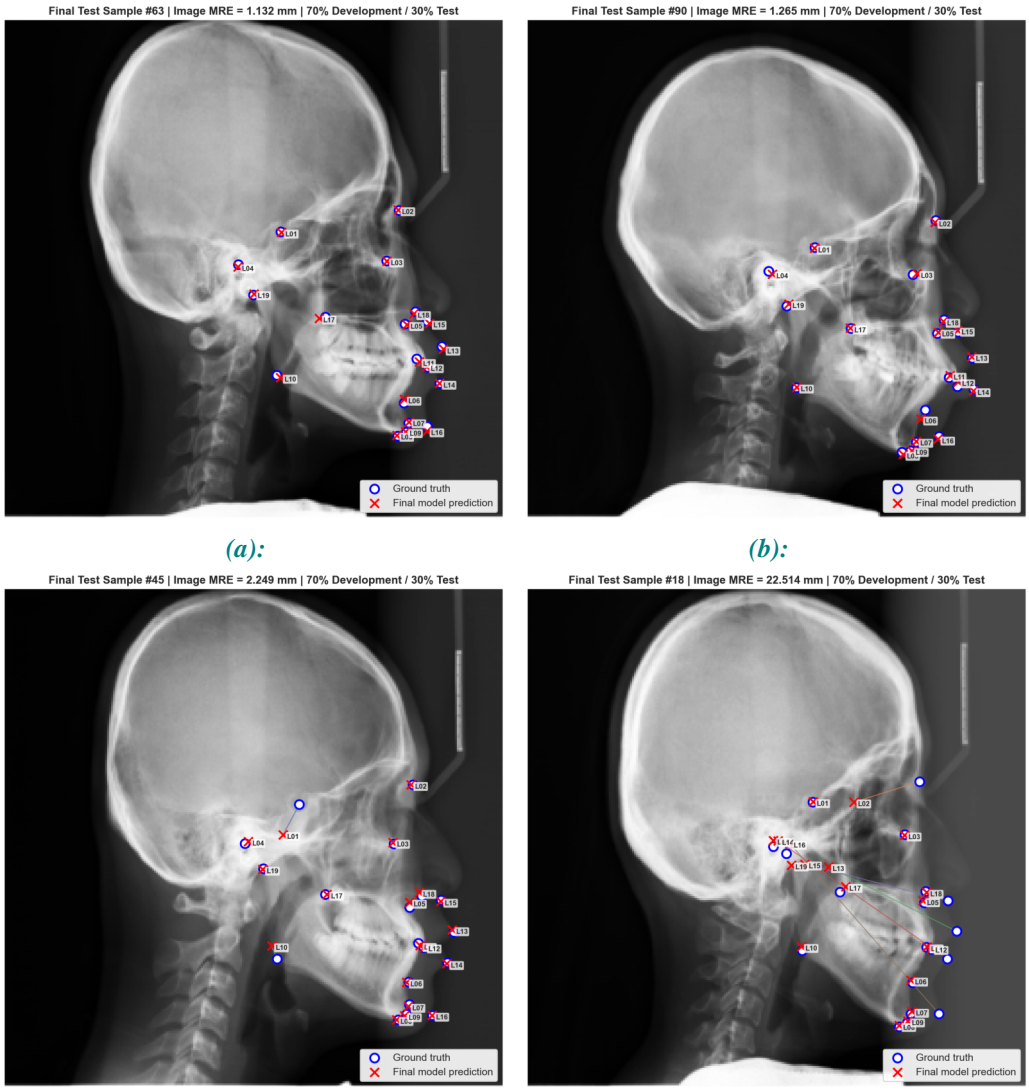


Figure 4.10: *Qualitative localization overlays for four representative test images from the primary Holdout 70/30 protocol:* ***(a)*** *ID 63,* ***(b)*** *ID 90,* ***(c)*** *ID 45, and* ***(d)*** *ID 18.*

an isolated but substantial failure. Therefore, the result for ID 18 should not be interpreted as evidence that all landmarks were poorly localized. Instead, it identifies a difficult radiograph in which a limited number of landmark predictions deviated considerably from their reference positions.

The four representative cases show that the localization performance varied across individual radiographs. Images ID 63 and ID 90 exhibited low image-level MRE values and relatively limited maximum errors. Image ID 45 showed a moderate increase in the mean error associated with a larger individual deviation, whereas ID 18 exhibited an extreme localization failure that strongly affected its image-level statistics.

This variability shows that aggregate metrics need to be supported by per-image and per-landmark analyzes.

Table 4.6. Landmark-level localization results for test image ID 63.

Landmark	x_{GT}	y_{GT}	x_{Pred}	y_{Pred}	e_{px}	e_{mm}	Δx_{mm}	Δy_{mm}
1	141.5414	108.6381	141.8792	108.8384	0.3927	0.3932	0.3445	0.1895
2	202.0795	97.5364	201.5820	97.3343	0.5370	0.5423	-0.5075	-0.1911
3	195.5845	123.4404	195.7887	123.6345	0.2817	0.2776	0.2083	0.1836
4	119.7281	125.2907	119.5724	126.6065	1.3250	1.2546	-0.1588	1.2446
5	205.1431	155.9525	206.0254	156.2496	0.9310	0.9428	0.9000	0.2810
6	204.6529	196.3944	204.8102	194.5091	1.8918	1.7903	0.1605	-1.7831
7	206.9813	206.7031	207.2205	206.0976	0.6511	0.6225	0.2440	-0.5727
8	201.3442	213.3113	201.0145	212.9938	0.4577	0.4508	-0.3362	-0.3003
9	205.7559	211.1967	205.4706	211.1116	0.2977	0.3019	-0.2910	-0.0805
10	139.4581	182.1208	141.2742	183.2526	2.1399	**2.1395**	1.8524	1.0705
11	211.1479	173.7945	212.4008	175.1832	1.8703	1.8325	1.2780	1.3134
12	216.6625	178.0237	216.2664	177.8177	0.4466	0.4486	-0.4041	-0.1949
13	223.8928	167.3185	224.7425	168.8296	1.7336	1.6715	0.8668	1.4292
14	222.9124	186.2179	222.7042	186.3183	0.2311	0.2326	-0.2123	0.0950
15	216.5400	154.4987	217.6494	155.5840	1.5520	1.5279	1.1316	1.0265
16	216.5400	208.4213	216.4673	210.7322	2.3120	**2.1869**	-0.0742	2.1856
17	164.2125	152.2519	161.2505	152.9482	3.0427	**3.0922**	-3.0213	0.6586
18	210.4126	149.6087	209.6640	150.6404	1.2748	1.2391	-0.7636	0.9758
19	127.3260	140.8859	127.6634	140.4059	0.5867	0.5697	0.3442	-0.4540

Image-level performance summary

MRE [mm]	1.1325	**Median [mm]**	0.9428	**STD [mm]**	0.8103
RMSE [mm]	1.3846	**Min [mm]**	0.2326	**Max [mm]**	3.0922
SDR@2 [%]	84.21	**SDR@2.5 [%]**	94.74	**SDR@3 [%]**	94.74
SDR@4 [%]	100.00				

Note: Ground-truth and predicted coordinates are reported in the resized image space (pixels). The pixel-domain localization error is computed as $e_{px} = \sqrt{(\hat{x} - x)^2 + (\hat{y} - y)^2}$. The millimeter-domain error is then obtained by anisotropic conversion: $e_{mm} = \sqrt{(M_x(\hat{x} - x))^2 + (M_y(\hat{y} - y))^2}$, where $M_x = 1.0200$ mm/pixel and $M_y = 0.9458$ mm/pixel. Landmark errors greater than 2 mm are shown in bold.

Image level MRE cannot tell us whether the error is spread over many landmarks or few hard predictions; numerical tables plus qualitative overlays tell us more about how the model behaves on individual radiographs.

The causes of these extreme deviations cannot be inferred from the results. They may be due to image quality, anatomical overlap, landmark ambiguity, atypical craniofacial structure or other reasons. Identifying them would require a case-by-case analysis which is not addressed in this study.

4.4 Additional Validation Results

This section reports the results of the two complementary validation protocols. The preprocessing, architecture, loss function, coordinate decoder, and evaluation metrics are those defined in Chapter 3; only the fold-specific configuration and the aggregated results are presented here.

Table 4.7. Landmark-level localization results for test image ID 90.

Landmark	x_{GT}	y_{GT}	x_{Pred}	y_{Pred}	e_{px}	e_{mm}	Δx_{mm}	Δy_{mm}
1	147.0560	116.5679	146.8719	117.0428	0.5093	0.4868	-0.1878	0.4492
2	208.9421	102.6908	208.7027	103.7926	1.1276	1.0704	-0.2442	1.0422
3	197.4227	130.4450	199.4545	130.0757	2.0651	**2.1017**	2.0725	-0.3493
4	123.5270	128.7269	125.3170	130.3805	2.4369	**2.4041**	1.8258	1.5640
5	209.7999	160.7104	210.3506	160.1552	0.7820	0.7690	0.5617	-0.5251
6	203.7951	199.9628	201.4065	204.6724	5.2807	**5.0771**	-2.4364	4.4543
7	199.0158	216.4832	199.0641	216.0902	0.3960	0.3750	0.0493	-0.3717
8	191.7855	221.5054	192.3253	222.7700	1.3750	1.3167	0.5505	1.1961
9	196.1972	220.8446	196.7755	220.8103	0.5793	0.5908	0.5899	-0.0324
10	137.7425	188.4646	137.8721	188.2834	0.2228	0.2165	0.1323	-0.1714
11	215.9272	183.3103	216.3537	182.2107	1.1794	1.1273	0.4350	-1.0400
12	220.2164	187.9360	220.5187	185.6111	2.3445	**2.2204**	0.3083	-2.1989
13	227.6917	173.3980	227.5900	172.7420	0.6639	0.6291	-0.1037	-0.6205
14	228.5495	189.3898	228.4073	190.1632	0.7864	0.7458	-0.1451	0.7315
15	220.4615	160.0496	220.2931	159.1208	0.9439	0.8951	-0.1717	-0.8784
16	210.6577	213.5756	210.2470	214.9577	1.4419	1.3727	-0.4190	1.3072
17	165.5606	158.0671	165.3321	158.0169	0.2339	0.2378	-0.2330	-0.0475
18	213.3538	153.7057	212.9256	154.5704	0.9649	0.9271	-0.4367	0.8178
19	132.7180	146.7011	133.6695	145.5220	1.5151	1.4784	0.9705	-1.1152

Image-level performance summary					
MRE [mm]	1.2654	**Median [mm]**	0.9271	**STD [mm]**	1.1894
RMSE [mm]	1.7338	**Min [mm]**	0.2165	**Max [mm]**	5.0771
SDR@2 [%]	78.95	**SDR@2.5 [%]**	94.74	**SDR@3 [%]**	94.74
SDR@4 [%]	94.74				

Note: Ground-truth and predicted coordinates are reported in the resized image space (pixels). The pixel-domain localization error is computed as $e_{px} = \sqrt{(\hat{x} - x)^2 + (\hat{y} - y)^2}$. The millimeter-domain error is then obtained by anisotropic conversion: $e_{mm} = \sqrt{(M_x(\hat{x} - x))^2 + (M_y(\hat{y} - y))^2}$, where $M_x = 1.0200$ mm/pixel and $M_y = 0.9458$ mm/pixel. Landmark errors greater than 2 mm are shown in bold.

4.4.1 K-Fold 70/30 Ensemble

This paragraph-level presentation reports the results of the complementary K-Fold 70/30 ensemble evaluation. The preprocessing, architecture, loss function, coordinate decoder, and evaluation metrics are those defined in Chapter 3; only the fold-specific configuration and the external results are presented here.

Fold-wise training configuration. During the K-Fold 70/30 training stage, the 72-image development set was divided into ten folds of seven or eight radiographs, each used once as an internal validation subset while the remaining folds supported optimization. The 30-image final test cohort remained external to the fold-wise model-selection process. Table 4.10 reports the configuration of each training fold.

The number of training images varied between 64 and 65, whereas the number of validation images varied between 7 and 8, for a total of 72 development radiographs in each fold. For each fold, the best epoch was selected

Table 4.8. Landmark-level localization results for test image ID 45.

Landmark	x_{GT}	y_{GT}	x_{Pred}	y_{Pred}	e_{px}	e_{mm}	Δx_{mm}	Δy_{mm}
1	150.9775	110.4884	142.6621	125.9821	17.5841	**16.9317**	-8.4818	14.6540
2	209.1872	100.3118	208.1486	100.2850	1.0389	1.0597	-1.0594	-0.0254
3	199.5060	130.7093	198.8955	129.9668	0.9613	0.9387	-0.6227	-0.7023
4	123.1594	130.4450	124.8197	129.4735	1.9237	1.9268	1.6936	-0.9189
5	207.4715	163.2215	207.5500	160.6618	2.5609	**2.4223**	0.0801	-2.4209
6	206.9813	201.4166	205.8671	201.6174	1.1321	1.1523	-1.1365	0.1899
7	207.3490	213.0470	206.8296	214.0549	1.1338	1.0906	-0.5297	0.9533
8	201.5893	220.8446	200.7197	220.8271	0.8698	0.8872	-0.8870	-0.0165
9	205.8784	217.9370	204.9802	218.8527	1.2827	1.2608	-0.9162	0.8661
10	139.5807	189.6541	136.8015	183.0398	7.1745	**6.8681**	-2.8348	-6.2558
11	211.8832	181.5921	212.8485	182.8544	1.5890	1.5475	0.9847	1.1938
12	214.7018	183.8389	215.0942	183.6255	0.4467	0.4483	0.4003	-0.2018
13	230.1427	175.9091	229.2647	174.7784	1.4316	1.3949	-0.8955	-1.0695
14	226.8339	191.6366	226.5721	192.1346	0.5626	0.5415	-0.2670	0.4710
15	223.5251	159.5209	223.5080	160.1701	0.6494	0.6142	-0.0175	0.6140
16	219.1134	218.7300	219.0032	218.6159	0.1587	0.1559	-0.1125	-0.1079
17	164.2125	156.6133	165.0937	156.6017	0.8812	0.8989	0.8988	-0.0110
18	213.5988	157.4063	212.4120	155.6310	2.1354	**2.0700**	-1.2106	-1.6791
19	132.5955	143.3970	132.1987	143.7416	0.5256	0.5197	-0.4048	0.3260

Image-level performance summary

MRE [mm]	2.2489	**Median [mm]**	1.0906	**STD [mm]**	3.7049
RMSE [mm]	4.3335	**Min [mm]**	0.1559	**Max [mm]**	16.9317
SDR@2 [%]	78.95	**SDR@2.5 [%]**	84.21	**SDR@3 [%]**	89.47
SDR@4 [%]	89.47				

Note: Ground-truth and predicted coordinates are reported in the resized image space (pixels). The pixel-domain localization error is computed as $e_{px} = \sqrt{(\hat{x} - x)^2 + (\hat{y} - y)^2}$. The millimeter-domain error is then obtained by anisotropic conversion: $e_{mm} = \sqrt{(M_x(\hat{x} - x))^2 + (M_y(\hat{y} - y))^2}$, where $M_x = 1.0200$ mm/pixel and $M_y = 0.9458$ mm/pixel. Landmark errors greater than 2 mm are shown in bold.

as the epoch with the minimum validation heatmap loss over the 250 training epoadd, and the corresponding checkpoint was retained for the ensemble. The selected epoadd ranged from 230 to 250, indicating that the fold-specific development subsets reached their best internal validation performance late in the optimization schedule. The lowest validation loss was obtained in Fold 6 at epoch 250, with a value of approximately 9.9649×10^{-5}, whereas the highest was observed in Fold 8 at approximately 1.5560×10^{-4}. This fold-to-fold variability is expected for small validation subsets and describes the internal training stage only; it is not an estimate of held-out performance.

External ensemble performance on the final test cohort. The ten fold-specific models were combined at the heatmap level, and Soft-Argmax was applied once to the averaged heatmaps, as described in Chapter 3. The ensemble was evaluated once on the 30-image final test cohort, which is identical to the final test cohort of the primary Holdout 70/30 protocol. This shared evaluation design enables the paired statistical comparisons reported in Section 4.5.2.

Table 4.9. Landmark-level localization results for test image ID 18.

Landmark	x_{GT}	y_{GT}	x_{Pred}	y_{Pred}	e_{px}	e_{mm}	Δx_{mm}	Δy_{mm}
1	146.0756	109.2989	146.1875	109.0938	0.2336	0.2251	0.1141	-0.1940
2	200.8540	98.7259	167.2689	109.2577	35.1977	**35.6763**	-34.2575	9.9610
3	193.2561	125.8193	193.0617	126.1986	0.4262	0.4098	-0.1983	0.3587
4	125.8554	132.0310	125.8688	128.9441	3.0869	**2.9196**	0.0136	-2.9196
5	203.1824	160.8425	202.5449	159.9087	1.1307	1.0968	-0.6503	-0.8832
6	197.1776	201.8131	196.4082	200.6072	1.4304	1.3845	-0.7848	-1.1405
7	196.3198	217.8049	196.9056	216.5854	1.3529	1.2990	0.5976	-1.1534
8	190.3150	224.5452	190.3075	223.8760	0.6692	0.6329	-0.0076	-0.6329
9	194.4816	222.4306	194.8191	221.5471	0.9458	0.9038	0.3443	-0.8356
10	140.5610	185.4249	140.2712	183.3755	2.0698	1.9607	-0.2956	-1.9383
11	204.4079	183.7068	204.5987	183.5713	0.2341	0.2331	0.1947	-0.1281
12	206.9813	184.6319	206.8923	184.3689	0.2777	0.2648	-0.0908	-0.2488
13	219.8487	175.5126	154.1715	142.4326	73.5376	**73.9379**	-66.9920	-31.2872
14	215.0694	189.6541	128.2247	129.0751	105.8858	**105.4978**	-88.5833	-57.2956
15	215.4371	159.9174	141.9867	141.0695	75.8301	**77.0124**	-74.9208	-17.8264
16	210.7803	217.8049	134.2852	130.4559	116.1091	**113.6368**	-78.0264	-82.6147
17	160.0460	155.2917	162.8350	152.5564	3.9065	**3.8453**	2.8449	-2.5871
18	203.9177	155.0274	204.3450	155.6071	0.7202	0.7005	0.4359	0.5483
19	132.5955	135.7315	135.1986	141.5797	6.4014	**6.1355**	2.6552	5.5312

Image-level performance summary					
MRE [mm]	22.5143	**Median [mm]**	1.3845	**STD [mm]**	35.8365
RMSE [mm]	41.2846	**Min [mm]**	0.2251	**Max [mm]**	113.6368
SDR@2 [%]	57.89	**SDR@2.5 [%]**	57.89	**SDR@3 [%]**	63.16
SDR@4 [%]	68.42				

Note: Ground-truth and predicted coordinates are reported in the resized image space (pixels). The pixel-domain localization error is computed as $e_{px} = \sqrt{(\hat{x} - x)^2 + (\hat{y} - y)^2}$. The millimeter-domain error is then obtained by anisotropic conversion: $e_{mm} = \sqrt{(M_x(\hat{x} - x))^2 + (M_y(\hat{y} - y))^2}$, where $M_x = 1.0200$ mm/pixel and $M_y = 0.9458$ mm/pixel. Landmark errors greater than 2 mm are shown in bold.

Table 4.10. Fold-wise training and validation configuration for the K-Fold 70/30 ensemble.

Fold	Training images	Validation images	Best epoch	Best validation loss
Fold 1	64	8	239	0.000114886
Fold 2	64	8	241	0.000118091
Fold 3	65	7	236	0.000127175
Fold 4	65	7	238	0.000119377
Fold 5	65	7	234	0.000119497
Fold 6	65	7	250	0.000099649
Fold 7	65	7	241	0.000100241
Fold 8	65	7	230	0.000155963
Fold 9	65	7	231	0.000111128
Fold 10	65	7	249	0.000133176

Table 4.11 summarizes the global localization performance of the ensemble. The mean radial error was 2.2423 mm, with an image-level standard deviation of 4.0365 mm and a mean test heatmap loss of 0.00008370. The successful detection rates were 81.23% at 2.0 mm, 87.89% at 2.5 mm, 91.58% at 3.0 mm, and 95.61% at

4.0 mm. These values correspond to $30 \times 19 = 570$ landmark-level errors evaluated on the same radiographs as the primary protocol.

Table 4.11. *Global performance of the K-Fold 70/30 ensemble on the 30-image final test set.*

Metric	Value
Number of test images	30
Number of landmark predictions	570
Mean test heatmap loss	0.00008370
Mean radial error (MRE)	2.242309 mm
Standard deviation of image-level MRE	4.036465 mm
Median image-level MRE	1.299727 mm
Minimum image-level MRE	1.025950 mm
Maximum image-level MRE	23.264553 mm
SDR@2 mm	81.228070%
SDR@2.5 mm	87.894737%
SDR@3 mm	91.578947%
SDR@4 mm	95.614035%

The image-level median MRE of 1.2997 mm was considerably lower than the image-level mean of 2.2423 mm, and the maximum image-level MRE reached 23.2646 mm. As under the primary protocol, this indicates a strongly right-skewed image-level distribution in which a limited number of difficult radiographs increase the average localization error. The SDR values increased consistently with the tolerance threshold, from 81.23% at 2.0 mm to 95.61% at 4.0 mm, indicating that most landmark predictions remained within the predefined error tolerances.

Landmark-level results. Table 4.12 reports the per-landmark mean radial errors, standard deviations, and SDR values obtained by the K-Fold 70/30 ensemble on the 570 landmark-level predictions.

Table 4.12. *Landmark-level results for the K-Fold 70/30 ensemble on the 30-image final test set.*

Landmark	MRE [mm]	SD [mm]	SDR@2	SDR@2.5	SDR@3	SDR@4
L01	1.668607	3.109276	83.33%	83.33%	93.33%	93.33%
L02	1.636400	3.282115	86.67%	96.67%	96.67%	96.67%
L03	1.662444	1.062637	66.67%	83.33%	86.67%	93.33%
L04	1.983932	1.341562	60.00%	70.00%	86.67%	93.33%
L05	1.020967	0.550957	96.67%	100.00%	100.00%	100.00%
L06	1.441853	0.914634	76.67%	86.67%	93.33%	96.67%
L07	1.294984	0.846745	86.67%	93.33%	93.33%	96.67%
L08	0.636073	0.302686	100.00%	100.00%	100.00%	100.00%
L09	0.819047	0.373000	96.67%	100.00%	100.00%	100.00%
L10	4.230256	7.106365	40.00%	53.33%	66.67%	73.33%
L11	1.111337	0.552734	90.00%	100.00%	100.00%	100.00%
L12	0.924831	0.996531	96.67%	96.67%	96.67%	96.67%
L13	4.226539	18.115280	90.00%	96.67%	96.67%	96.67%
L14	4.129398	18.635954	96.67%	96.67%	96.67%	96.67%
L15	3.749662	15.883375	96.67%	96.67%	96.67%	96.67%
L16	7.052161	22.509754	60.00%	83.33%	83.33%	93.33%
L17	2.068807	1.144458	60.00%	63.33%	76.67%	96.67%
L18	1.550011	0.909817	73.33%	80.00%	86.67%	100.00%
L19	1.396563	1.231920	86.67%	90.00%	90.00%	96.67%

The per-landmark MRE values ranged from 0.64 mm for L08 to 7.05 mm for L16. Three landmarks were localized with an MRE below 1 mm (L08, L09, and L12), eight landmarks achieved an MRE below 1.5 mm, and thirteen landmarks achieved an MRE below 2 mm. The landmarks with the largest average errors were L16, L10,

L13, L14, and L15, the same group identified under the primary protocol; the distinction between occasional and recurrent error behaviors within this group is analyzed in Section 4.5.3. L10 exhibits the lowest SDR@2 of all landmarks (40.00%), confirming its status as the most consistently difficult structure across evaluation designs.

Overall, the K-Fold 70/30 ensemble reproduced the localization behavior observed under the primary protocol on the same evaluation cohort, with a slightly higher mean radial error (2.2423 mm versus 2.1663 mm) and closely matched SDR values. Because both evaluations share the same 30 test radiographs, the formal assessment of this difference is deferred to the paired statistical analysis reported in Section 4.5.2.

4.4.2 Nested Leave-One-Out Cross-Validation

This subsection reports the results of the complementary Nested Leave-One-Out Cross-Validation protocol, which provides a robustness assessment across the complete dataset. The preprocessing, architecture, loss function, coordinate decoder, and evaluation metrics are those defined in Chapter 3; only the outer-iteration organization and the aggregated out-of-sample results are presented here.

Validity of the Nested LOOCV design. The Nested LOOCV design maximizes the use of the available annotated data. In the implemented protocol, every radiograph contributes to model development in all outer iterations except one and contributes exactly once to the final held-out evaluation, yielding one out-of-sample prediction per radiograph and $102 \times 19 = 1{,}938$ landmark-level predictions.

Strict out-of-sample evaluation also depends on the checkpoint-selection procedure. If the excluded radiograph were used to select the epoch with the lowest loss, the excluded image would influence model selection, and the resulting prediction would not be fully independent of the evaluation image. In the implemented protocol, the excluded radiograph is not used for training, inner validation, learning-rate scheduling, early stopping, checkpoint selection, or best-epoch selection; the training duration is selected exclusively by the inner validation subset, and the model is refitted from scratch on the remaining 101 radiographs before the excluded radiograph is evaluated once. The protocol therefore satisfies the strict Nested LOOCV design defined in Chapter 3, and each of the 102 stored predictions is a strictly out-of-sample prediction.

Global out-of-sample performance. Table 4.13 summarizes the localization performance obtained under the Nested LOOCV protocol across the 102 held-out radiographs.

Table 4.13. Overall performance of the Nested LOOCV protocol across the 102 held-out images.

Statistic	MRE [mm]	SDR@2 [%]	SDR@2.5 [%]	SDR@3 [%]	SDR@4 [%]
Mean	2.254088	79.772962	87.770898	90.970072	95.717234
Median	1.383033	–	–	–	–
STD	5.248308	–	–	–	–
Min	0.823389	–	–	–	–
Max	49.924923	–	–	–	–

The mean image-level MRE was 2.2541 mm, with a 95% confidence interval of $[1.2232, 3.2850]$ mm, whereas the image-level median MRE was considerably lower at 1.3830 mm. The median radial error across the 1,938 landmark-level predictions was 1.1148 mm. This difference between mean and median, together with the relatively large image-level standard deviation of 5.2483 mm and the maximum image-level MRE of 49.9249 mm, observed for image ID 1, indicates a strongly right-skewed distribution in which a limited number of difficult radiographs substantially increase the average localization error. The median therefore represents the typical prediction error better than the mean alone. The mean outer-test heatmap loss over the 102 evaluations was 0.00012035.

The SDR values show that localization accuracy improved consistently as the tolerance threshold increased. Overall, 79.77% of the 1,938 landmark predictions were localized within 2 mm, increasing to 87.77%, 90.97%, and 95.72% for the 2.5, 3.0, and 4.0 mm thresholds, respectively. These values are closely aligned with those obtained under the primary Holdout 70/30 protocol, although the two evaluations differ in their evaluated populations: the Holdout protocol reports 30 predictions generated by a single refitted model, whereas the Nested LOOCV protocol reports 102 predictions generated by 102 independently refitted models, each trained without the radiograph it evaluates. The Nested LOOCV results are therefore interpreted as a robustness assessment of the same pipeline rather than as a competing performance estimate.

Landmark-level results. Table 4.14 reports the per-landmark mean radial errors, standard deviations, and SDR values obtained across the 1,938 out-of-sample landmark predictions, with $N = 102$ evaluations per landmark.

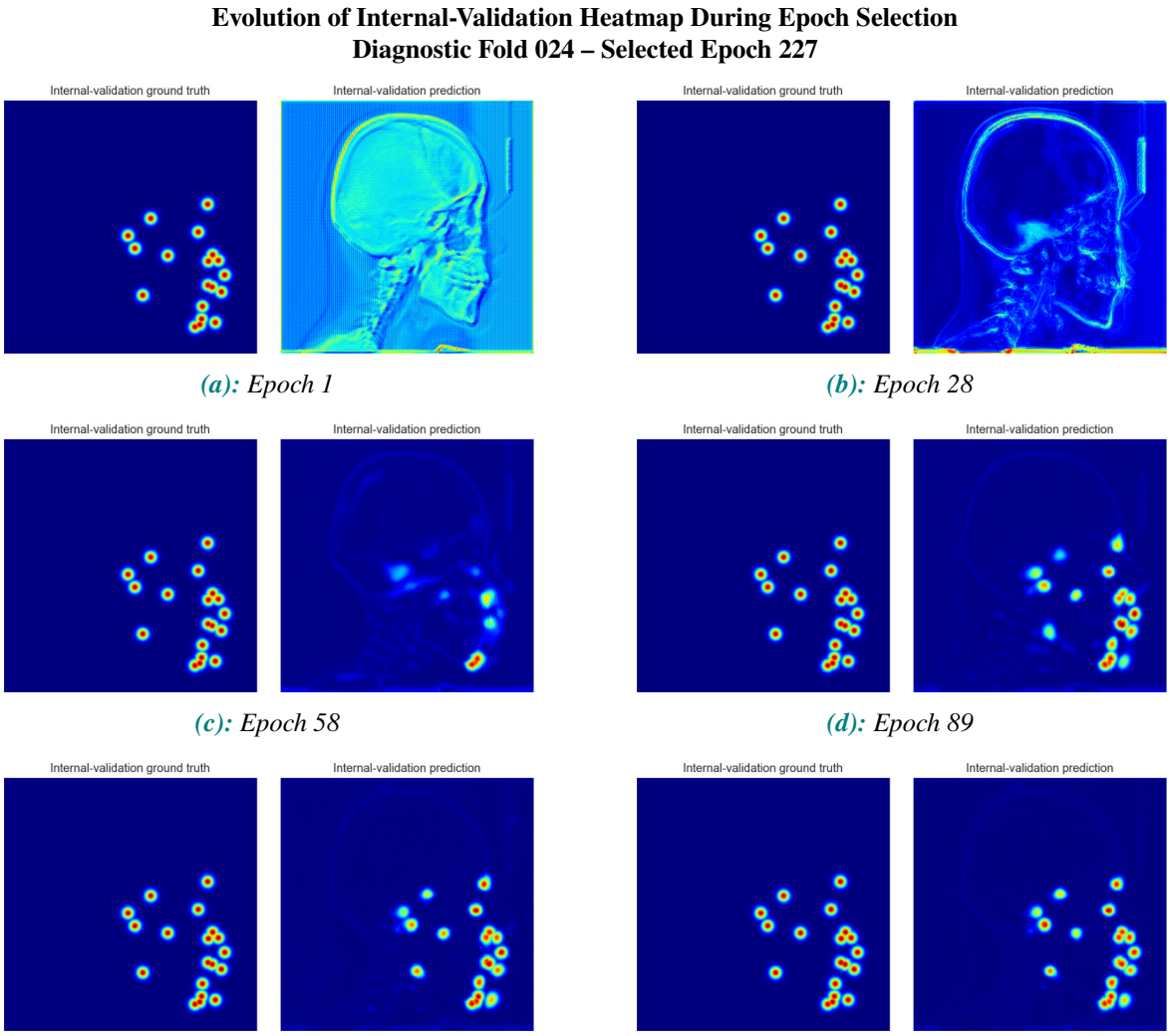
Table 4.14. *Landmark-level results for the Nested LOOCV protocol across the 102 held-out images.*

Landmark	MRE [mm]	SD [mm]	SDR@2	SDR@2.5	SDR@3	SDR@4
L01	1.119031	0.937339	95.10%	97.06%	97.06%	97.06%
L02	2.097926	7.416127	83.33%	91.18%	93.14%	96.08%
L03	2.881802	8.434034	69.61%	85.29%	87.25%	92.16%
L04	2.010455	1.134757	58.82%	78.43%	84.31%	92.16%
L05	2.155363	8.451868	82.35%	88.24%	95.10%	98.04%
L06	1.544736	1.077805	73.53%	83.33%	91.18%	98.04%
L07	1.313638	0.923708	83.33%	89.22%	93.14%	98.04%
L08	0.765531	0.519319	96.08%	99.02%	99.02%	100.00%
L09	0.930941	0.523269	97.06%	100.00%	100.00%	100.00%
L10	4.403468	8.941880	43.14%	54.90%	62.75%	78.43%
L11	2.515393	10.340548	77.45%	89.22%	93.14%	97.06%
L12	2.097204	10.805719	89.22%	93.14%	95.10%	98.04%
L13	3.081665	14.937722	87.25%	96.08%	97.06%	98.04%
L14	3.576738	15.867499	89.22%	94.12%	96.08%	96.08%
L15	2.681294	12.760507	96.08%	98.04%	98.04%	98.04%
L16	4.065378	16.016065	69.61%	79.41%	84.31%	95.10%
L17	2.009439	1.905194	65.69%	75.49%	76.47%	91.18%
L18	2.316018	8.228139	73.53%	84.31%	87.25%	96.08%
L19	1.261664	0.831803	85.29%	91.18%	98.04%	99.02%

The per-landmark MRE values ranged from 0.77 mm for L08 to 4.40 mm for L10. Two landmarks were localized with an MRE below 1 mm (L08 and L09), five landmarks achieved an MRE below 1.5 mm (L01, L07, L08, L09, and L19), and six landmarks achieved an MRE below 2 mm. The higher per-landmark dispersion under this protocol, compared with the primary evaluation, is consistent with the broader out-of-sample coverage of this design; the evaluated populations differ between protocols, so this difference is descriptive rather than causal. The landmarks with the largest average errors were L10, L16, L14, L13, and, to a lesser extent, L03; the distinction between occasional and recurrent error behaviors within this group, together with the influence of

extreme radiographs on the aggregate statistics, is analyzed in Section 4.5.3.

Training dynamics and fold-specific training durations. Each of the 102 outer folds was trained for up to 250 epoadd. Figure 4.11 presents the evolution of the training loss and the inner validation heatmap loss across the 102 completed folds.



(a): Epoch 1 *(b): Epoch 28*

(c): Epoch 58 *(d): Epoch 89*

(e): Epoch 136 *(f): Epoch 227*

Figure 4.11: *Predicted and reference inner-validation heatmaps at six training epoadd of outer fold 024 of the Nested LOOCV protocol, whose training duration was selected at epoch 227. Each panel shows the predicted heatmap (left) and the corresponding Gaussian reference heatmap (right).*

The training curves describe optimization behavior and fold-to-fold variability. A decreasing training loss indicates that the network is fitting the Gaussian heatmap targets, but this decrease alone does not guaranty improved landmark localization on unseen radiographs. The inner validation loss provides complementary

information concerning the transfer of the learned representation to data not used for parameter updates, and it constitutes the sole criterion for training-duration selection within each outer fold.

For each outer fold, the training duration was determined from the inner validation heatmap loss. The selected training duration varied across the 102 outer folds. The mean selected epoch was 232.8 epoadd, with a standard deviation of 15.9, a median of 238 epoadd, an interquartile range of 224-244 epoadd, and a range of 164-250 epoadd. These values were obtained from the completed outer-fold training records and describe internal training-duration selection only.

Because each outer fold contains only one excluded radiograph and a small inner validation subset, the validation loss may vary substantially from one fold to another. Such variability may be caused by differences in image quality, anatomical complexity, landmark visibility, and annotation ambiguity. The training curves should therefore be interpreted as evidence of optimization behavior rather than as an additional held-out evaluation, because the held-out performance is determined exclusively from the predictions obtained for the excluded radiographs.

The relatively concentrated distribution of the selected epoadd, with an interquartile range of only 224-244 epoadd, indicates that the different development subsets reached their best inner validation performance at similar stages of optimization and supports the use of a fixed training schedule in future experiments.

4.5 Inter-Protocol Comparison

This section compares the three evaluation protocols at two levels. The global comparison aggregates all landmark-level errors of each protocol, whereas the paired comparison restricts the analysis to images evaluated under both protocols of a given pair, treating their image-level MRE values as matched observations.

4.5.1 Global Performance Comparison

Table 4.15 summarizes the global performance of the three protocols. The Holdout 70/30 protocol provides the primary estimate on its 30-image final test set. The K-Fold 70/30 ensemble is evaluated on the same 30 radiographs. The Nested LOOCV protocol evaluates all 102 radiographs, each once, through fold-specific refitting.

Table 4.15. *Global comparison of the evaluation protocols. Medians and standard deviations are computed at the image level.*

Protocol	*N*	**MRE [mm]**	**SD [mm]**	**Median [mm]**	**SDR@2 [%]**	**SDR@4 [%]**
Holdout 70/30 (primary)	30	2.1663	3.8597	1.3600	80.35	95.79
K-Fold 70/30 ensemble	30	2.2423	4.0365	1.2997	81.23	95.61
Nested LOOCV	102	2.2541	5.2483	1.3830	79.77	95.72

The three protocols produced closely aligned global results, with mean radial errors between 2.1663 mm and 2.2541 mm and SDR@4 values between 95.61% and 95.79%. The Holdout 70/30 protocol yielded the lowest

mean radial error, and the Nested LOOCV protocol the highest image-level standard deviation. This observation is consistent with the stricter out-of-sample design and the broader evaluated population of the Nested LOOCV protocol. Because the three protocols evaluate different populations, however, the difference in dispersion cannot be attributed to the validation design alone. Because the sample sizes and the evaluated populations differ, these global values must not be interpreted as a direct competition between protocols.

4.5.2 Paired Statistical Comparison

Because the Holdout 70/30 and K-Fold 70/30 protocols share the same 30-image test cohort, and because these radiographs are also included among the 102 images evaluated under Nested LOOCV, image-level MRE values can be compared as matched observations. Three comparisons were performed on the same 30 radiographs: Holdout versus K-Fold, Holdout versus Nested LOOCV, and K-Fold versus Nested LOOCV.

For each comparison, the paired difference was defined as the image-level MRE of the first-listed protocol minus that of the second-listed protocol. The paired Student's t-test and the Wilcoxon signed-rank test were used as complementary parametric and nonparametric procedures. Cohen's d_z quantified the standardized paired mean difference. Cliff's δ was computed as a descriptive dominance-based effect size from the two distributions of image-level MRE restricted to the 30 common radiographs. The implemented calculation of Cliff's δ compares all $30 \times 30 = 900$ cross-protocol pairs and does not preserve image-by-image matching. A positive value indicates that an MRE drawn from the first-listed protocol tends to exceed an MRE drawn from the second-listed protocol; the value $\delta = 0.1644$ obtained for Holdout versus K-Fold corresponds, in the absence of tied MRE values across protocols, to approximately 58% of the 900 cross-protocol comparisons exhibiting a larger error under the first-listed protocol. Because Cliff's δ is a rank-based dominance measure computed over all $30 \times 30 = 900$ cross-protocol pairs, whereas the paired mean difference is magnitude-weighted on the matched images, the two statistics can differ in sign when the error distributions are right-skewed and heavy-tailed. In the present comparison, Holdout versus K-Fold exhibits a positive dominance ($\delta = 0.1644$, the first-listed protocol producing the larger MRE in a larger proportion of cross-protocol comparisons) together with a negative paired mean difference (−0.0760 mm, the first-listed protocol being lower on average on the matched radiographs). This coexistence is a descriptive property of the two image-level MRE distributions; the present analyzes do not identify which radiographs are responsible for it. Its direction is nevertheless consistent with the strongly right-skewed error distributions reported in Section 4.5.3. A 95% confidence interval was calculated for each paired mean difference. The results are summarized in Table 4.16.

Table 4.16. *Statistical comparison of the evaluation protocols on the 30 common radiographs. Mean differences, confidence intervals, t-tests, Wilcoxon tests, and Cohen's d_z use matched image-level observations. Cliff's δ is a descriptive dominance measure computed on the same 30 image-level MRE values per protocol, over all cross-protocol pairs. Differences are defined as the first-listed protocol minus the second-listed protocol.*

Comparison	Mean diff. [mm]	95% CI	t(29)	p (t)	p (Wilcoxon)	d_z	δ
Holdout vs. K-Fold	−0.0760	[−0.2391, 0.0871]	−0.953	0.349	0.140	−0.174	0.1644
Holdout vs. Nested LOOCV	−0.1457	[−0.3761, 0.0846]	−1.294	0.206	0.839	−0.236	−0.0178
K-Fold vs. Nested LOOCV	−0.0698	[−0.2514, 0.1118]	−0.786	0.438	0.135	−0.144	−0.1711

None of the three paired comparisons reached statistical significance at the 5% level under either the paired t-test or the Wilcoxon signed-rank test. All 95% confidence intervals for the paired mean differences included zero, and the absolute values of Cohen's d_z did not exceed 0.24. The mean paired difference was −0.0760 mm for Holdout versus K-Fold, −0.1457 mm for Holdout versus Nested LOOCV, and −0.0698 mm for K-Fold versus Nested LOOCV. Thus, no statistically significant differences in image-level MRE were detected on the 30 common radiographs. These findings do not, however, establish statistical equivalence between the protocols.

Exact McNemar tests provided a complementary comparison of binary landmark-level success and failure at the 2.0 mm and 2.5 mm thresholds. For each threshold and protocol pair, the two-sided exact binomial test was applied to the discordant classifications among the 570 matched landmark predictions. No statistically significant difference was detected for Holdout versus K-Fold ($p = 0.4869$ at 2 mm; $p = 0.6636$ at 2.5 mm), Holdout versus Nested LOOCV ($p = 1.0000$ at 2 mm; $p = 0.3105$ at 2.5 mm), or K-Fold versus Nested LOOCV ($p = 0.5515$ at 2 mm; $p = 0.6177$ at 2.5 mm).

For Holdout versus Nested LOOCV at the 2 mm threshold, both protocols achieved 458 successful predictions out of 570 on the common radiographs. Of these predictions, 426 were successful under both protocols, 80 failed under both protocols, and the remaining 64 were discordant, with 32 successes exclusive to each protocol. The equal numbers of discordant outcomes yielded an exact McNemar p-value of 1.0000.

Because the 19 landmark predictions within each radiograph may be correlated, the landmark-level McNemar tests should be interpreted as complementary analyzes that do not explicitly account for within-image clustering. Together with the paired image-level analyzes, they provide no evidence of a statistically detectable difference at the examined thresholds, but they do not demonstrate identical predictions or statistical equivalence.

4.5.3 Landmark-Specific Errors and Extreme Cases

The per-landmark analyzes of the three protocols identified a consistent group of difficult structures exhibiting two distinct error behaviors. L10 and L16 exhibited the largest mean radial errors under all three validation designs (L10: 3.36 to 4.40 mm; L16: 4.07 to 7.05 mm), combining large mean errors with substantially lower SDR@2 values (40.00% to 69.61%), which indicates that their localization errors are both more frequent and more variable. In contrast, L13, L14, and L15 combined high SDR@2 values (87.25% to 96.67% across protocols) with very large standard deviations, indicating that these landmarks are localized accurately in most radiographs but occasionally produce extreme deviations. This contrast is particularly clear under the Nested LOOCV evaluation, where the per-landmark maximum errors exceeded 100 mm for L13 (116.23 mm), L14 (121.50 mm), and L16 (122.71 mm) while their SDR@2 values remained between 69.61% and 89.22%. L03 and L04 also exhibited elevated errors and reduced SDR@2 values under several protocols, with L03 reaching an MRE of 2.8818 mm and an SDR@2 of 69.61% under the Nested LOOCV evaluation, placing it in an intermediate category.

The landmark-specific statistics differ across validation designs. Under the primary Holdout 70/30 and K-Fold 70/30 evaluations, three landmarks achieved an MRE below 1 mm (L08, L09, and L12) and eleven to thirteen landmarks achieved an MRE below 2 mm, whereas under the Nested LOOCV evaluation only two landmarks (L08

and L09) remained below 1 mm and six landmarks below 2 mm. This shift is consistent with the broader evaluated population of the Nested LOOCV protocol, in which each landmark is assessed across all 102 radiographs through models trained without them; because the evaluated populations differ between protocols, this difference is descriptive rather than causal. Across all three protocols, however, L08 and L09 remained the most accurately localized structures, and L10, L16, L13, and L14 (with L15 close behind) remained the most challenging ones, indicating that these difficulty patterns recur across the three validation protocols and are therefore largely driven by specific anatomical structures, although the present design cannot exclude an influence of the validation protocol on the precise ranking.

The influence of extreme radiographs on the aggregate statistics was substantial. Under the primary Holdout 70/30 protocol, image ID 18 exhibited an image-level MRE of 22.5143 mm, and its removal reduced the mean image-level MRE from 2.1663 mm to 1.4647 mm, a decrease of 32%. The same pattern appeared under the K-Fold 70/30 ensemble, which reached its maximum image-level MRE on the same image ID 18 (23.2646 mm), with its mean decreasing from 2.2423 mm to 1.5174 mm without this radiograph. Under the Nested LOOCV protocol, image ID 1 reached an image-level MRE of 49.9249 mm, and its removal reduced the mean from 2.2541 mm to 1.7821 mm.

The distribution of the 1,938 landmark-level radial errors of the Nested LOOCV evaluation quantifies this concentration: the errors were strongly right-skewed (skewness = 10.71), with a median of 1.1148 mm, a first quartile of 0.6689 mm, a third quartile of 1.8031 mm, and a maximum of 122.715 mm. Nearly half of the landmark predictions (44.43%) were localized within 1 mm and 79.77% within 2 mm; beyond the conventional thresholds, the tail decreased rapidly, with 98.61% of predictions below 10 mm and only 8 of the 1,938 predictions (0.41%) exceeding 100 mm. At the image level, the same pattern was observed under the three protocols, with compact interquartile ranges ($[1.2603, 1.5466]$ mm for the Holdout, $[1.1834, 1.4636]$ mm for the K-Fold ensemble, $[1.2213, 1.5516]$ mm for the Nested LOOCV) around medians between 1.30 and 1.38 mm, while the means (2.1663 to 2.2541 mm) were pulled upward by a small number of extreme radiographs. These results explain why a relatively moderate global mean can coexist with a large standard deviation: the typical radiograph is localized with a sub-2 mm accuracy, whereas one to three difficult cases per protocol dominate the mean and the standard deviation. The outlier radiographs must therefore be described as observations of the evaluation set that concentrate severe localization failures, rather than as evidence of a general limitation of the pipeline. Their common characteristics and candidate explanations, including image quality, anatomical overlap, landmark ambiguity, and atypical craniofacial structure, are discussed as directions for future work in Chapter 5; a case-by-case causal analysis of the individual extreme radiographs was not performed in this study.

4.5.4 Comparison with the State of the Art

To position the proposed pipeline within the current literature, Table 4.17 summarizes representative approaches for automatic cephalometric landmark localization. The compared values follow the reporting convention introduced in section 2.8.3: each value is reported as published by the original study, and the dataset on which it was obtained is indicated in parentheses.

Table 4.17. Comparison with representative methods reported in the literature. Performance values are reported as published by the original studies; the dataset on which each value was obtained is indicated in parentheses. For studies reporting Test1 and Test2 separately, values are given as Test1 / Test2.

Method	Main approach	MRE [mm]	SDR@2 mm
Proposed pipeline	**Lightweight U-Net, Gaussian heatmap regression, Soft-Argmax, Holdout 70/30**	**2.1663**	**80.35% (PKU)**
Self-CephaloNet (2025) [Sumon et al., 2025]	HRNetV2 + Self-ONN (external validation)	1.87 ± 2.79	75.95% (PKU)
Self-CephaloNet (2025) [Sumon et al., 2025]	Two-stage heatmap regression	1.08 / 1.44	87.82 / 76.68% (ISBI Test1 / Test2)
Zeng et al. [2021]	Cascaded CNNs (external validation)	2.02 ± 1.89	64.81% (PKU)
Zeng et al. [2021]	Cascaded CNNs	1.34 / 1.64	81.37 / 70.58% (ISBI Test1 / Test2)
CephXNet (2023) [Neeraja y Anbarasi, 2023]	SE-CNN	1.41 ± 0.91	78.72% (ISBI Test2)
Rashmi et al. [2023]	HOG + LightGBM	N/A	84.00% (PKU)

Direct numerical comparisons must be interpreted with caution because the published methods use different datasets, annotation protocols, preprocessing strategies, coordinate systems, and validation procedures.

Under the primary Holdout 70/30 protocol, the proposed pipeline achieves an SDR@2 of 80.35% on the PKU dataset. This value exceeds the 64.81% reported by Zeng et al. [2021] on the same dataset and the 75.95% reported for Self-CephaloNet under its external-validation design [Sumon et al., 2025], whereas the 84.00% reported by Rashmi et al. [2023] remains higher than the present result. The MRE of 2.1663 mm obtained on the 30-image held-out test set is the highest among the PKU values listed in Table 4.17; this average is, however, strongly influenced by a small number of extreme cases. Excluding the single most affected radiograph (ID 18) reduces the MRE to 1.4647 mm. This sensitivity to individual outliers is analyzed in detail in Section 4.5.3.

The complementary Nested LOOCV evaluation, which covers all 102 radiographs, obtained an MRE of 2.2541 mm with an SDR@2 of 79.77%. This value is close to the primary estimate, which supports the consistency of the observed performance across the three evaluation designs applied to the PKU dataset; it does not exclude all possible partitioning effects, since all three evaluations draw on the same 102 radiographs. Because Nested LOOCV reuses the same 102 radiographs through internal resampling,it does not constitute an independent external validation, and the positioning relative to the literature should be read as indicative rather than as a controlled ranking.

Results obtained on other datasets, such as ISBI 2015, should be interpreted qualitatively. Differences in image characteristics, landmark definitions, annotation procedures, preprocessing operations, and validation strategies prevent strict numerical benchmarking.

The comparison nevertheless suggests that a relatively lightweight U-Net can achieve competitive localization performance when combined with Gaussian heatmap regression, differentiable Soft-Argmax coordinate extraction, anisotropic millimeter conversion, and a carefully controlled evaluation protocol. This statement is descriptive: without matched case-level observations from the published methods, the numerical differences in the table cannot establish statistical superiority or equivalence. The most appropriate comparisons involve the same dataset, the same landmark definitions, comparable physical calibration, and comparable validation procedures. Even when nominal MRE or SDR values are similar, differences in data partitioning, observer annotations, preprocessing, image resolution, and checkpoint selection can materially affect the reported performance. The literature comparison should therefore be interpreted as methodological positioning rather than as a formal statistical ranking of algorithms.

4.5.5 Methodological and Clinical Interpretation

Contribution of Gaussian heatmap regression. A central methodological choice was to use Gaussian heatmap regression rather than direct coordinate regression. It preserves the spatial relationship between input image and predicted landmark location, dense supervision during training (localization errors can generate smooth changes in optimization objective) and Gaussian heatmaps are particularly suitable when target landmark is small, partially occluded or surrounded by similar structures. Gaussian width acts as a spatial supervision parameter: very narrow target gets point-like supervision while wide target tolerates larger neighborhood around annotation. Localization results under all three validation protocols are consistent with effectiveness of this representation but no ablation experiment was performed so individual contribution of heatmap regression cannot be disentangled from other parts of pipeline.

Contribution of differentiable Soft-Argmax. The differentiable Soft-Argmax coordinate decoder computes the expected coordinates of the normalized heatmap instead of selecting the pixel with maximum activation. This operation produces continuous sub-pixel predictions and is differentiable with respect to the heatmap values. The parameter β controls the concentration of the normalized response: increasing β makes the operator behave more similarly to hard Argmax, whereas lower values distribute greater weight over neighboring locations. With $\beta = 25$, the present implementation favors a concentrated response while retaining continuous coordinate estimation. The use of Soft-Argmax is especially relevant after image resizing because the predicted coordinate may lie between two resized pixels. Although the present study does not include an ablation analysis comparing Soft-Argmax with alternative decoding methods, the obtained localization performance supports its use within the proposed framework.

Contribution of anisotropic millimeter conversion. Image resizing introduces different horizontal and vertical scaling factors because the original radiographs are not square. The adopted anisotropic conversion applies $M_x^{(j)}$ to the horizontal coordinate difference and $M_y^{(j)}$ to the vertical coordinate difference before computing Euclidean norm. This ordering is important: pixel-space Euclidean distance should not be first collapsed to a scalar and then multiplied by an averaged millimeter per pixel factor, because resize operation produces different physical scales along each of the two axes. Conversion step directly affects magnitude of MRE and SDR so it must be performed if comparisons with clinically meaningful thresholds are to be made.

Contribution of the validation hierarchy. The evaluation hierarchy is the core of our proposed methodology. The primary Holdout 70/30 protocol provides the reference held-out estimate on a fully isolated 30-image test cohort. The K-Fold 70/30 ensemble and the Nested LOOCV protocol provide complementary validation and robustness analyzes, and the paired statistical comparisons on the 30 common radiographs showed no statistically significant differences between the three protocols. The Nested LOOCV design is appropriate for a relatively small medical imaging dataset because each outer fold uses the maximum possible number of remaining radiographs for model development, ultimately yielding one out-of-sample prediction for every radiograph. Nevertheless, Nested LOOCV does not eliminate sampling uncertainty: the resulting performance remains conditional on the available cohort, and external validation on independent datasets is still necessary before generalization to other populations or acquisition conditions can be claimed.

Clinical interpretability and scope. The reported errors are expressed in physical millimeters through image-specific anisotropic conversion, which allows the SDR thresholds to be compared with the tolerance levels used in the cephalometric literature. The present results remain conditional on the studied PKU cohort and on its expert annotations, and they do not establish clinical superiority over published methods or generalization to external populations, acquisition devices, or imaging protocols. These considerations, together with the limitations of the present evaluation, are developed in Chapter 5.

4.6 Chapter Summary

This chapter presented the experimental evaluation of the proposed lightweight U-Net pipeline for cephalometric landmark localization on the PKU Cephalometric Dataset. The primary Holdout 70/30 protocol, evaluated once on 30 fully isolated test radiographs, yielded a mean radial error of 2.1663 mm with an SDR@2 of 80.35% and an SDR@4 of 95.79%. The complementary K-Fold 70/30 ensemble and Nested LOOCV protocol achieved closely aligned results (2.2423 mm and 2.2541 mm, respectively), and the paired statistical comparisons on the 30 common radiographs found no statistically significant differences between any pair of protocols, either by the paired t-tests or by the Wilcoxon signed-rank tests (all $p > 0.05$, all $|d_z| \leq 0.24$, and all 95% confidence intervals for the paired mean differences included zero).

The error analysis identified a small group of anatomically difficult landmarks, primarily L10 and L16, whose

errors are both frequent and variable, and a second group, L13, L14, and L15, whose mean errors are dominated by occasional extreme deviations. Similarly, one to three difficult radiographs per protocol, notably ID 18 and ID 1, concentrated the largest localization failures and dominated the mean-based statistics, while the median image-level MRE remained between 1.30 and 1.38 mm in all protocols. These findings, together with their clinical interpretability and the limitations arising from the studied cohort, are discussed in Chapter 5.

Chapter 5

General Conclusions

5.1 Introduction

In this thesis, we focused on the automatic localization of cephalometric landmarks in lateral radiographs with a heatmap-regression framework based on a simple two-level U-Net. We were motivated not only to localize landmarks accurately but also to obtain performance estimates that are scientifically defensible, even when the annotated dataset is small.

From there, we went on to implement a landmark-detection model and develop a complete, reproducible localization pipeline that includes deterministic image preprocessing, Gaussian heatmap regression, continuous coordinate decoding with Soft-Argmax, image-specific anisotropic physical conversion, and strict held-out validation. The final methodological setup puts the Holdout 70/30 protocol at the center of our experimental design as the main evaluation protocol but keeps K-Fold 70/30 ensemble and Nested LOOCV (Nested Leave-One-Out Cross-Validation) protocol as the complementary validation and robustness analyzes of the same pipeline.

The results demonstrate that the proposed approach provides accurate localization for the majority of cephalometric landmarks, while the exhaustive and complementary evaluations reveal the importance of a rigorous validation design for identifying difficult radiographs and landmark-specific failure patterns.

5.2 Summary of the Work Performed

The study started by reviewing the main methodological approaches for automatic cephalometric landmark localization (especially convolutional encoder-decoder architectures and heatmap-regression methods). This review led us to choose a lightweight U-Net architecture that can preserve spatial information but is still adapted to small datasets.

The resulting pipeline is composed of five major methodological components. First, lateral cephalometric radiographs (native width 2089 pixels, native height 1937 pixels, spatial resolution 0.125 mm/pixel) were resized to 256×256 pixels by bilinear interpolation, preserving the information necessary to recover their physical geometry during evaluation.

Second, each anatomical landmark was represented by a Gaussian heatmap, transforming direct coordinate prediction into a dense spatial regression problem. A two-level U-Net was used to estimate the 19 landmark heatmaps simultaneously.

Third, the predicted heatmaps were converted into continuous landmark coordinates using Soft-Argmax with $\beta = 25$ rather than relying exclusively on discrete maximum-response extraction. This decoding strategy preserved the spatial information contained in the heatmap distribution and enabled sub-pixel coordinate estimation.

Fourth, the physical localization error was calculated using independent horizontal and vertical scaling factors computed from the actual native dimensions of each radiograph. For a predicted coordinate $(\hat{x}, \hat{y})$ and a reference coordinate (x, y) in resized space, the radial error was calculated according to Eq. (5.1):

$$e_{\text{mm}} = \sqrt{\left[M_x^{(j)}(\hat{x} - x)\right]^2 + \left[M_y^{(j)}(\hat{y} - y)\right]^2}, \tag{5.1}$$

where $M_x^{(j)}$ and $M_y^{(j)}$ denote the image-specific horizontal and vertical millimeter-per-resized-pixel factors. This formulation avoided replacing the two spatial scales with a mean conversion factor and maintained consistency with the native radiographic geometry.

Fifth, the validation methodology was organized around a three-level hierarchy. The primary Holdout 70/30 protocol used 72 radiographs for model development and reserved 30 radiographs as a fully isolated held-out test set. The complementary K-Fold 70/30 protocol trained ten fold-specific models on the development data, combined their predicted heatmaps before Soft-Argmax coordinate extraction, and evaluated the resulting ensemble on the same 30-image held-out cohort. The Nested LOOCV protocol excluded each of the 102 radiographs once from model development: the training duration was selected exclusively by the inner validation subset, a newly initialized model was refitted from scratch on the remaining 101 radiographs, and the excluded radiograph was evaluated exactly once. The excluded radiograph was never used for training, inner validation, learning-rate scheduling, early stopping, checkpoint selection, or best-epoch selection.

The complete Nested LOOCV procedure produced $102 \times 19 = 1{,}938$ strictly out-of-sample landmark predictions, while the Holdout and K-Fold protocols each produced $30 \times 19 = 570$ held-out landmark predictions on their common evaluation cohort.

5.3 Main Findings

The primary Holdout 70/30 evaluation achieved a global Mean Radial Error of 2.1663 mm on the 30-image final test set, with a median image-level MRE of 1.3600 mm. The corresponding Successful Detection Rates were

80.35%, 87.37%, 91.40%, and 95.79% at localization thresholds of 2.0, 2.5, 3.0, and 4.0 mm, respectively. These values constitute the principal performance reference of this research.

These results demonstrate that the majority of landmark predictions were concentrated within relatively small physical errors: approximately 80% of the 570 held-out predictions were localized within 2 mm of the reference position, and more than 95% were localized within 4 mm. At the same time, the difference between the mean and the median errors revealed a right-skewed error distribution in which a limited number of difficult radiographs and landmark–radiograph combinations produced large deviations.

The complementary K-Fold 70/30 ensemble achieved an MRE of 2.2423 mm on the same 30 held-out images, with SDR values of 81.23%, 87.89%, 91.58%, and 95.61% at the same thresholds. The close agreement between the primary and complementary estimates on the identical evaluation cohort indicates that the observed performance is consistent across the three evaluation designs applied to the PKU dataset. This consistency does not exclude all possible partitioning effects, since all three evaluations draw on the same 102 radiographs.

The Nested LOOCV protocol achieved an MRE of 2.2541 mm across all 102 radiographs, with an image-level median MRE of 1.3830 mm and SDR values of 79.77%, 87.77%, 90.97%, and 95.72%. This exhaustive evaluation extends the held-out assessment from the 30-image cohort to the complete dataset and indicates that the performance observed on the fixed test set is consistent with the behavior observed across the full population, while its broader coverage exposes a larger image-level dispersion driven by a limited number of difficult radiographs.

The maximum inter-protocol MRE difference was 0.088 mm, and the SDR values differed by at most 1.5 percentage points at each threshold. The paired statistical analyzes on the 30 common radiographs, including the paired t-test, the Wilcoxon signed-rank test, Cohen's d_z, Cliff's δ, and the exact McNemar tests on landmark-level classifications (two-sided exact binomial tests on the discordant pairs), did not identify statistically significant differences between the protocols, and all effect sizes remained small. The absence of statistical significance does not demonstrate exact equivalence between the protocols, but it supports the interpretation that the three validation designs measure the same underlying localization behavior within the resolution of the matched sample.

The landmark-level analysis demonstrated that localization difficulty was not uniformly distributed across the 19 anatomical points. L08 and L09 were consistently among the most accurately localized landmarks across all three protocols, with L12, L11, and L05 among the most accurate under the two fixed-cohort protocols, whereas L10, L16, L14, and L13 (with L15 close behind) exhibited the largest average errors across the protocols. Under the primary protocol, 29 of the 30 test radiographs achieved an image-level MRE below 2.71 mm, while a single difficult radiograph contributed disproportionately to the image-level dispersion. These findings show that the global MRE must be interpreted together with the image-level distributions, the landmark-level statistics, and the qualitative examples rather than as a single summary value.

5.4 Main Contributions of the Study

The contributions of this research are methodological and pipeline-based rather than architectural. The study does not claim the introduction of a new neural-network architecture or a new algorithmic component. Instead, it demonstrates how established components can be reorganized and integrated into a rigorous framework for cephalometric landmark localization on a limited dataset.

5.4.1 Integrated Localization Pipeline

The first contribution is the development of an end-to-end localization pipeline that integrates image preprocessing, Gaussian heatmap regression, Soft-Argmax decoding, anisotropic physical conversion, and held-out performance evaluation within a coherent framework, following the complete path from the input radiograph to the final physical localization error. The methodological effort focused on controlling this complete path rather than increasing architectural complexity, which is particularly relevant in the context of a limited dataset.

5.4.2 Continuous Coordinate Extraction

Soft-Argmax keeps information about the spatial response distribution and produces continuous sub-pixel coordinates. Its contribution in this work is that it was always used as part of the heatmap-regression pipeline (as the main decoder) across all three validation protocols so no controlled ablation compared Soft-Argmax with other decoders, hence the study cannot claim independent causal effect only from this component.

5.4.3 Image-Specific Anisotropic Evaluation

The physical error calculation uses separate horizontal and vertical conversion factors (computed from the actual native dimensions of each radiograph) before computing Euclidean distance: this keeps the geometry of rectangular radiographs after resizing intact (unlike the approximation given by a single averaged conversion factor).

5.4.4 Three-Level Validation Hierarchy with Matched Statistical Analysis

The study evaluated the same localization pipeline under a primary fixed partition (Holdout 70/30), a complementary ensemble design evaluated on the identical held-out cohort (K-Fold 70/30), and an exhaustive out-of-sample protocol (Nested LOOCV). Rather than comparing aggregate values from unrelated test sets, the paired statistical analyzes used identical Sample IDs and the corresponding $30 \times 19 = 570$ matched landmark pairs, combining hypothesis tests with effect-size measures and with exact McNemar analyzes of

the landmark-level classifications. This distinction prevents differences in test-set composition from being misinterpreted as differences in model behavior.

5.4.5 Reproducible and Traceable Workflow

The experimental workflow records fold-specific training durations, inner validation behavior, held-out predictions, image-level metrics, landmark-level errors, and the arrays required for subsequent statistical analysis. These records support the traceability of the relationship between model development, held-out prediction, physical-error computation, and statistical interpretation, and they allow every reported value to be verified against the stored experimental outputs.

5.5 Limitations of the Study

The conclusions of this research must be interpreted within the following limitations:

- the PKU dataset contains only 102 radiographs, limiting the anatomical and radiographic variability available for training and evaluation;
- the study does not include an independent external institutional cohort; the Nested LOOCV predictions are out-of-sample within the PKU dataset but do not constitute external-domain validation;
- a limited number of radiographs and landmarks produces severe localization errors despite favorable performance for the majority of predictions;
- localization accuracy and variability remain strongly landmark dependent, so a global MRE cannot represent the reliability of every anatomical point;
- the primary and K-Fold held-out test cohort contains 30 radiographs, which limits the precision of their aggregate estimates;
- the paired inter-protocol statistical analyzes are consequently restricted to 30 matched images and 570 matched landmark pairs, which limits their statistical power for small effects;
- differences between validation protocols cannot be attributed causally to individual methodological components without additional controlled ablation experiments; and
- the study evaluates a single compact U-Net-based architectural family and does not constitute an exhaustive benchmark against alternative deep-learning architectures.

These limitations define the scope of the conclusions rather than invalidate the experimental findings. In particular, the combined primary and exhaustive evaluation provides a rigorous assessment of the available dataset and exposes difficult cases that could remain hidden under a smaller held-out test partition.

5.6 Future Work and Perspectives

5.6.1 Independent External Validation

The most important extension is the evaluation of the proposed pipeline on independent external datasets acquired from different institutions, imaging systems, and patient populations. Such an evaluation would test external-domain generalization beyond the PKU cohort, ideally while preserving the same image-specific anisotropic physical evaluation principles so that differences between datasets are not confounded by inconsistent coordinate or millimeter conversion procedures.

5.6.2 Larger and More Diverse Populations

Larger datasets would expose the model to a broader range of craniofacial morphology and radiographic variability, would allow larger independent validation and test populations, and could support stratified analyzes according to clinically relevant sources of variability. Such extensions would help determine whether the difficult cases identified in this research represent isolated observations or recurring patterns within broader populations.

5.6.3 Analysis of Difficult Cases and Landmarks

The difficult radiographs and landmarks identified across the three protocols should be examined together with their predicted heatmaps, reference annotations, image characteristics, and localization errors. The objective would be to distinguish model-localization failures from cases in which the reference landmark itself is difficult to identify consistently, and to transform the extreme errors from aggregate observations into information that can guide targeted methodological improvements.

5.6.4 Prediction Uncertainty and Failure Detection

The coexistence of accurate predictions for the majority of cases with isolated severe errors motivates the development of uncertainty or quality-control mechanisms. Heatmap concentration, disagreement among independently trained models, or calibrated uncertainty estimates could be used to flag predictions for manual review before any clinical use, thereby introducing an automatic control layer between landmark prediction and clinical decision support.

5.6.5 Controlled Ablation Studies

Future experiments should compare Soft-Argmax with alternative decoding strategies using identical predicted heatmaps and identical anisotropic physical-error calculations, and should isolate the contribution of heatmap-level

ensemble averaging by comparing individual fold-specific models with their averaged ensemble under the same held-out test set. Controlled studies of the Gaussian heatmap width, alternative loss weightings for difficult landmarks, and coarse-to-fine refinement strategies would likewise allow the individual contribution of each pipeline component to be established. Finally, systematic variations of the network depth, that is, the number of encoder and decoder levels, and of the feature-channel configuration (64–128–256), whose selection was justified in Section 3.7.1 as a fixed design setting, would directly test whether the two-level configuration employed here is optimal.

5.6.6 Toward Clinical Evaluation

The present results do not constitute a clinical validation of an autonomous diagnostic system. Before clinical use can be considered, future work should evaluate the pipeline on independent populations, relate landmark errors to their downstream effect on derived cephalometric measurements, and establish mechanisms for detecting unreliable predictions. Because identical radial localization errors do not necessarily produce identical effects on every derived measurement, a complete clinical evaluation would extend the sequence from landmark localization to clinical agreement on the derived cephalometric measurements.

5.7 Final Conclusion

In this work, we developed and evaluated a complete, reproducible framework for automatic landmark localization on cephalometric images. The framework consists of a small two-level U-Net, Gaussian heatmap regression, continuous Soft-Argmax decoding, image specific anisotropic millimeter conversion, and three level validation hierarchy where the main Holdout 70/30 is supported by a K-Fold ensemble evaluation and an exhaustive Nested LOOCV robustness analysis.

With the primary protocol, the pipeline localized 19 cephalometric landmarks with MRE of 2.1663 mm and SDR@2 of 80.35% on a fully isolated held-out test set. The complementary evaluations were very close to each other and matched statistical analyzes did not find statistically significant differences between validation designs on the 30 common radiographs.

The main contribution of this work is not in a new neural-network architecture or any particular algorithmic piece but rather in bringing spatial prediction, continuous coordinate extraction, physically consistent measurement, strict held-out validation, and matched statistical interpretation into one reproducible framework. The study demonstrates that validation design is part of the methodology (not an after-the-fact technical step) and that strict control of the held-out evaluation is necessary to obtain defensible estimates of localization performance.

Future extensions should preserve this rigor but address the limitations mentioned above: validation on independent and larger populations; systematic study of difficult cases/landmarks; uncertainty estimation; ablation studies controlled for landmark errors downstream; evaluation of theeffect of landmark errors on derived cephalometric measurements. These are the main directions toward stronger generalization and eventually clinical

assessment.

Bibliography

Kim Martin, Christos Katsaros, Robert Brylka, Ulrich Schwanecke, y Ralf Schulze. Influence of patient motion on definition of typical cephalometric reference points in digital horizontally scanning cephalometric radiography. *Head & Face Medicine*, 21(1):18, 2025. doi:10.1186/s13005-025-00491-z.

Rossana Izzetti, Maria Pisano, Chiara Cinquini, C. Moreno Lorenzo, Antonio Barone, y Cosimo Nardi. Lateral cephalometric radiography: Principles, common positioning errors, and ai-driven quality control. *Diagnostics*, 16(4):543, 2026. doi:10.3390/diagnostics16040543.

Md. Shaheenur Islam Sumon, Khandaker Reajul Islam, Tanzila Rafique, Gazi Shamim Hassan, Md. Sakib Abrar Hossain, Kanchon Kanti Podder, Noha Barhom, Faleh Tamimi, Abdulrahman Alqahtani, y Muhammad E. H. Chowdhury. Self-cephalonet: A two-stage novel framework using operational neural network for cephalometric analysis. *Neural Computing and Applications*, 37(16):9777–9805, 2025. doi:10.1007/s00521-025-11097-6.

Naveen Bansal, Jeetinder Singla, Gurmeet Gera, Monika Gupta, y Gurpreet Kaur. Reliability of natural head position in orthodontic diagnosis: A cephalometric study. *Contemporary Clinical Dentistry*, 3(2):180–183, 2012. doi:10.4103/0976-237X.96824.

A. R. Durão, P. Pittayapat, M. I. B. Rockenbach, R. Olszewski, S. Ng, A. P. Ferreira, y R. Jacobs. Validity of 2d lateral cephalometry in orthodontics: a systematic review. *Progress in Orthodontics*, 14(1):31, 2013. doi:10.1186/2196-1042-14-31.

Nuha Junaid, Niha Khan, Naseer Ahmed, Maria Shakoor Abbasi, Gotam Das, Afsheen Maqsood, Abdul Razzaq Ahmed, Anand Marya, Mohammad Khursheed Alam, y Artak Heboyan. Development, application, and performance of artificial intelligence in cephalometric landmark identification and diagnosis: A systematic review. *Healthcare*, 10(12):2454, 2022. doi:10.3390/healthcare10122454.

Sheldon Baumrind y Robert C. Frantz. The reliability of head film measurements: 1. Landmark identification. *American Journal of Orthodontics*, 60(2):111–127, 1971. doi:10.1016/0002-9416(71)90028-5.

Falk Schwendicke, Akhilanand Chaurasia, Lubaina Arsiwala, Jae-Hong Lee, Karim Elhennawy, Paul-Georg Jost-Brinkmann, Flavio Demarco, y Joachim Krois. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. *Clinical Oral Investigations*, 25(7):4299–4309, 2021. doi:10.1007/s00784-021-03990-w.

Germana de Queiroz Tavares Borges Mesquita, Walbert de Andrade Vieira, Maria Tereza Campos Vidigal, Bruno Augusto Nassif Travençolo, Thiago Leite Beaini, Rubens Spin-Neto, Luiz Renato Paranhos, y Rui Barbosa de Brito Júnior. Artificial intelligence for detecting cephalometric landmarks: A systematic review and meta-analysis. *Journal of Digital Imaging*, 36(3):1158–1179, 2023. doi:10.1007/s10278-022-00766-w.

Mu Chen, Zhi-Cai Feng, Xue Liu, Zheng-Ming Li, Bin Cai, y Da-Wei Wang. Impact of malocclusion on oral health-related quality of life in young adults. *Angle Orthodontist*, 85(6):986–991, 2014. doi:10.2319/101714-743.1.

Lucero Mendoza-Oropeza, Ana Fabiola Meléndez-Rodríguez, Araceli Ortíz-Ocón, y Héctor Pérez-García. Prevalence of malocclusions associated with pernicious oral habits in a Mexican sample. *Revista Mexicana de Ortodoncia*, 2(4):220–227, 2014. doi:10.1016/S2395-9215(16)30151-9.

B. Holly Broadbent. A new X-ray technique and its application to orthodontia. *The Angle Orthodontist*, 1(2): 45–66, 1931. doi:10.1043/0003-3219(1931)001<0045:ANXTAI>2.0.CO;2.

Alexander Jacobson y Richard L. Jacobson. *Radiographic Cephalometry: From Basics to 3-D Imaging*. Quintessence Publishing Co, Inc, Batavia, IL, 2nd edition, 2006. ISBN 978-0867154610.

Rakesh Avadesh Singh, Merin Sara Sojan, Sameer Narkhede, y Pranita Jadhav. Comparative evaluation of accuracy and time efficiency of upper airway analysis using manual tracing method and artificial intelligence application - a cross-sectional, retrospective study. *Journal of Carcinogenesis*, 24(2s):1132–1141, 2025. doi:10.64149/j.carcinog.24.2s.1132-1141.

Rumeesha Zaheer, Hafiza Zainab Shafique, Zahra Khalid, Rooma Shahid, Abdullah Jan, Tooba Zahoor, Ramsha Nawaz, y Mehak Hassan. Comparison of semi and fully automated artificial intelligence driven softwares and manual system for cephalometric analysis. *BMC Medical Informatics and Decision Making*, 24, 2024. doi:10.1186/s12911-024-02664-3.

Julie Hendrickx, Rellyca Sola Gracea, Michiel Vanheers, Nicolas Winderickx, Flavia Preda, Sohaib Shujaat, y Reinhilde Jacobs. Can artificial intelligence-driven cephalometric analysis replace manual tracing? A systematic review and meta-analysis. *European Journal of Orthodontics*, 46(4), 2024. doi:10.1093/ejo/cjae029.

A. R. Durão, A. Alqerban, A. P. Ferreira, y R. Jacobs. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study. *Imaging Science in Dentistry*, 45(4):213–220, 2015. doi:10.5624/isd.2015.45.4.213.

Lan Nan, Min Tang, Bohui Liang, Shuixue Mo, Na Kang, Shaohua Song, Xuejun Zhang, y Xiaojuan Zeng. Automated sagittal skeletal classification of children based on deep learning. *Diagnostics*, 13(10):1719, 2023. doi:10.3390/diagnostics13101719.

Gauthier Dot, Thomas Schouman, Shang-Hung Chang, Frédéric Rafflenbeul, Adeline Kerbrat, Philippe Rouch, y Laurent Gajny. Automatic 3-dimensional cephalometric landmarking via deep learning. *Journal of Dental Research*, 101(11):1380–1387, 2022. doi:10.1177/00220345221112333.

Symeon Sitaras, Ioannis A. Tsolakis, Marina Gelsini, Apostolos I. Tsolakis, Falk Schwendicke, Thomas Gerhard Wolf, y Paula Perlea. Applications of Artificial Intelligence in Dental Medicine: A Critical Review. *International Dental Journal*, 75(2):474–486, 2025. doi:10.1016/j.identj.2024.11.009.

Alejandro Newell, Kaiyu Yang, y Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016. doi:10.1007/978-3-319-46484-8_29.

Christian Payer, Darko Štern, Horst Bischof, y Martin Urschler. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical Image Analysis*, 54:207–219, 2019. doi:10.1016/j.media.2019.03.007.

Franz Thaler, Christian Payer, Martin Urschler, y Darko Stern. Modeling annotation uncertainty with gaussian heatmaps in landmark localization. *Machine Learning for Biomedical Imaging*, 2021. doi:10.59275/j.melba.2021-77a7.

Olaf Ronneberger, Philipp Fischer, y Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, pages 234–241. Springer, 2015. doi:10.48550/arXiv.1505.04597.

S Rashmi y Vani Ashok. A novel method for cephalometric landmark regression using convolutional neural networks and local binary pattern. In *International Conference on Computer Vision and Image Processing*, pages 315–326. Springer, 2020. doi:10.1007/978-981-16-1086-8_28.

Ruben Pauwels. Cone beam CT for dental and maxillofacial imaging: dose matters. *Radiation Protection Dosimetry*, 165(1-4):156–161, 2015. doi:10.1093/rpd/ncv057.

Touko Kaasalainen, Marja Ekholm, Teemu Siiskonen, y Mika Kortesniemi. Dental cone beam ct: An updated review. *Physica Medica*, 88:193–217, 2021. doi:10.1016/j.ejmp.2021.07.007.

Mamta Juneja, Poojita Garg, Ravinder Kaur, Palak Manocha, Prateek, Shivam Batra, Pradeep Singh, Shaswat Singh, y Prashant Jindal. A review on cephalometric landmark detection techniques. *Biomedical Signal Processing and Control*, 66:102486, 2021. doi:10.1016/j.bspc.2021.102486.

Jesús Fernández Sánchez. *Manual de prácticas de odontopediatría, ortodoncia y odontología preventiva*. Editorial Médica Ripano, Madrid, Spain, 2006. ISBN 84-609-7414-6.

Martin Šavc, Gašper Sedej, y Božidar Potočnik. Cephalometric landmark detection in lateral skull x-ray images by using improved spatialconfiguration-net. *Applied Sciences*, 12(9), 2022. doi:10.3390/app12094644.

Hannah Kim, Eungjune Shim, Jungeun Park, Yoon-Ji Kim, Uilyong Lee, y Youngjun Kim. Web-based fully automated cephalometric analysis by deep learning. *Computer Methods and Programs in Biomedicine*, 194: 105513, 2020. doi:10.1016/j.cmpb.2020.105513.

Su Yang, Eun Sun Song, Eun Seung Lee, Se-Ryong Kang, Won-Jin Yi, y Seung-Pyo Lee. Ceph-net: automatic detection of cephalometric landmarks on scanned lateral cephalograms from children and

adolescents using an attention-based stacked regression network. *BMC Oral Health*, 23(1):803, 2023. doi:10.1186/s12903-023-03452-7.

Rosalia Leonardi, Daniela Giordano, Flavia Maiorana, y Concetto Spampinato. Automatic cephalometric analysis: A systematic review. *The Angle Orthodontist*, 78(1):145–151, 2008. doi:10.2319/120506-491.1.

R. Neeraja y L. Jani Anbarasi. A critical review of artificial intelligence based techniques for automatic prediction of cephalometric landmarks. *Artificial Intelligence Review*, 58(5):1–56, 2025. doi:10.1007/s10462-025-11135-8.

Falk Schwendicke, Wojciech Samek, y Joachim Krois. Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, 99(7):769–774, 2020. doi:10.1177/0022034520915714.

Andrew Innes, Victor Ciesielski, John Mamutil, y Sabu John. Landmark detection for cephalometric radiology images using pulse coupled neural networks. In *Proceedings of the International Conference on Artificial Intelligence (IC-AI'02)*, pages 511–517, 2002.

Jia-Kuang Liu, Yen-Ting Chen, y Kuo-Sheng Cheng. Accuracy of computerized automatic identification of cephalometric landmarks. *American Journal of Orthodontics and Dentofacial Orthopedics*, 118(5):535–540, 2000. doi:10.1067/mod.2000.110168.

Vicente Grau, M Alcaniz, MC Juan, Carlos Monserrat, y Christian Knoll. Automatic localization of cephalometric landmarks. *Journal of Biomedical Informatics*, 34(3):146–156, 2001. doi:10.1006/jbin.2001.1014.

Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, y Jim Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. doi:10.1006/cviu.1995.1004.

Timothy F. Cootes, Gareth J. Edwards, y Christopher J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. doi:10.1109/34.927467.

A. A. Saad, A. El-Bialy, A. H. Kandil, y A. A. Sayed. Automatic cephalometric analysis using active appearance model and simulated annealing. *ICGST International Journal on Graphics, Vision and Image Processing*, 6: 51–67, 2006. Special Issue on Image Retrieval and Representation.

Jesús Montúfar, Marcelo Romero, y Rogelio J Scougall-Vilchis. Automatic 3-dimensional cephalometric landmarking based on active shape models in related projections. *American Journal of Orthodontics and Dentofacial Orthopedics*, 153(3):449–458, 2018. doi:10.1016/j.ajodo.2017.06.028.

Claudia Lindner, C. William Wang, Ching-Tang Huang, Cheng-Hsuan Li, Shih-Wen Chang, y Timothy F. Cootes. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. *Scientific Reports*, 6(1):33581, 2016. doi:10.1038/srep33581.

Yu Song, Xu Qiao, Yutaro Iwamoto, y Yen-wei Chen. Automatic cephalometric landmark detection on x-ray images using a deep-learning method. *Applied Sciences*, 10(7), 2020. doi:10.3390/app10072547.

Bulat Ibragimov, Boštjan Likar, Franjo Pernuš, y Tomaž Vrtovec. Automatic cephalometric X-ray landmark detection by applying game theory and random forests. In *Proceedings of the IEEE ISBI 2014 Grand Challenge: Automatic Cephalometric X-Ray Landmark Detection Challenge*, pages 1–8, 2014.

Leonardo Rundo y Carmelo Militello. Image biomarkers and explainable ai: handcrafted features versus deep learned features. *European Radiology Experimental*, 8(1):130, 2024. doi:10.1186/s41747-024-00529-y.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud A. A. Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, y Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi:10.1016/j.media.2017.07.005.

Dinggang Shen, Guorong Wu, y Hi-I Suk. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017. doi:10.1146/annurev-bioeng-071516-044442.

Heidi Lindroth, Keivan Nalaie, Roshini Raghu, Ivan N Ayala, Charles Busch, Anirban Bhattacharyya, Pablo Moreno Franco, Daniel A Diedrich, Brian W Pickering, y Vitaly Herasevich. Applied artificial intelligence in healthcare: a review of computer vision technology application in hospital settings. *Journal of Imaging*, 10(4): 81, 2024. doi:10.3390/jimaging10040081.

Chao Ma, Yuan Gu, y Ziyang Wang. Triconvunext: a pure cnn-based lightweight symmetrical network for biomedical image segmentation. *Journal of Imaging Informatics in Medicine*, 37(5):2311–2323, 2024. doi:10.1007/s10278-024-01116-8.

Yann LeCun, Léon Bottou, Yoshua Bengio, y Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi:10.1109/5.726791.

Ian Goodfellow, Yoshua Bengio, y Aaron Courville. *Deep Learning*. MIT Press, 2016. URL `https://www.deeplearningbook.org/`.

D. H. Hubel y T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962. doi:10.1113/jphysiol.1962.sp006837.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980. doi:10.1007/BF00344251.

Yann LeCun, Bernhard Boser, John S. Denker, Donnie Howard, William Habbard, Lawrence D. Jackel, y Donna Henderson. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989. doi:10.1162/neco.1989.1.4.541.

Prajit Ramachandran, Barret Zoph, y Quoc V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. doi:10.48550/arXiv.1710.05941.

Matthew D. Zeiler y Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014. doi:10.1007/978-3-319-10590-1_53.

Christian Payer, Darko Štern, Horst Bischof, y Martin Urschler. Regressing heatmaps for multiple landmark localization using CNNs. In *Deep Learning and Data Labeling for Medical Applications (DLMIA), MICCAI*

2016 Workshops, volume 10008 of *Lecture Notes in Computer Science*, pages 230–238. Springer, 2016. doi:10.1007/978-3-319-46723-8_27.

Ke Sun, Bin Xiao, Dong Liu, y Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. doi:10.1109/CVPR.2019.00584.

Jonathan Long, Evan Shelhamer, y Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. doi:10.1109/CVPR.2015.7298965.

Vijay Badrinarayanan, Alex Kendall, y Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12): 2481–2495, 2017. doi:10.1109/TPAMI.2016.2644615.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, y Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. doi:10.1109/CVPR.2016.511.

Reeha Khan, Muhammad Anwaar Khalid, Kanwal Zulfiqar, Ulfat Bashir, y Muhammad Moazam Fraz. Enhanced cephalometric landmark detection using multi-scale feature learning and heatmap regression. In *Asia Pacific Advanced Network (APAN)*, Communications in Computer and Information Science, pages 29–41. Springer, 2025. doi:10.1007/978-3-031-89813-6_3.

Jiuxiang Gu, Zhenhua Wang, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, y Gang Wang. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018. doi:10.1016/j.patcog.2017.10.013.

Kaisei Takahashi, Yui Shimamura, Chie Tachiki, Yasushi Nishii, y Masafumi Hagiwara. Cephalometric landmark detection without x-rays combining coordinate regression and heatmap regression. *Scientific Reports*, 13(1): 20011, 2023. doi:10.1038/s41598-023-46919-x.

Qian Wu, Si Yong Yeo, Yufei Chen, y Jun Liu. Revisiting cephalometric landmark detection from the view of human pose estimation with lightweight super-resolution head. *arXiv preprint arXiv:2309.17143*, 2023. doi:10.48550/arXiv.2309.17143.

Yankai Jiang, Yiming Li, Xinyue Wang, Yubo Tao, Jun Lin, y Hai Lin. Cephalformer: Incorporating global structure constraint into visual features for general cephalometric landmark detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 227–237. Springer, 2022. doi:10.1007/978-3-031-16437-8_22.

Heqin Zhu, Qingsong Yao, y S. Kevin Zhou. DATR: Domain-adaptive transformer for multi-domain landmark detection. *arXiv preprint arXiv:2203.06433*, 2022. doi:10.48550/arXiv.2203.06433.

Aiden Nibali, Zhen He, Stuart Morgan, y Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018. doi:10.48550/arXiv.1801.07372.

Minmin Zeng, Zhenlei Yan, Shuai Liu, Yanheng Zhou, y Lixin Qiu. Cascaded convolutional networks for automatic cephalometric landmark detection. *Medical Image Analysis*, 68:101904, 2021. doi:10.1016/j.media.2020.101904.

S. Rashmi, Prashanth Murthy, Vani Ashok, y S. Srinath. Cephalometric skeletal structure classification using convolutional neural networks and heatmap regression. *SN Computer Science*, 3(5):336, 2022. doi:10.1007/s42979-022-01230-w.

Yueyuan Ao y Hong Wu. Feature aggregation and refinement network for 2d anatomical landmark detection. *Journal of Digital Imaging*, 36(2):547–561, 2023. doi:10.1007/s10278-022-00718-4.

Chao Dai, Yang Wang, Chaolin Huang, Jiakai Zhou, Qilin Xu, y Minpeng Xu. A cephalometric landmark regression method based on dual-encoder for high-resolution x-ray image. In *Computer Vision – ECCV 2024*, volume 15087 of *Lecture Notes in Computer Science*, pages 93–109. Springer, 2024. doi:10.1007/978-3-031-73397-0_6.

Dmitrii Lachinov, Alexandra Getmanskaya, y Vadim Turlapov. Cephalometric landmark regression with convolutional neural networks on 3d computed tomography data. *Pattern Recognition and Image Analysis*, 30: 355–363, 2020. doi:10.1134/S1054661820030165.

Muhammad Anwaar Khalid, Kanwal Zulfiqar, Ulfat Bashir, Areeba Shaheen, Rida Iqbal, Zarnab Rizwan, Ghina Rizwan, y Muhammad Moazam Fraz. A benchmark dataset for automatic cephalometric landmark detection and cvm stage classification. *Scientific Data*, 12(1):1336, 2025. doi:10.1038/s41597-025-05542-3.

Zhusi Zhong, Jie Li, Zhenxi Zhang, Zhicheng Jiao, y Xinbo Gao. An attention-guided deep regression model for landmark detection in cephalograms. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, y Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11769 of *Lecture Notes in Computer Science*, pages 540–548, Cham, 2019. Springer International Publishing. doi:10.1007/978-3-030-32226-7_60.

R. Neeraja y L. Jani Anbarasi. CephXNet: A deep convolutional squeeze-and-excitation model for landmark prediction on lateral cephalograms. *IEEE Access*, 11:90780–90800, 2023. doi:10.1109/ACCESS.2023.3307636.

S Rashmi, S Srinath, Karthikeya Patil, Prashanth Murthy, y Seema Deshmukh. Lateral cephalometric landmark annotation using histogram oriented gradients extracted from region of interest patches. *Journal of Maxillofacial and Oral Surgery*, 22(4):806–812, 2023. doi:10.1007/s12663-023-02025-z.