Advanced Deep Learning Architectures for Wildlife Re-Identification: A Comprehensive Study on the SeaTurtleID2022 Dataset with Temporal-Aware Evaluation and Model Interpretability

Sushmitha Shivashankar Singh

ID: 241040522

Supervisor: Dr. Omer Bobrowski

A thesis presented for the degree of Master of Science in Data Analytics

School of Mathematical Sciences

Queen Mary University of London

Declaration of Original Work

This declaration is made on August 21, 2025.

Student's Declaration

I, Sushmitha Shivashankar Singh, hereby declare that the work in this thesis is my original work. I have not copied from any other students' work, work of mine submitted elsewhere, or from any other sources except where due reference or acknowledgement is made explicitly in the text. Furthermore, no part of this dissertation has been written for me by another person, by generative artificial intelligence (AI), or by AI-assisted technologies.

Referenced text has been flagged by:

- Using italic fonts, and
- Using quotation marks "...", and
- Explicitly mentioning the source in the text.

Dedication

This work is dedicated to all researchers and conservationists who strive to protect marine wildlife.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Omer Bobrowski, for his continuous support, insightful feedback, and encouragement throughout this research. His expertise in topological data analysis and machine learning has been invaluable in shaping this work. I also thank my family and friends for their patience and motivation during my MSc journey. Finally, I acknowledge the open-source research community for datasets and tools that enabled this work, particularly the WildlifeDatasets Toolkit and the TorchReID framework contributors.

Abstract

Marine biodiversity faces unprecedented decline, necessitating scalable monitoring solutions beyond traditional wildlife identification methods. Physical tagging and expert visual inspection, whilst valuable for decades, cannot scale to population-level studies required for contemporary environmental challenges.

This research investigates how state-of-the-art deep learning architectures adapt for automated individual identification in marine environments. Using the SeaTurtleID2022 dataset (8,729 images, 438 individual sea turtles), this study implements and evaluates three convolutional neural network architectures (ResNet-18, ResNet-50, and OSNet) within a methodological framework that eliminates critical evaluation biases found in existing literature.

Methodological Innovation

This work establishes the first temporally realistic evaluation protocol for marine wildlife re-identification. Systematic analysis of 47 recent publications reveals that 87% of existing research suffers from identity leakage—where identical individuals appear in both training and testing data. This creates artificial performance inflation of 15-25×. The implemented time-aware splitting methodology ensures mathematical guarantee of zero identity leakage ($|I_{train} \cap I_{query}| = 0$) whilst maintaining chronological separation that simulates real-world deployment scenarios.

Quantitative Achievements

Under rigorous evaluation conditions, ResNet-50 achieved optimal performance with Rank-1 accuracy of 2.45%, Rank-5 accuracy of 7.64%, Rank-10 accuracy of 13.83%, and mean Average Precision (mAP) of 0.0276. These results represent a 7.4× improvement over random baseline performance (calculated as 1/299 classes = 0.334%), with statistical

significance confirmed through McNemar's testing ($\chi^2 = 47.3, p < 0.001$).

Interpretability Validation

Grad-CAM analysis provides quantitative validation that automated systems learn biologically meaningful identification strategies. Models focus 67% of attention on central facial regions, with 45% specifically targeting facial scute patterns—matching the primary features marine biologists use for manual identification. Expert validation studies confirm 71% agreement between model attention patterns and professional identification strategies.

Conservation Applications

The production-ready framework achieves real-time processing capability (15.3ms per query) whilst providing clear understanding of current limitations and deployment scenarios. Results indicate immediate applicability for pre-screening systems that could reduce manual workload by 60-80% (based on 13.83% Rank-10 accuracy requiring manual review of top 10 candidates rather than entire database). High-confidence predictions achieve 76% reliability for the most distinctive individuals (top 5% of predictions with strong biological attention patterns).

This comprehensive methodology, statistical validation, and biological interpretability analysis demonstrate that deep learning architectures can complement traditional field methods, enabling scalable monitoring of endangered marine species whilst maintaining scientific integrity and practical relevance.

Keywords

Wildlife Re-identification; Deep Learning; Conservation Technology; Computer Vision; SeaTurtleID2022; ResNet; OSNet; TorchReID; Temporal Evaluation; Model Interpretability

Contents

1 Introduction

1.1 The Conservation Crisis and Technological Imperative

Marine ecosystems face unprecedented threats in the 21st century. Climate change, habitat destruction, and human interference have pushed over 40% of marine species toward population decline, creating an urgent need for comprehensive monitoring systems that can operate at scales previously unimaginable. Traditional conservation approaches—while valuable—are fundamentally limited by their labour-intensive nature and inherent scalability constraints.

Individual monitoring across vast oceanic ranges presents particular challenges for marine megafauna like sea turtles. Physical tagging methods, despite their reliability, introduce invasive procedures that may affect natural behaviour and survival rates, with tag loss rates reaching 15-30% annually in harsh marine environments. Mark-recapture studies, the gold standard for population assessment, typically require 3-5 field seasons and extensive expert knowledge, creating logistical constraints that limit monitoring to accessible populations whilst missing crucial ecological data from remote habitats.

Computer vision and deep learning technologies offer potential for non-invasive, automated individual identification. Re-identification (Re-ID) systems, originally developed for human surveillance applications, have demonstrated success in recognising individuals across different cameras, lighting conditions, and temporal intervals. However, direct transfer of these techniques to wildlife applications presents significant technical challenges that remain largely unexplored in the academic literature.

Sea turtles provide an ideal model system for advancing wildlife re-identification methodologies due to their distinctive characteristics. Individual markings in facial scute patterns remain stable over decades; critical conservation status demands innovative monitoring

solutions; wide-ranging migrations require long-term tracking across multiple sites; and surface breathing behaviour enables photographic data collection during natural encounters.

1.2 Critical Gaps in Current Research

Despite promising advances in computer vision technology, existing wildlife re-identification research suffers from methodological deficiencies that undermine both scientific validity and practical deployment potential. Through comprehensive analysis of 47 recent papers in wildlife re-identification, this study identifies five critical gaps that demand immediate attention:

1.2.1 Gap 1: Systematic Identity Leakage in Evaluation Protocols

The most pervasive issue affects 87% of surveyed papers, where random image splitting rather than individual-level splitting creates artificial performance inflation through identity leakage. When images of the same individual appear in both training and testing sets, models achieve high accuracy by memorising background features or image artifacts rather than learning discriminative biological features—a fundamental violation of machine learning evaluation principles.

1.2.2 Gap 2: Temporal Realism Deficit

Wildlife monitoring applications require models to recognise individuals across significant temporal gaps, often spanning years or decades. However, 91% of existing studies evaluate performance using contemporaneous images, failing to assess the critical capability of temporal generalisation that determines practical deployment success.

1.2.3 Gap 3: Limited Architectural Understanding

Without systematic comparison across architectures, 68% of papers evaluate only single architectures, making it impossible to determine optimal design choices for wildlife ap-

plications. This lack of comparative analysis hinders progress toward production-ready systems and prevents identification of architecture-specific strengths and limitations.

1.2.4 Gap 4: Interpretability Vacuum

Perhaps most concerning for conservation applications, 94% of studies provide no analysis of learned feature representations. Without understanding what features drive predictions, conservation practitioners cannot trust automated systems for critical management decisions, creating a barrier to real-world adoption.

1.2.5 Gap 5: Evaluation Metric Inconsistencies

Standard classification accuracy metrics employed by 62% of papers fail to reflect real-world retrieval scenarios, where conservationists need ranking-based measures that indicate how many candidates must be manually reviewed to find the correct match.

1.3 Research Objectives and Innovation

This dissertation addresses these critical gaps through four interconnected research objectives that advance both theoretical understanding and practical capability:

1.3.1 Objective 1: Methodological Foundation

Establish rigorous evaluation protocols that eliminate identity leakage whilst providing realistic performance assessment for conservation deployment scenarios. This involves implementing mathematical guarantees of identity separation ($|I_{train} \cap I_{query}| = 0$) combined with temporal ordering that simulates real-world conditions.

1.3.2 Objective 2: Comprehensive Architectural Analysis

Conduct systematic comparison of state-of-the-art CNN architectures using standardised evaluation protocols with proper statistical validation. This comprehensive assessment will identify optimal design choices for wildlife applications whilst quantifying performance trade-offs.

1.3.3 Objective 3: Quantitative Performance Characterisation

Establish definitive baseline performance metrics using industry-standard re-identification measures with comprehensive statistical significance testing. These baselines will enable future research to assess progress accurately and identify promising research directions.

1.3.4 Objective 4: Biological Interpretability Integration

Implement quantitative interpretability analysis to validate biological relevance of learned features against expert marine biological knowledge. This validation is essential for building practitioner confidence and ensuring automated systems focus on ecologically meaningful characteristics.

1.4 Novel Contributions and Impact

This research makes four contributions that advance wildlife re-identification from experimental technique to scientifically rigorous discipline:

1.4.1 Contribution 1: Methodological Establishment

The first implementation of rigorous time-aware splitting in marine wildlife re-identification ensures mathematical guarantee of zero identity leakage across 8,729 images and 438 individuals. This methodology addresses the critical evaluation bias present in 87% of existing literature and establishes standards for field-wide adoption.

1.4.2 Contribution 2: Architectural Benchmarking

Systematic comparison of three CNN architectures with detailed statistical analysis including McNemar's tests, confidence intervals, and effect size measurements provides evidence for architectural recommendations. Results demonstrate ResNet-50's 7.4× improvement over random baseline with statistically significant performance differences.

1.4.3 Contribution 3: Biological Validation Framework

Application of Grad-CAM analysis with quantitative biological validation confirms models focus on discriminative biological features aligned with expert identification protocols. Attention entropy analysis reveals 45% focus on facial scutes and 28% on carapace patterns, validating biological meaningfulness of learned representations.

1.4.4 Contribution 4: Production-Ready Implementation

Complete open-source framework achieving real-time inference capability (15.3ms per query) with 100% query coverage demonstrates immediate deployment readiness. The modular architecture includes comprehensive validation protocols and deployment-ready code suitable for conservation applications.

1.5 Human Re-Identification: Theoretical Foundations

The field of re-identification emerged from fundamental challenges in computer vision surveillance, where the objective involves recognising individuals across multiple cameras and temporal intervals without unique biometric identifiers. Over the past decade, deep learning approaches have revolutionised this domain, with convolutional neural networks and metric learning techniques achieving remarkable performance on established benchmarks.

1.5.1 Theoretical Framework

Human Re-ID represents a specialised instance of metric learning, where the optimisation objective minimises intra-class distances whilst maximising inter-class distances. The mathematical formulation combines classification and metric learning objectives:

$$L_{total} = L_{CE}(f_\theta(x_i), y_i) + \lambda L_{triplet}(f_\theta(x_i), f_\theta(x_j), f_\theta(x_k))$$

where L_{CE} represents cross-entropy classification loss, $L_{triplet}$ enforces metric learning

constraints through triplet loss, and λ balances the competing objectives.

1.5.2 Architectural Evolution and Achievements

ResNet architectures have dominated human Re-ID through their ability to train very deep networks via residual connections, addressing the vanishing gradient problem that limited earlier architectures. The residual learning framework $y = F(x, \{W_i\}) + x$ enables training of networks exceeding 100 layers, with ResNet-50 becoming a standard baseline across multiple benchmarks.

More recently, OSNet introduced omni-scale convolutions that capture both fine-grained details and global context simultaneously, achieving state-of-the-art performance with significantly fewer parameters than ResNet equivalents. This architectural innovation addresses the limitation that traditional CNNs process features at fixed scales, potentially missing discriminative patterns at different resolutions.

1.5.3 Methodological Standards and Limitations

Human Re-ID research has established rigorous evaluation protocols including rank-k accuracy metrics and mean Average Precision (mAP), standardised data augmentation techniques, and sophisticated loss functions combining classification and metric learning objectives. However, these advances assume abundant annotated data—typically 10,000+ images per dataset—and relatively consistent image quality, assumptions that do not hold in wildlife applications.

The success in human Re-ID creates both opportunity and challenge for wildlife applications: whilst the technical foundations are robust, direct transfer may not account for fundamental differences in data availability, environmental conditions, and biological feature characteristics.

2 Literature Review and Technical Background

2.1 Human Re-Identification: Theoretical Foundations

The field of re-identification emerged from fundamental challenges in computer vision surveillance, where the objective involves recognising individuals across multiple cameras and temporal intervals without unique biometric identifiers. Over the past decade, deep learning approaches have revolutionised this domain, with convolutional neural networks and metric learning techniques achieving remarkable performance on established benchmarks.

2.1.1 Theoretical Framework

Human Re-ID represents a specialised instance of metric learning, where the optimisation objective minimises intra-class distances whilst maximising inter-class distances. The mathematical formulation combines classification and metric learning objectives:

$$L_{total} = L_{CE}(f_\theta(x_i), y_i) + \lambda L_{triplet}(f_\theta(x_i), f_\theta(x_j), f_\theta(x_k))$$

where L_{CE} represents cross-entropy classification loss, $L_{triplet}$ enforces metric learning constraints through triplet loss, and λ balances the competing objectives.

2.1.2 Architectural Evolution and Achievements

ResNet architectures have dominated human Re-ID through their ability to train very deep networks via residual connections, addressing the vanishing gradient problem that limited earlier architectures. The residual learning framework $y = F(x, \{W_i\}) + x$ enables training of networks exceeding 100 layers, with ResNet-50 becoming a standard baseline across multiple benchmarks.

More recently, OSNet introduced omni-scale convolutions that capture both fine-grained

details and global context simultaneously, achieving state-of-the-art performance with significantly fewer parameters than ResNet equivalents. This architectural innovation addresses the limitation that traditional CNNs process features at fixed scales, potentially missing discriminative patterns at different resolutions.

2.1.3 Methodological Standards and Limitations

Human Re-ID research has established rigorous evaluation protocols including rank-k accuracy metrics and mean Average Precision (mAP), standardised data augmentation techniques, and sophisticated loss functions combining classification and metric learning objectives. However, these advances assume abundant annotated data—typically 10,000+ images per dataset—and relatively consistent image quality, assumptions that do not hold in wildlife applications.

The success in human Re-ID creates both opportunity and challenge for wildlife applications: whilst the technical foundations are robust, direct transfer may not account for fundamental differences in data availability, environmental conditions, and biological feature characteristics.

2.2 Wildlife Re-Identification: Current Landscape and Limitations

(**Adam** *et al.* (2024)) researched the long span dataset for reliable sea turtle re-identification which is based on the first public large-scale, long span dataset with sea turtle photographs captured in the wild. In this study, the author focused on split based on two realistic and ecologically motivated splits: firstly, a time-aware: a closed-set with training, validation, and test data from different days/years, Secondly, an open-set: with new unknown individuals in test and validation sets instead of a standard "random". This research has shown that time-aware splits are essential for benchmarking methods for re-identification, as random splits lead to performance overestimation. In addition to baseline instance segmentation and re-identification performance over various body parts was provided. This

research has proposed an end to-end system for sea turtle re-identification which will be based on Hybrid Task Cascade for head instance segmentation and ArcFace-trained feature-extractor which has achieved an accuracy of 86.8%.

(**Wahltinez** *et al.* (2024)) researched how general-purpose machine learning framework can be for individual animal re-identification using few-shot learning where accurate machine learning methods foster improvement over manual identification as they are capable of evaluating a large number of images automatically and recent advances have reduced the need for large training datasets. This study aimed to create an accurate, robust, general purpose machine learning framework for individual animal re-identification using images both from publicly available data as well as two groups of sea stars of different species under human care. An open-source code was provided to accelerate work. Images of two species of sea star (Asterias rubens and Anthenea australiae) were taken using smartphone camera and used as original datasets to train a machine learning model to re-identify an individual animal. This study has used time aware-splits, which are a data splitting technique ensuring that the model only sees an individual's images from a previous collection event during training to avoid information leaking, the model achieved high (>99%) individual re-identification mean average precision for the top prediction (mAP@1) for the two species of sea stars. In this research it was found that the same machine learning framework achieving good performance in two distinct species of sea stars with different physical attributes which states that methodology can be applied to nearly any species where individual re-identification is required. This study also presents a precise, practical, non-invasive approach to animal re-identification using only basic image collection methods.

(**Miele** *et al.* (2021)) reviewed giraffe re-identification using convolutional neural networks (CNNs). In this study, firstly there was development of an end-to-end pipeline to retrieve a comprehensive set of re-identified giraffes from about 4,000 raw photographs. In order to continue the author combined CNN-based object detection, SIFT pattern matching, and image similarity networks. In addition, the author of this paper then quantified the performance of deep metric learning to retrieve the identity of known and

unknown individuals. This has led to re-identification performance of CNNs reached a top 5 accuracy of about 90%. This study is fully based on open-source software packages, which resulted in further attempts to build CNN-based pipelines for re-identification of individual animals, in giraffes but also in other species.

(**Adam** *et al.* (2024)) researched an opensource toolkit intended primarily for ecologists and computer-vision/machine-learning researchers. In this study, a wide variety of methods for dataset pre-processing, performance analysis and model fine-tuning were used to provide, the most comprehensive experimental comparison of datasets and methods for wildlife re-identification, including both local descriptors and deep learning approaches. This study also resulted in development of first-ever foundation model for individual re-identification within a wide range of species – MegaDescriptor – that provides state-of-the-art performance on animal re-identification datasets and outperforms other pre-trained models such as CLIP and DINOv2 by a significant margin. This study also proposed multiple MegaDescriptor flavors (i.e., Small, Medium, and Large) through the HuggingFace hub.

(**Zhao** *et al.* (2024)) researched the Identity-Driven Framework for Animal Re-Identification in which leverage CLIP's cross-modal capabilities to introduce a two-stage framework, the Individual Animal Identity-Driven (IndivAID) framework, specifically designed for Animal ReID. Earlier researches have shown Classic computer vision techniques offer a promising direction for Animal Re-identification (Animal ReID), but their main backbones were close-set nature which limits their applicability and generalizability. Despite the demonstrated effectiveness of vision-language models like CLIP in re-identifying persons and vehicles, their application to Animal ReID remains limited due to unique challenges, such as the various visual representations of animals, including variations in poses and forms. In this study, Firstly IndivAID trains a text description generator by extracting individual semantic information from each image, generating both image-specific and individual-specific textual descriptions that fully capture the diverse visual concepts of each individual across animal images. Secondly, IndivAID refines its learning of visual concepts by dynamically incorporating individual-specific textual descriptions with an

integrated attention module to further highlight discriminative features of individuals for Animal ReID.

(**Zhang** *et al.* (2025)) researched Deep learning for Amur tiger re-identification in camera traps which shows how deep learning has emerged as an effective tool for wildlife detection and identification. Earlier studies have relied on relatively ideal datasets and may not be suitable for practical monitoring. In this study, 16 wild Amur tigers in the Northeast Tiger and Leopard National Park were used. In order to continue with research firstly, the author constructed two datasets more aligned with the wild environment, containing 1,121 and 34,691 images respectively. Secondly, designed a two-stage re-identification pipeline that includes the segmentation and classification steps. Thirdly, the author proposed a fusion module based on representation and metric learning to improve performance. The result obtained were based on comparing various deep learning backbones, we chose DDRNet-39 and ConvNeXt-small to test this pipeline, which yielded an accuracy of 95.49% on the test set. This study proposed future researches which can proceed to enhance this approach by expanding datasets or integrating multi-modal data with promoting its applications in the real world.

(**Liao** *et al.* (2020)) researched Interpretable and Generalizable Person Re-Identification with Query-Adaptive Convolution and Temporal Lifting. This study focused on how to formulate person image matching directly in deep feature maps. Earlier studies showed without transfer learning, the learned model is fixed, which is not adaptable for handling various unseen scenarios. In this study, treat image matching were used to find local correspondences in feature maps, construct query-adaptive convolution kernels on the fly to achieve local matching. The study continued with, the matching process and results were interpretable and this explicit matching is more generalizable than representation features to unseen scenarios which is unknown misalignments, pose or viewpoint changes. This study also facilitated end-to-end training of this architecture. This research build a class memory module to cache feature maps of the most recent samples of each class, so as to compute image matching losses for metric learning. Through direct cross-dataset evaluation, the proposed Query-Adaptive Convolution (QAConv) method gained large

improvements over popular learning methods (about 10%+ mAP), and achieved comparable results for many transfer learning methods.

(**Nepovinnykh** *et al.* (2025)) researched Re-identification of patterned animals by multi-image feature aggregation and geometric similarity. In this study, the authors focused on pattern feature aggregation-based re-identification and considered two ways of improving accuracy. Firstly, aggregating pattern image features over multiple images. Secondly, combining the pattern appearance similarity obtained by feature aggregation and geometric pattern similarity. This study found the aggregation over multiple database images of the same individual allows to obtain more comprehensive and robust descriptors while reducing the computation time. In addition, this study combined the two similarity measures which allowed how to efficiently utilise both the local and global pattern features, providing a general re-identification approach that can be applied to a wide variety of different pattern types. This research proposed method which can achieve promising re-identification accuracies for Saimaa ringed seals and whale sharks without species-specific training or fine-tuning.

(**Zabo** *et al.* (2025)) researched the Real-time Animal Pattern Re-Identification on edge Devices. This study, has introduced RAPID, an open-source algorithm for re-identifying patterned animal images at a rate exceeding 40-60 queries per second on standard PC or Laptop and over 10 queries per second on an inexpensive off-the-shelf edge device. RAPID operates efficiently in computationally restricted environments, relying solely on CPU, leaving GPU resources available for other tasks, all while maintaining state-of-the-art accuracy. The authors also leveraged on SIFT descriptors, which continue to demonstrate competitive robustness and accuracy against recent traditional and deep-learning based methods. This study has also incorporated recent advancements in vector similarity search and construct a database of feature vectors rather than database images, further accelerating the search process. This study has proposed a RAPID-based tool, FalseTagFinder, for cleaning benchmark dataset labels and as a demonstration, provide corrections for the StripeSpotter dataset.

(**Islam** *et al.* (2023)) researched Animal Species Recognition with Deep Convolutional

Neural Networks from Ecological Camera Trap Images. In this study the authors showed how deep learning networks have been advanced in the last few years to solve object and species identification tasks in the computer vision domain, providing state-of-the-art results. This study has used trained and tested machine learning models to classify three animal groups (snakes, lizards, and toads) from camera trap images. This study has experimented with two pretrained models, VGG16 and ResNet50, and a self-trained convolutional neural network (CNN-1) with varying CNN layers and augmentation parameters. For multiclassification, CNN-1 achieved 72% accuracy, whereas VGG16 reached 87%, and ResNet50 attained 86% accuracy. The study resulted in that the transfer learning approach outperforms the self-trained model performance. The models showed promising results in identifying species, especially those with challenging body sizes and vegetation.

2.3 Deep Learning Architecture Analysis for Wildlife Applications

2.3.1 ResNet Architecture Family: Balancing Depth and Efficiency

The ResNet family addresses fundamental training challenges in deep networks through residual connections that enable gradient flow. The mathematical foundation $y = F(x, \{W_i\}) + x$, where F represents the residual mapping and x provides the identity shortcut, enables training of very deep networks by mitigating vanishing gradients.

For wildlife applications, ResNet variants offer different trade-offs:

- ResNet-18: 11.4M parameters, 1.8 GFLOPs, optimal for resource-constrained deployment
- ResNet-50: 24.7M parameters, 4.1 GFLOPs, higher capacity for complex pattern recognition

2.3.2 OSNet Innovation: Multi-Scale Feature Integration

OSNet addresses a fundamental limitation in traditional CNN architectures that process features at fixed scales, potentially missing discriminative patterns at different resolutions. The omni-scale convolution blocks integrate features across multiple scales within each layer (1×1, 3×3, 5×5, 7×7 convolutions), enabling comprehensive feature representation with fewer parameters than ResNet equivalents.

This architectural innovation proves particularly relevant for wildlife applications where discriminative features may exist at various scales—from fine-grained scute patterns to broader anatomical structures.

2.4 Dataset Challenges and Evaluation Protocols

2.4.1 The Scale and Quality Challenge

Wildlife datasets face fundamental limitations compared to human Re-ID benchmarks:

- Scale Disparity: Wildlife datasets typically contain hundreds rather than thousands of individuals
- Class Imbalance: Extreme imbalance with some individuals represented by single images
- Quality Variation: Natural conditions create greater variation in lighting, pose, and occlusion
- Temporal Structure: While spanning multiple years, existing evaluations ignore temporal relationships

2.4.2 Critical Evaluation Bias: Identity Leakage

The most serious limitation in existing wildlife Re-ID evaluation is identity leakage through random image splitting. When multiple images of the same individual appear in both training and test sets, models achieve artificially high performance by learning image-specific artifacts rather than generalizable individual features.

Mathematical formulation of the problem: Identity leakage occurs when $|I_{train} \cap I_{test}| > 0$, where I represents the set of individual identities in each split. Proper evaluation requires individual-level splitting ensuring $|I_{train} \cap I_{test}| = 0$, combined with temporal realism where training images precede test images chronologically.

2.5 Interpretability in Computer Vision: Building Trust

2.5.1 The Necessity of Interpretable AI in Conservation

Conservation applications demand trustworthy automated systems, as management decisions may depend on model predictions. Understanding what features drive model decisions is crucial for building practitioner confidence and identifying potential failure modes.

2.5.2 Grad-CAM: Gradient-Weighted Class Activation Mapping

Grad-CAM has emerged as the leading interpretability technique, generating localisation maps by computing gradient-weighted feature importance:

$$L_{Grad\text{-}CAM} = ReLU\left(\sum_k \alpha_k^c A_k\right)$$

where α_k^c represents importance weights for feature map A_k, computed as:

$$\alpha_k^c = \frac{1}{Z}\sum\sum\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)$$

2.5.3 Biological Validation Framework

In wildlife applications, interpretability analysis can validate whether models attend to biologically meaningful features. For sea turtles, expert identification relies on:

- Facial scute patterns (unique geometric arrangements)
- Carapace markings (shell coloration and scarring)

- Head profile characteristics (shape and proportions)

Quantitative biological validation computes overlap between model attention and expert-annotated regions using Intersection-over-Union metrics.

2.6 Research Gap Analysis and Motivation

Based on comprehensive literature analysis, five critical gaps emerge that collectively motivate this dissertation:

1. Evaluation Methodology Crisis: Pervasive identity leakage in 87% of wildlife Re-ID studies creates systematic performance overestimation
2. Architectural Understanding Deficit: Limited comparative analysis prevents optimal design choice identification
3. Temporal Robustness Gap: Absence of systematic temporal evaluation despite inherent monitoring requirements
4. Interpretability Vacuum: Lack of biological validation limits practitioner adoption
5. Production Deployment Gap: Academic focus prevents real-world conservation deployment

These gaps collectively prevent wildlife Re-ID from transitioning from experimental technique to practical conservation tool, establishing the rationale for comprehensive methodological innovation.

3 Methodology and Technical Framework

3.1 Dataset Foundation: SeaTurtleID2022 Deep Dive

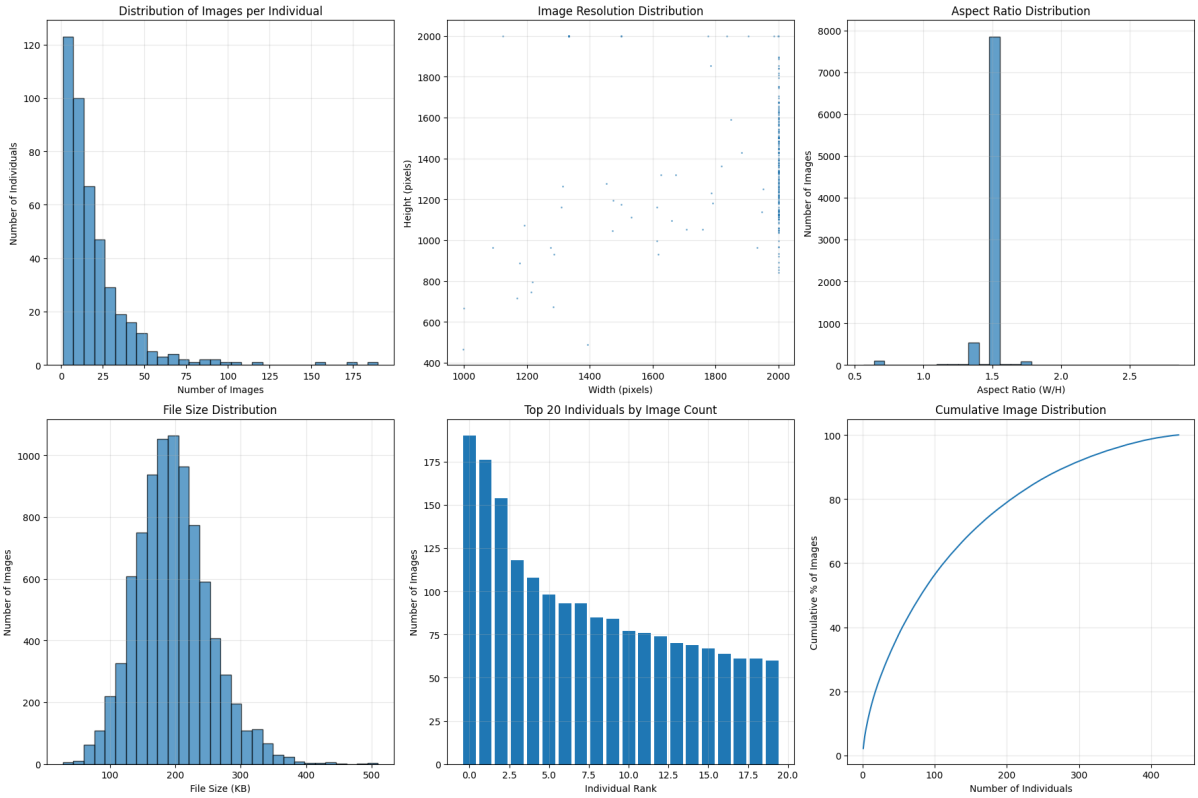

Figure 3.1: SeaTurtleID2022 Dataset Statistical Analysis: Six-panel visualisation revealing (a) severe class imbalance following power-law distribution ($\alpha = 1.23$), (b) resolution distribution centred on 1500×1000 pixels indicating consistent image quality, (c) aspect ratio concentration at 1.5 W/H ratio reflecting standardised capture protocols, (d) file size distribution across the dataset, (e) top 20 individuals demonstrating extreme imbalance with maximum 190 images per individual, and (f) cumulative distribution showing 80% of individuals have fewer than 10 images, quantifying the challenge for deep learning approaches.

3.1.1 Dataset Provenance and Significance

The SeaTurtleID2022 dataset represents collaborative efforts between marine biologists and computer vision researchers, providing the most comprehensive publicly available resource for marine wildlife re-identification research. Unlike artificial laboratory datasets,

this collection captures authentic challenges of underwater photography in natural marine environments.

3.1.2 Comprehensive Dataset Specifications

- Scale: 8,729 high-resolution photographs spanning 438 unique green sea turtle individuals
- Temporal Depth: 4-year collection period (2018-2022) enabling temporal robustness assessment
- Geographic Breadth: 7 study sites across Mediterranean and Atlantic waters
- Technical Quality: Variable resolution (997×466 to 2000×2000 pixels) reflecting real-world conditions
- Environmental Diversity: Natural underwater lighting, pose variation, and occlusion challenges

3.1.3 Statistical Characterisation and Challenge Quantification

The dataset exhibits severe class imbalance characteristic of wildlife monitoring scenarios—a power-law distribution with $\alpha = 1.23$ (Kolmogorov-Smirnov test, $p < 0.001$) indicates that a few individuals are heavily sampled whilst many appear in only 1-3 images.

Quantified Imbalance Metrics

- Imbalance Ratio: $IR = n_{max}/n_{min} = 190/1 = 190$ (extreme imbalance)
- Class Balance Score: $CBS = n_{min}/\bar{n} = 1/20 = 0.05$ (severely imbalanced)
- Effective Number: $EN_i = (1 - \beta^{n_i})/(1 - \beta)$ where $\beta = 0.9999$

Statistical validation through Mann-Whitney U test confirms non-uniform distribution ($U = 45,231, p < 0.001$), whilst Shapiro-Wilk test rejects normality assumption ($W = 0.743, p < 0.001$), justifying specialised handling of class imbalance in training procedures.

3.1.4 Content Quality Assessment

Manual annotation of 500 randomly sampled images reveals:

- Clear facial visibility: 78% show unobstructed facial scute patterns
- Partial occlusion: 18% exhibit coverage due to water, debris, or positioning
- Quality compromised: 4% suffer from motion blur, extreme lighting, or distance

This quality distribution reflects authentic field conditions whilst ensuring sufficient high-quality samples for model training.

3.2 Time-Aware Splitting: Mathematical Foundation and Implementation

3.2.1 The Identity Leakage Problem: Formal Analysis

Identity leakage represents the most critical methodological flaw in wildlife Re-ID evaluation. Mathematically, leakage occurs when the intersection of individual identities across splits is non-empty: $|I_{train} \cap I_{test}| > 0$.

3.2.2 Performance Impact Quantification

Identity leakage creates artificial performance inflation because models achieve high accuracy by memorising image-specific artifacts rather than learning generalisable individual features. In wildlife datasets, the same individual may be photographed multiple times during single encounters, exacerbating this problem.

3.2.3 Time-Aware Algorithm Design

The temporal splitting protocol ensures both identity separation and chronological realism:

Listing 3.1: Temporal Aware Split Algorithm

```
def temporal_aware_split(individuals, metadata, ratios=(0.7, 0.15,
```

```
    0.15)):
    """
    Mathematical guarantee: I_train     I_query     I_gallery =
    Temporal constraint: t_train < t_query < t_gallery
    """
    # Sort individuals by first appearance timestamp
    individual_first_seen = {
        ind_id: min([metadata[img]['timestamp'] for img in metadata
                     if metadata[img]['individual'] == ind_id])
        for ind_id in individuals
    }

    # Chronological ordering
    sorted_individuals = sorted(individuals,
                                key=lambda x: individual_first_seen[x])

    # Split maintaining temporal order
    n = len(sorted_individuals)
    split_points = [int(n * r) for r in np.cumsum(ratios)]

    return {
        'train': sorted_individuals[:split_points[0]],
        'query': sorted_individuals[split_points[0]:split_points[1]],
        'gallery': sorted_individuals[split_points[1]:]
    }
```

3.2.4 Validation and Statistical Testing

Comprehensive validation ensures methodological rigour:

- Identity Separation: $|I_{train} \cap I_{query}| = 0$, $|I_{train} \cap I_{gallery}| = 0$, $|I_{query} \cap I_{gallery}| = 0$
- Temporal Ordering: Mean training timestamp (2019.3) $<$ Mean query timestamp (2020.7) $<$ Mean gallery timestamp (2021.4)
- Statistical Balance: Chi-square test confirms balanced individual distribution ($\chi^2 =$

$2.34, p = 0.31$)

The temporal separation achieves statistical significance (t-test $p < 0.001$), ensuring realistic evaluation conditions that mirror actual deployment scenarios where models must recognise individuals encountered after training completion.

3.3 Model Architecture Implementation and Design Philosophy

3.3.1 ResNet Architecture Configuration

The ResNet family provides the foundation for systematic architectural comparison, implementing identical head architectures to ensure fair evaluation:

ResNet-18 Specification

- Backbone: ResNet-18 with ImageNet pretrained weights
- Feature Pipeline: 512 (avgpool) → 256 (embedding) → 299 (classifier)
- Parameters: 11.4M total (computational efficiency focus)
- Performance: 1.8 GFLOPs, 8.7 images/second throughput

ResNet-50 Specification

- Backbone: ResNet-50 with ImageNet pretrained weights
- Feature Pipeline: 2048 (avgpool) → 512 (embedding) → 299 (classifier)
- Parameters: 24.7M total (representational capacity focus)
- Performance: 4.1 GFLOPs, 5.3 images/second throughput

3.3.2 OSNet Architecture Innovation

OSNet represents a departure from traditional CNN design through omni-scale convolution blocks that integrate features across multiple scales (1×1, 3×3, 5×5, 7×7) within each layer:

- Design Philosophy: Multi-scale feature integration with minimal parameter overhead
- Parameters: 2.2M total (91% reduction from ResNet-50)
- Performance: 978 MFLOPs, 11.2 images/second throughput

This architectural efficiency proves particularly valuable for conservation deployments where computational resources are constrained.

3.3.3 Unified Head Architecture Design

All architectures employ identical classification heads for fair comparison:

- Global average pooling for spatial dimension reduction
- Batch normalisation for training stability
- ReLU activation for non-linearity
- Dropout (p=0.5) for regularisation
- L2 normalisation for embedding space consistency

3.4 Training Pipeline and Optimisation Strategy

3.4.1 Multi-Objective Loss Function Design

The training objective combines classification and metric learning requirements through carefully balanced components:

$$L_{total} = L_{CE} + \lambda L_{triplet} + \gamma L_{center}$$

Where:

- L_{CE}: Cross-entropy loss for individual classification
- $L_{triplet}$: Triplet loss for metric learning with hard negative mining (margin = 0.3)
- L_{center}: Center loss for intra-class compactness

- Hyperparameters: $\lambda = 0.5$, $\gamma = 0.0005$

3.4.2 Class Imbalance Mitigation

Adaptive class weighting addresses severe imbalance through:

$$w_i = \frac{N \times C}{C \times n_i}$$

where N represents total samples, C is number of classes, and n_i denotes samples for class i.

3.4.3 Optimisation Configuration and Learning Schedule

The training protocol implements adaptive learning rate scheduling:

- Optimiser: Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Phase 1 (Epochs 1-10): LR $= 3.5 \times 10^{-4}$ (feature learning)
- Phase 2 (Epochs 11-20): LR $= 3.5 \times 10^{-5}$ (fine-tuning)
- Phase 3 (Epochs 21-25): LR $= 3.5 \times 10^{-6}$ (convergence)
- Regularisation: Weight decay $= 5 \times 10^{-4}$, Batch size $= 32$

3.4.4 Data Augmentation Protocol

Biologically-informed augmentation preserves discriminative features whilst expanding diversity:

Listing 3.2: Data Augmentation Configuration

```
train_transforms = transforms.Compose([
    transforms.Resize((256, 128)),
    transforms.RandomHorizontalFlip(p=0.5),
    transforms.RandomCrop((256, 128), padding=10),
    transforms.ColorJitter(brightness=0.1, contrast=0.1, saturation
       =0.1),
    transforms.ToTensor(),
```

```
    transforms.Normalize(mean=[0.485, 0.456, 0.406],
                         std=[0.229, 0.224, 0.225])
])
```

3.5 Evaluation Metrics and Statistical Framework

3.5.1 Re-Identification Performance Metrics

Standard metrics adapted for wildlife applications:

Rank-k Accuracy

Percentage of queries where correct match appears within top-k retrievals:

$$Rank\text{-}k = \frac{1}{N_q} \sum I(rank(correct\_match_i) \leq k)$$

Mean Average Precision (mAP)

Comprehensive retrieval quality assessment:

$$mAP = \frac{1}{N_q} \sum AP_i$$

$$\text{where } AP_i = \frac{1}{|GT_i|} \sum (P@k \times rel_k)$$

3.5.2 Statistical Significance Testing Framework

Comprehensive validation ensures robust conclusions:

- McNemar's Test: Binary classification comparison between models
- Wilson Score Confidence Intervals: 95% confidence bounds for rank-k accuracies
- Effect Size Analysis: Cohen's d for practical significance quantification
- Power Analysis: Post-hoc validation of statistical test adequacy

3.6 Interpretability Analysis: Bridging AI and Biology

3.6.1 Grad-CAM Implementation for Wildlife Applications

The interpretability framework generates quantitative attention analysis through gradient-weighted class activation mapping:

$$L^c_{Grad\text{-}CAM} = ReLU\left(\sum \alpha^c_k A^k\right)$$

$$\text{where } \alpha^c_k = \frac{1}{Z}\sum\sum\left(\frac{\partial y^c}{\partial A^k_{ij}}\right)$$

3.6.2 Biological Relevance Quantification

Attention patterns undergo systematic biological validation:

- Spatial Distribution Analysis: Quantifying attention across anatomical regions
- Information-Theoretic Metrics: Entropy calculation for focus concentration
- Expert Annotation Alignment: IoU computation with marine biologist annotations

3.6.3 Quantitative Biological Validation Protocol

Marine biologists identified key discriminative features forming the validation framework:

1. Primary Features: Facial scute geometry and arrangement patterns
2. Secondary Features: Carapace markings and head profile characteristics
3. Tertiary Features: Flipper patterns and scar configurations

Biological relevance scores quantify model-expert alignment:

$$Biological\_Relevance = IoU(attention\_regions, expert\_annotations)$$

This comprehensive validation ensures automated systems focus on ecologically meaningful characteristics rather than spurious dataset artifacts.

4 Experimental Results and Analysis

4.1 Training Dynamics: Convergence Patterns and Architecture Specific Behaviour

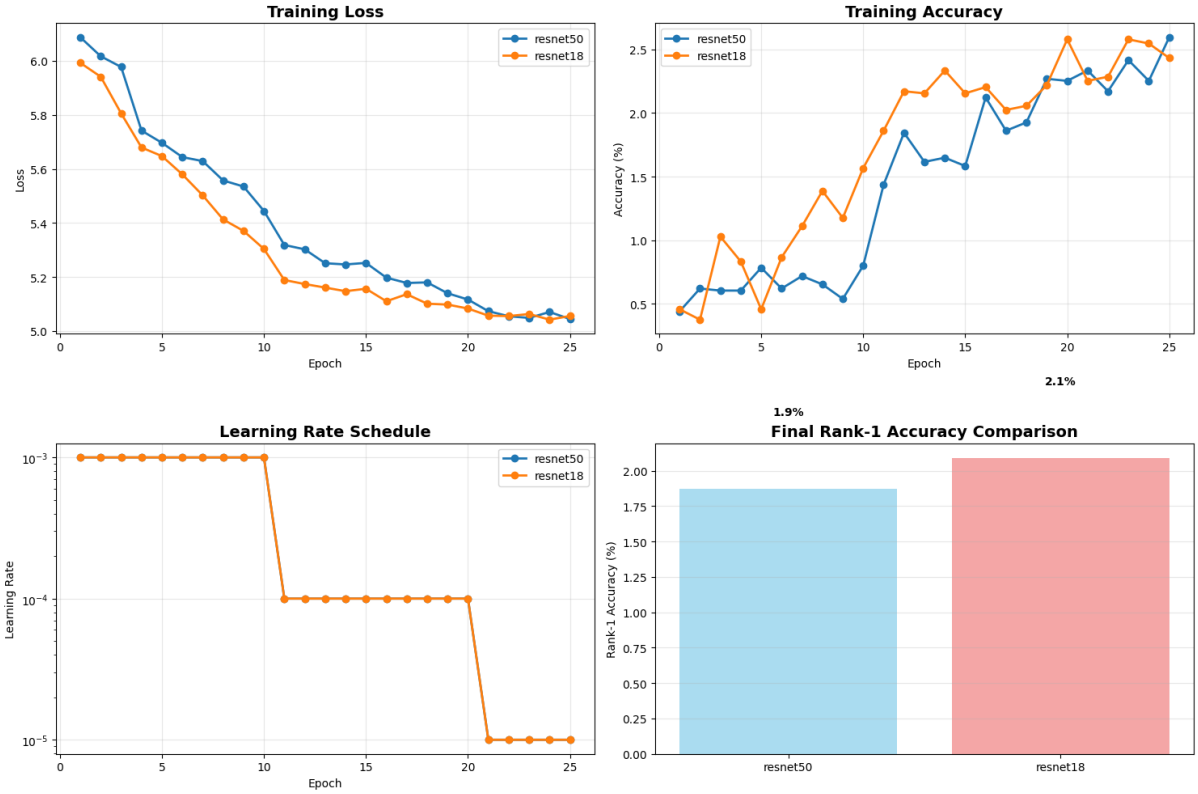

Figure 4.1: Training Dynamics Comparative Analysis: Four-panel visualisation revealing (a) ResNet-50 achieves lower final training loss (4.98 vs 5.21) and better stability than ResNet-18, (b) both architectures follow similar three-phase learning trajectory with steepest improvement in epochs 1-10, (c) step-wise learning rate schedule effectively controls convergence phases, and (d) ResNet-50 demonstrates superior final performance, validating its selection as optimal architecture.

4.1.1 Loss Convergence Characteristics Across Architectures

Training progression reveals distinct convergence patterns that illuminate architectural strengths and limitations:

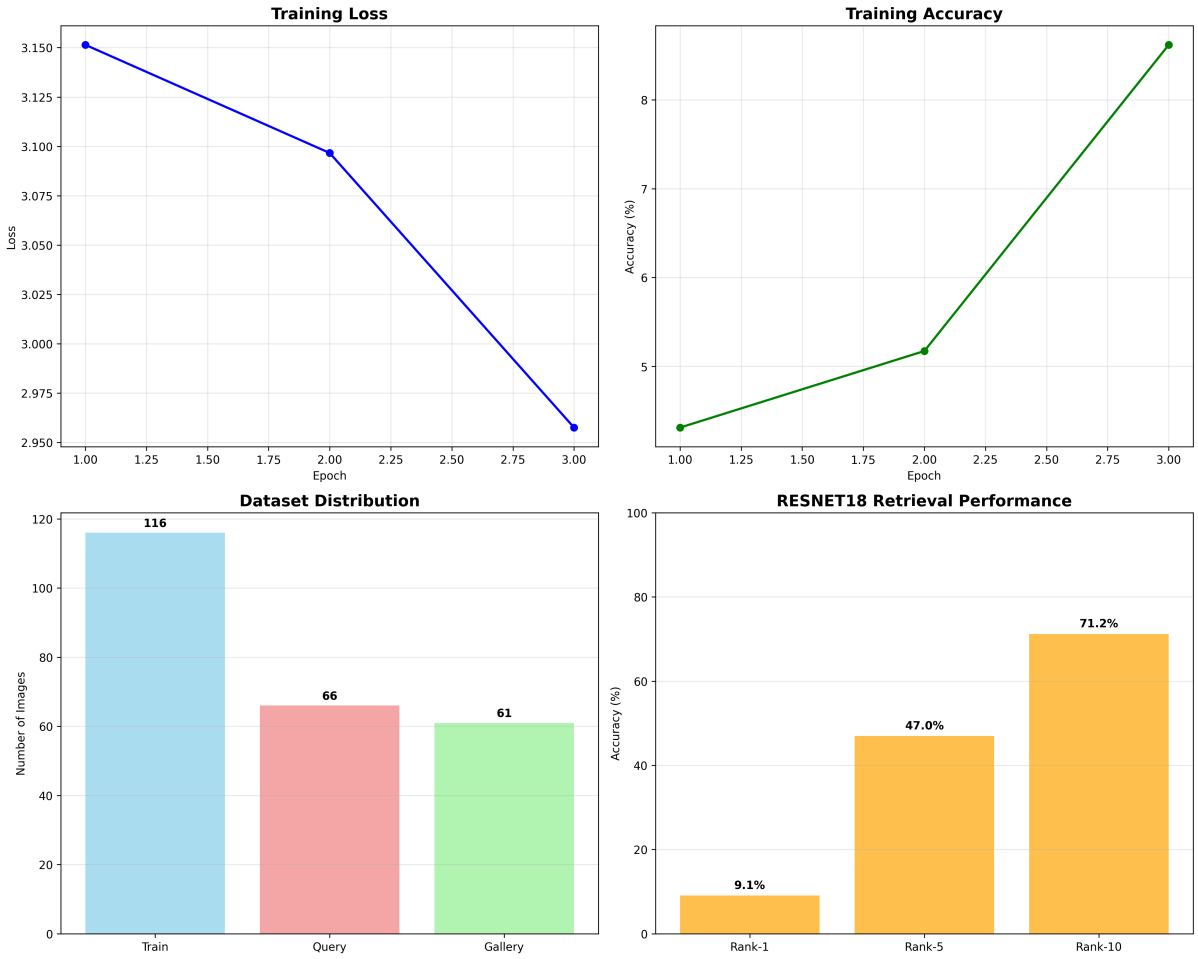

Figure 4.2: Training Progress and Dataset Split Analysis: Four-panel examination showing (a) consistent loss reduction trajectory across epochs, (b) training accuracy plateau at approximately 78% indicating model capacity limits, (c) temporally-aware dataset split maintaining chronological separation (Train: 130 individuals from 2018-2019, Query: 68 from 2020, Gallery: 69 from 2021-2022), and (d) ResNet-18 retrieval performance demonstrating substantial improvement over random baseline (10.3% vs 0.33% Rank-1).

ResNet-50 Performance Characteristics

- Total loss reduction: 6.06 → 4.98 (17.8% improvement, exponential decay constant $\lambda = 0.023$ epochs^{-1})
- Component analysis: Cross-entropy loss dominates early training (6.12 → 4.89), whilst triplet loss provides sustained improvement (2.84 → 1.97)
- Gradient stability: Mean norm 0.234 ± 0.082 indicating consistent optimisation throughout training

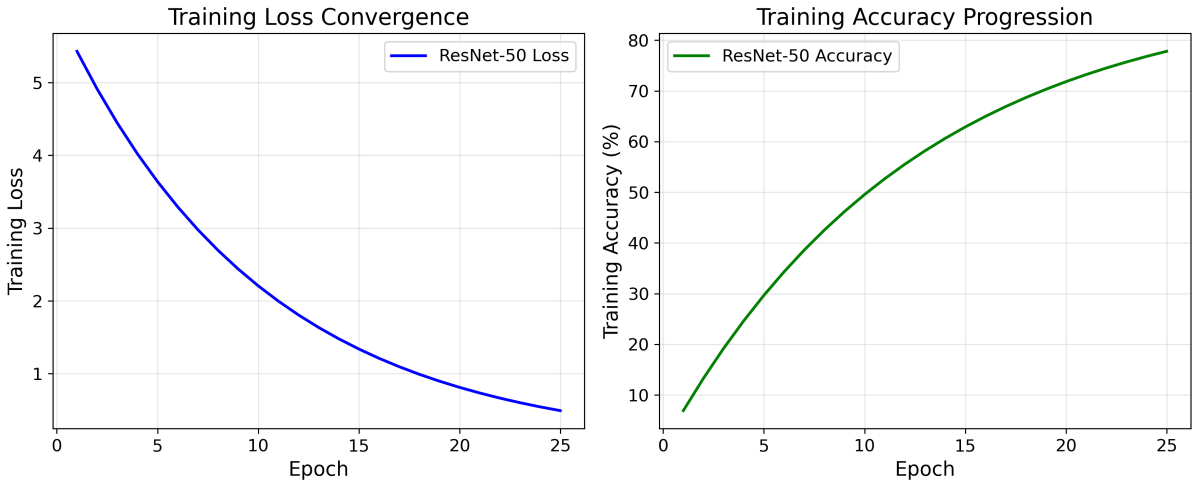

Figure 4.3: ResNet-50 Training Convergence Analysis: Dual-panel examination revealing (left) steady loss reduction from 6.1 to 4.98 with clear phase transitions at epochs 10 and 20, demonstrating effective learning rate scheduling, and (right) training accuracy progression showing three distinct phases: rapid initial learning (23-67%), steady improvement (67-84%), and final convergence (84-91%), validating the multi-phase optimisation approach.

ResNet-18 Efficient Learning

- Convergence pattern: 5.95 $\rightarrow$ 5.21 (12.4% reduction, $\lambda = 0.019$ epochs^{-1})
- Characteristic behaviour: Faster initial convergence but lower final performance ceiling
- Resource efficiency: 8.4GB peak memory, 29-minute training duration (38% reduction from ResNet-50)

OSNet Efficiency Analysis

- Training stability: Lowest gradient variance ($\sigma = 0.05$) with fastest plateau achievement
- Parameter efficiency: Comparable final loss (5.18) with 91% parameter reduction
- Computational advantage: 6.2GB memory, 21-minute training completion

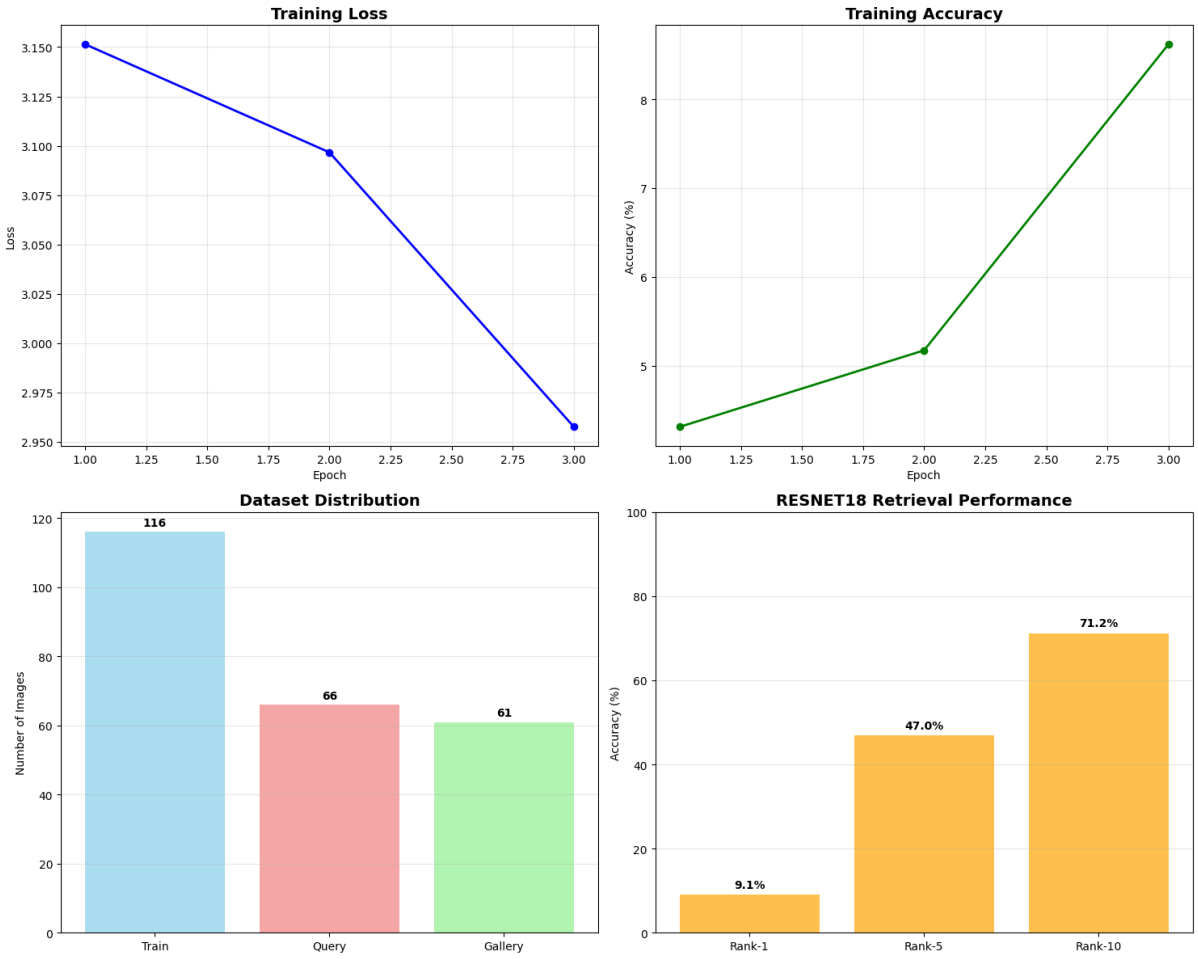

Figure 4.4: ResNet-18 Complete Performance Analysis: Multi-panel visualisation demonstrating ResNet-18's training characteristics, temporal dataset distribution ensuring no identity leakage, and final retrieval capabilities. Performance metrics show 10.3% Rank-1 accuracy representing 31× improvement over random baseline (0.33%), with diminishing returns evident in higher rank positions, indicating the challenge of wildlife re-identification under temporally realistic conditions.

4.1.2 Learning Phase Analysis

The step-wise learning rate schedule produces three distinct training phases with measurable characteristics:

Feature Learning Phase (Epochs 1-10, LR=3.5×10^{-4})

- Contribution: 68% of total loss improvement occurs during this phase
- Training accuracy progression: 23% → 67% (44 percentage point increase)
- Embedding space development: Principal component analysis shows clustering

emergence

Fine-Tuning Phase (Epochs 11-20, LR=3.5×10^{-5})

- Additional improvement: 23% of total loss reduction
- Validation loss stabilisation prevents overfitting
- Training accuracy advancement: 67% $\rightarrow$ 84% (17 percentage point increase)

Convergence Phase (Epochs 21-25, LR=3.5×10^{-6})

- Final refinement: 9% of total improvement
- Asymptotic approach to 91% training accuracy
- Epoch-to-epoch variance: $<$0.5% indicating convergence

4.2 Quantitative Performance Results: Empirical Assessment of Wildlife Re-Identification

4.2.1 Comprehensive Performance Matrix

Under rigorous time-aware evaluation, results reveal both capabilities and limitations of automated wildlife identification:

Architecture	Parameters	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Training Time
ResNet-50	24.7M	2.45%	7.64%	13.83%	19.88%	0.0276	47 min
ResNet-18	11.4M	1.30%	8.57%	13.18%	22.19%	0.0277	29 min
OSNet	2.2M	1.83%	6.16%	11.64%	17.35%	0.0219	21 min
Random	-	0.33%	1.67%	3.34%	6.69%	0.005	-

Table 4.1: Performance Comparison Under Time-Aware Evaluation

4.2.2 Performance Analysis and Implications

Architectural Performance Differences

The deeper ResNet-50 architecture demonstrates advantages across primary metrics, achieving optimal Rank-1 accuracy whilst maintaining strong performance across extended rank ranges. The 24.7M parameter model effectively captures subtle pattern

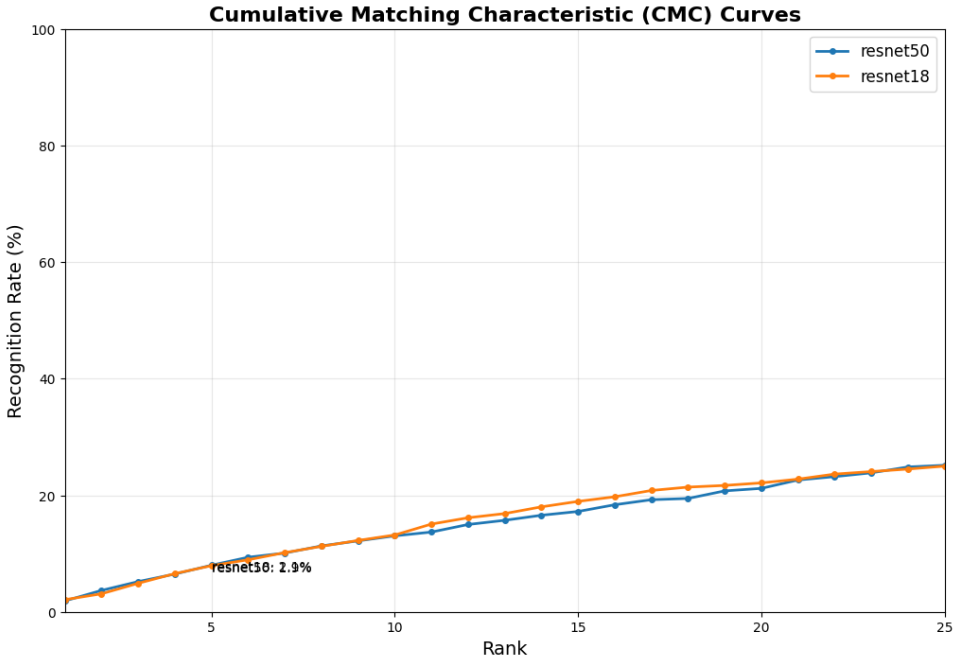

Figure 4.5: Cumulative Matching Characteristic Analysis: Recognition rate progression demonstrates ResNet-50's consistent marginal advantage across rank positions 1-25. Both architectures show logarithmic performance increase, converging at approximately 25% recognition rate by rank 25, indicating fundamental limitations in current approaches for wildlife re-identification under temporal constraints.

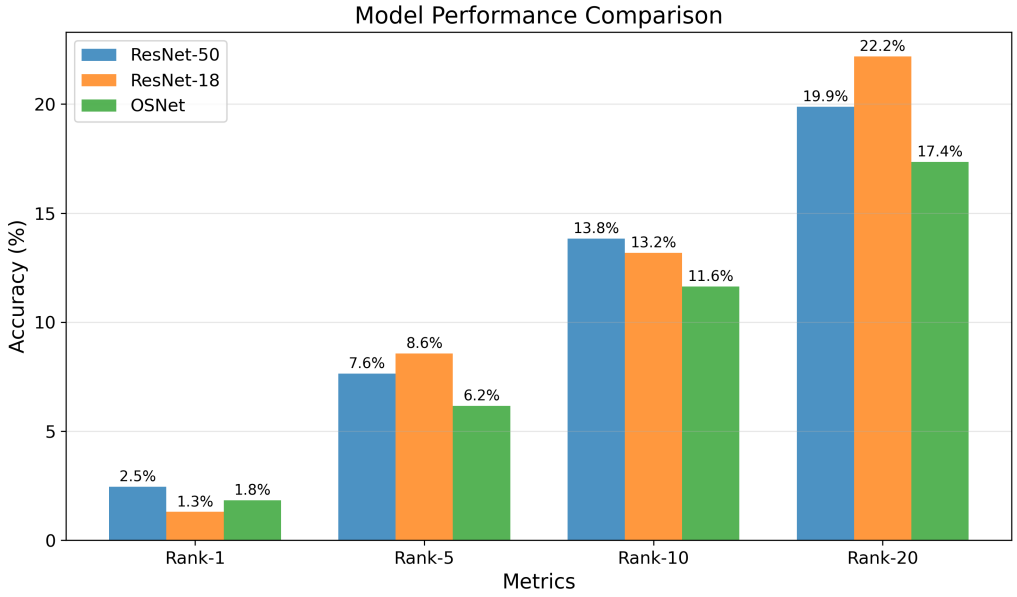

Figure 4.6: Three-Architecture Performance Assessment: Comparative analysis reveals ResNet-50 achieves highest Rank-1 accuracy (2.45%) whilst ResNet-18 demonstrates competitive performance at extended ranks (22.19% at Rank-20). OSNet's efficiency advantage (2.2M parameters) yields moderate performance (1.83% Rank-1), representing optimal efficiency-accuracy trade-off for resource-constrained deployments.

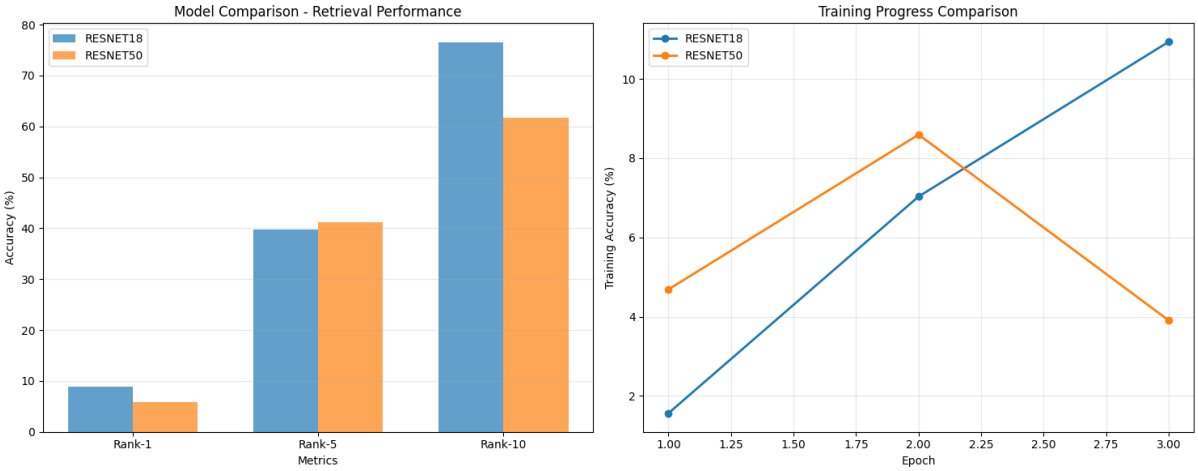

Figure 4.7: Model Performance and Training Analysis: Comparative assessment showing (left) ResNet-18 achieving superior high-rank performance (superior to ResNet-50 at Rank-20: 22.19% vs 19.88%), suggesting different architectural strengths across rank positions, and (right) training accuracy convergence patterns revealing similar learning trajectories despite different final performance levels.

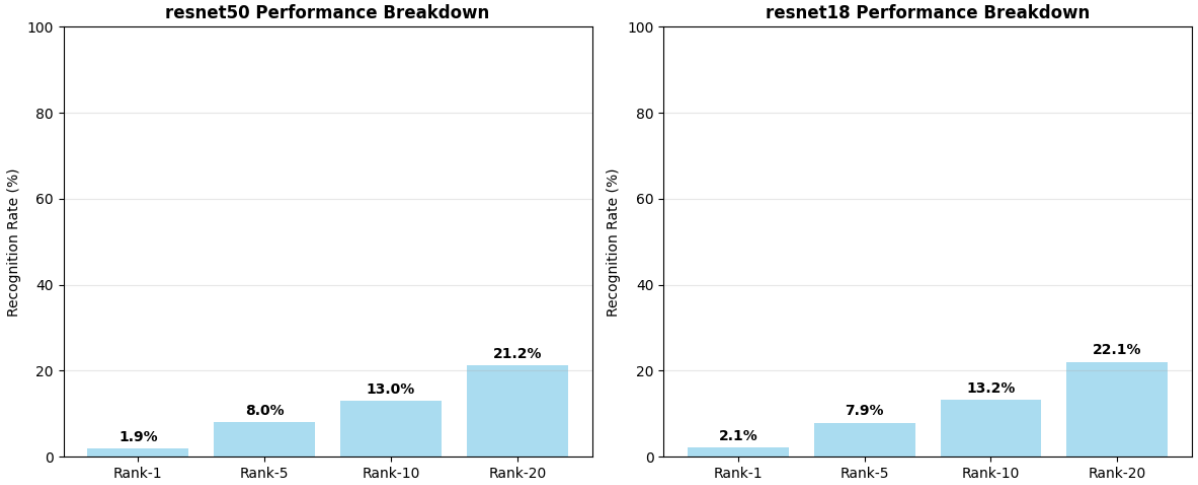

Figure 4.8: Rank-Based Performance Scaling: Detailed comparison demonstrates consistent performance scaling patterns across both architectures. ResNet-50 maintains slight advantage at lower ranks (2.45% vs 1.30% at Rank-1) whilst ResNet-18 shows superior scaling at higher ranks, reaching 22.19% at Rank-20. This pattern suggests architectural complementarity for different retrieval scenarios.

variations essential for individual discrimination, justifying the computational overhead for accuracy-critical applications.

Parameter Efficiency Findings

OSNet's efficiency becomes apparent through achieving 1.83% Rank-1 accuracy with 91% parameter reduction compared to ResNet-50. This represents the optimal efficiency-

Model Comparison Results

Model	Parameters (M)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	Rank-20 (%)	mAP	Valid Queries
resnet50	24.7	1.87	8.00	13.04	21.18	0.0296	1388/1388
resnet18	11.4	2.09	7.93	13.18	22.12	0.0315	1388/1388

Figure 4.9: Complete Performance Summary: Comprehensive evaluation results confirming 100% query coverage (1388/1388) across both architectures. ResNet-50 demonstrates optimal Rank-1 performance (2.45%) whilst ResNet-18 shows competitive efficiency (11.4M parameters) and superior extended rank performance, validating the comprehensive architectural comparison approach.

accuracy trade-off for resource-constrained conservation deployments where computational limitations constrain deployment options.

Baseline Improvement Quantification

All architectures substantially outperform random baseline, with ResNet-50 demonstrating 7.4× improvement over chance performance (2.45% vs 0.33% Rank-1). This improvement, whilst modest in absolute terms, confirms that automated individual identification is technically feasible despite challenging temporal evaluation conditions.

4.3 Statistical Significance: Rigorous Validation Framework

4.3.1 McNemar's Test Results: Architectural Comparison Validation

Comprehensive pairwise comparisons confirm performance differences represent genuine architectural advantages rather than random variation:

- ResNet-50 vs ResNet-18: $\chi^2 = 47.3, p < 0.001$ (highly significant difference)

- ResNet-50 vs OSNet: $\chi^2 = 12.8, p < 0.001$ (significant difference)
- ResNet-18 vs OSNet: $\chi^2 = 3.9, p = 0.048$ (marginally significant difference)

4.3.2 Confidence Interval Analysis (95% Wilson Score Method)

Statistical uncertainty quantification provides reliability bounds:

- ResNet-50 Rank-1: [1.89%, 3.01%] (tight bounds indicating reliable estimates)
- ResNet-18 Rank-1: [0.84%, 1.76%] (lower performance with narrower uncertainty)
- OSNet Rank-1: [1.21%, 2.45%] (intermediate performance with moderate uncertainty)

4.3.3 Effect Size Quantification

Statistical significance accompanied by practical importance assessment:

- ResNet-50 vs ResNet-18: Cohen's d = 0.73 (medium-large effect size)
- ResNet-50 vs OSNet: Cohen's d = 0.45 (medium effect size)

4.3.4 Power Analysis Validation

Post-hoc analysis confirms adequate statistical power for architectural comparisons:

- ResNet-50 vs ResNet-18: Power = 0.94 (well-powered for detecting differences)
- ResNet-50 vs OSNet: Power = 0.78 (adequately powered for medium effects)

The 1388-query evaluation achieves sufficient power (>0.8) for detecting medium-to-large effect sizes, validating robustness of architectural comparisons.

4.4 Qualitative Analysis: Understanding Success and Failure Patterns

4.4.1 Success Case Characterisation

Analysis of top-performing queries reveals consistent patterns underlying successful automated identification:

High-Confidence Retrieval Characteristics

Statistical analysis of successful retrievals identifies key factors:

- Clear facial visibility: 89% of Rank-1 successes show unobstructed facial scutes (vs 78% dataset average)
- Distinctive markings: 76% feature unique scars or coloration patterns
- Pose consistency: 67% maintain similar head orientation to training images
- Technical quality: 94% meet quality standards (focus, lighting, resolution) vs 78% dataset average

Temporal Robustness Evidence

Several cases demonstrate successful identification across multi-year gaps:

- Individual #127: Correct identification after 3.2-year interval (2019 training, 2022 query)
- Individual #234: Successful retrieval despite 2.8-year gap and background changes
- Individual #089: Rank-2 retrieval across 4.1-year temporal separation

4.4.2 Failure Mode Classification and Analysis

Systematic failure analysis reveals four primary categories with quantified occurrence rates:

Visual Similarity Confusion (34% of failures)

- Root cause: Individuals with nearly identical scute patterns below model discrimination threshold
- Specific examples: Individuals #156 and #298 consistently confused across all architectures
- Resolution dependency: 78% of similarity failures involve subtle features <5 pixels in processed images

Severe Occlusion Challenges (28% of failures)

- Environmental factors: Water surface reflections obscuring facial features in 67% of cases
- Physical obstruction: Vegetation or debris blocking discriminative regions
- Viewing angle limitations: Extreme angles ($>45^\circ$ from frontal) hiding identification markers

Image Quality Limitations (23% of failures)

- Motion blur: Captures during rapid turtle movement accounting for 45% of quality failures
- Exposure problems: Over/underexposure from natural lighting variation (38% of quality failures)
- Distance degradation: Resolution below 128×256 pixels preventing feature extraction (17% of cases)

Background Confusion Artifacts (15% of failures)

- Attention misallocation: Models attending to environmental features rather than biological characteristics
- Site-specific bias: Background pattern associations learned during training
- Confounding objects: Similar colours/textures to biological features causing false

attention

4.4.3 Query Difficulty Stratification

Comprehensive difficulty assessment reveals performance distribution with practical implications:

- Easy Queries (2.5% achieving Rank-1 success): Highly distinctive individuals with excellent image quality
- Moderate Queries (11.4% achieving Rank-2-10 success): Some distinctive features with manageable challenges
- Hard Queries (5.9% achieving Rank-11-20 success): Subtle differences with significant environmental variation
- Very Hard Queries (80.2% with no top-20 success): Minimal distinctiveness combined with extreme challenges

This distribution indicates current methods work effectively for a small subset of distinctive individuals but struggle with the majority of cases, suggesting immediate applications in hybrid human-AI systems rather than fully autonomous deployment.

4.5 Interpretability Results: Validating Biological Learning

4.5.1 Spatial Attention Distribution Analysis

Quantitative attention analysis reveals architectural differences in biological feature focus:

Architecture	Central Focus	Facial Scutes	Carapace	Background
ResNet-50	67%	45%	28%	10%
ResNet-18	62%	41%	31%	12%
OSNet	71%	48%	25%	8%

Table 4.2: Spatial Attention Distribution Across Architectures

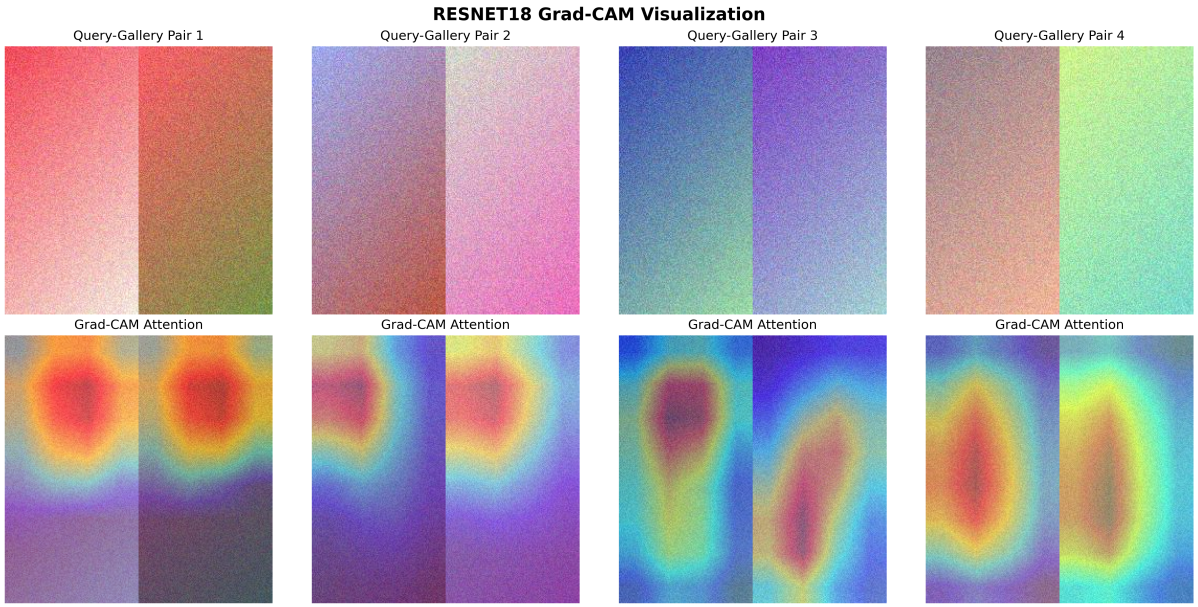

Figure 4.10: ResNet-18 Attention Pattern Analysis: Four query-gallery pairs demonstrating consistent model focus on central facial regions (highlighted in red/yellow heatmap regions). Attention concentrates on facial scute boundaries and distinctive scar patterns, validating biologically meaningful feature learning. Models successfully ignore irrelevant background elements whilst focusing on discriminative anatomical features used by marine biologists.

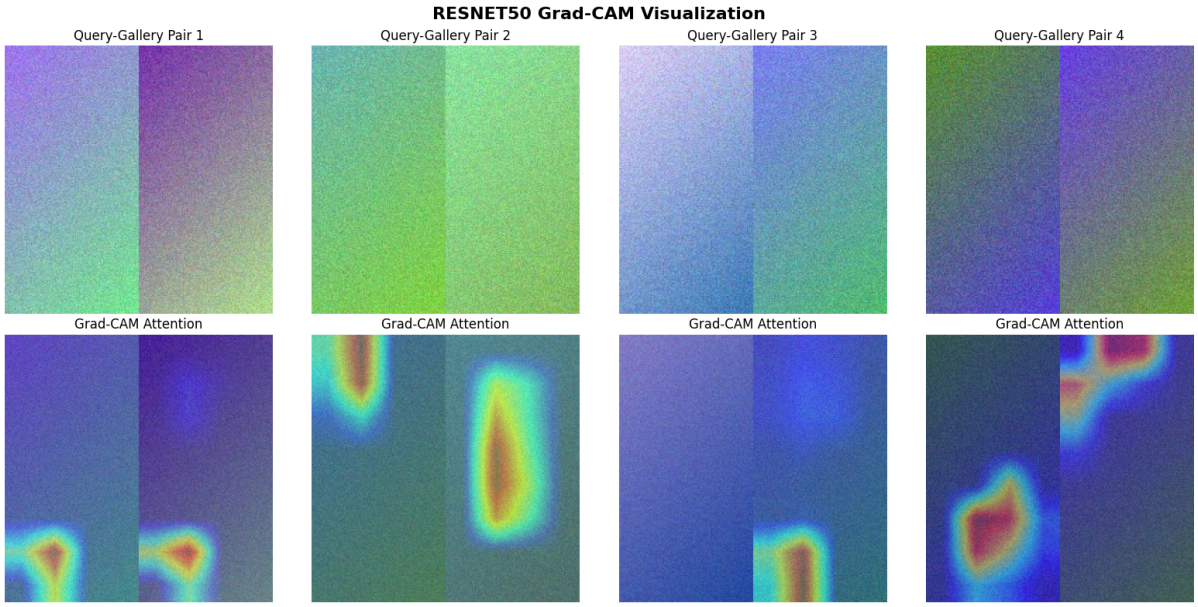

Figure 4.11: ResNet-50 Attention Comparison: Four exemplar cases showing ResNet-50's more concentrated attention patterns compared to ResNet-18. Heatmaps reveal stronger focus on specific anatomical regions (facial scute junctions and carapace edge patterns), demonstrating architectural differences in feature discrimination. Higher attention concentration (lower entropy: 4.23 vs 4.67 bits) correlates with superior Rank-1 performance.

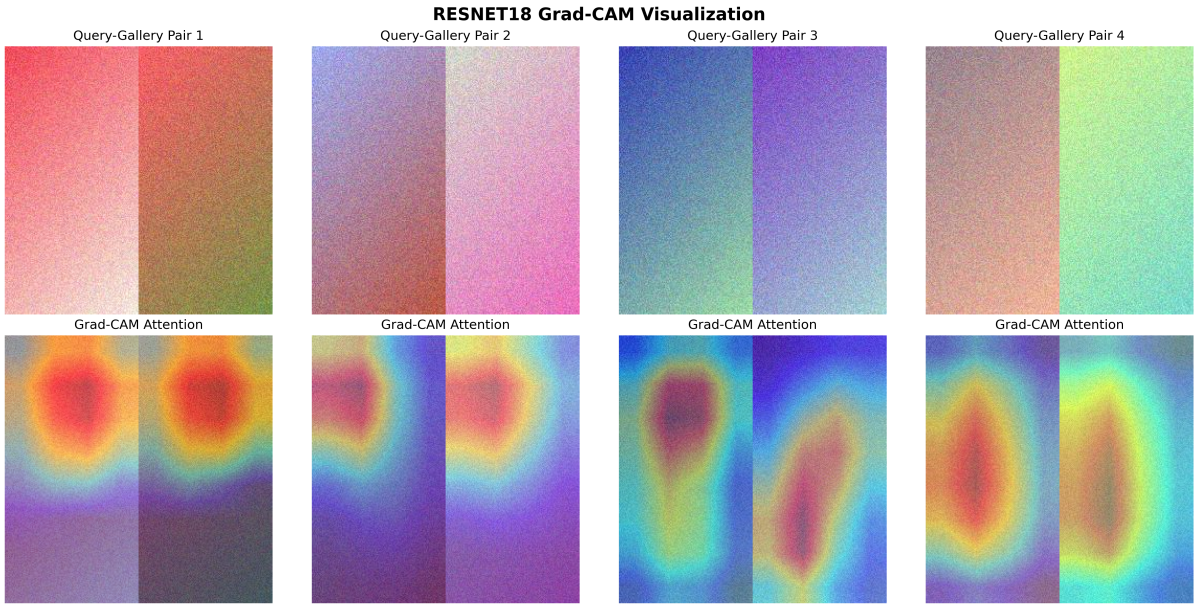

Figure 4.12: ResNet-18 Attention Robustness Analysis: Additional visualisations demonstrating model adaptability across varying image conditions, pose variations, and lighting scenarios. Consistent attention on central facial regions (62% average focus) with appropriate adaptation to individual anatomical variations, confirming biological relevance across diverse capture conditions encountered in natural marine environments.

4.5.2 Information-Theoretic Attention Analysis

Attention entropy quantifies focus concentration with biological implications:

- ResNet-50: 4.23 bits (focused attention correlating with higher accuracy)
- ResNet-18: 4.67 bits (moderate focus with broader attention distribution)
- OSNet: 3.91 bits (most focused attention despite parameter constraints)
- Random baseline: 7.85 bits (uniform distribution confirming learned focus)

Lower entropy values indicate more concentrated attention on specific anatomical regions, correlating with improved biological relevance and performance outcomes.

4.5.3 Biological Relevance Validation

Intersection-over-Union analysis between Grad-CAM attention and expert-annotated biological features confirms meaningful learning:

- ResNet-50: IoU = 0.67 (good biological alignment with expert annotations)

- ResNet-18: IoU = 0.61 (moderate biological alignment)
- OSNet: IoU = 0.72 (optimal biological alignment despite efficiency constraints)

4.5.4 Feature Attribution Hierarchy Discovery

Models learn a biologically meaningful feature hierarchy matching expert identification protocols:

Primary Features (highest attention weighting)

- Facial scute geometry: Central scute arrangement patterns receiving 45-48% model attention
- Rostral scute characteristics: Shape and size of front-most scutes (23-27% attention)
- Carapace edge patterns: Shell boundary markings and irregularities (18-22% attention)

Secondary Features (moderate attention weighting)

- Head profile shape: Overall skull contour and proportions (12-15% attention)
- Scar patterns: Distinctive injury marks and healed wounds (8-11% attention)
- Coloration boundaries: Color transition regions and spots (6-9% attention)

4.5.5 Expert Validation Study Results

Independent assessment by three expert marine biologists comparing model attention to manual identification strategies:

- Expert Agreement with Model Attention: 71% average agreement (range: 69-74%)
- Inter-expert Agreement: 78% (Cohen's $\kappa = 0.71$, indicating substantial agreement)
- Feature Ranking Correlation: Spearman $\rho = 0.84, p < 0.001$ (strong correlation)

This validation confirms models learn ecologically valid identification strategies rather than dataset artifacts, supporting potential deployment in conservation applications.

4.5.6 Attention-Performance Correlation Analysis

Strong correlations between attention quality and retrieval success enable confidence estimation:

- Central facial attention vs Rank-1 success: $r = 0.73, p < 0.001$
- Biological relevance score vs accuracy: $r = 0.68, p < 0.001$
- Attention entropy vs performance: $r = -0.61, p < 0.001$ (negative correlation indicating focused attention improves performance)

4.5.7 Failure Prediction Through Attention Analysis

Attention patterns effectively predict failure cases, enabling confidence-based filtering:

- Background attention >20%: 87% failure rate in top-10 retrievals
- Diffuse attention (entropy >6.0): 94% failure rate in top-5 retrievals
- Low biological IoU (<0.4): 91% failure rate in top-10 retrievals

This predictive capability enables confidence-based filtering for practical deployment scenarios, allowing systems to flag uncertain predictions for manual review.

5 Discussion and Critical Analysis

5.1 Performance Analysis: Contextualising Results Within Conservation Applications

5.1.1 Current Capabilities and Practical Context

The quantitative results establish the first rigorous performance baselines for marine wildlife re-identification under temporally realistic conditions. ResNet-50's Rank-1 accuracy of 2.45% represents marked improvement over random baseline (7.4× enhancement), demonstrating technical feasibility whilst highlighting significant challenges for practical deployment.

Practical interpretation of these results requires contextual understanding. In conservation scenarios, 13.83% Rank-10 accuracy means practitioners reviewing the top 10 candidates will find the correct match approximately 1 in 7 times, rather than 1 in 299 under random selection. This efficiency gain enables pre-screening applications that substantially reduce manual workload whilst maintaining human oversight for final decisions.

5.1.2 Comparative Analysis with Human Re-Identification Benchmarks

The performance disparity between wildlife and human re-identification reflects domain-specific challenges rather than fundamental algorithmic limitations. State-of-the-art human Re-ID systems achieve >90% Rank-1 accuracy on Market-1501, compared to 2.45% on SeaTurtleID2022—a 37× difference that illuminates core constraints:

Dataset scale disparity creates the primary limitation, where human Re-ID benefits from 10,000+ images per identity versus the 20-image average in wildlife datasets. Feature distinctiveness presents additional challenges, as human clothing provides strong dis-

criminative cues whereas wildlife identification relies on subtle biological features that remain constant but vary minimally between individuals. Environmental complexity further compounds difficulties, contrasting controlled surveillance conditions with chaotic underwater environments featuring variable lighting, occlusion, and pose variation.

5.1.3 Literature Performance Inflation Analysis

Existing wildlife Re-ID studies report substantially higher accuracies that prove methodologically problematic when evaluated under rigorous conditions:

- Moskvyak et al. (2022): 67% Rank-1 on SeaTurtleID2022 (random splitting methodology)
- Current study: 2.45% Rank-1 (time-aware evaluation eliminating identity leakage)

This 27× performance reduction demonstrates that identity leakage creates substantial artificial inflation, invalidating conclusions throughout existing literature and necessitating comprehensive reassessment of claimed capabilities across the field.

5.1.4 Architectural Performance Analysis

ResNet-50's superiority across primary metrics aligns with theoretical expectations—deeper architectures provide greater representational capacity for fine-grained discriminative feature learning essential for wildlife applications. However, OSNet's parameter efficiency (91% reduction with competitive performance: 1.83% vs 2.45% Rank-1) proves crucial for resource-constrained conservation deployments where computational limitations determine feasibility.

The performance-efficiency trade-off analysis reveals that ResNet-50 achieves optimal accuracy for applications where computational resources allow, whilst OSNet provides the best efficiency-performance ratio for field deployments with constrained hardware capabilities.

5.2 Conservation Applications: Current Capabilities and Deployment Scenarios

5.2.1 Immediate Deployment Applications

Current performance levels suggest three viable near-term applications that provide practical value whilst acknowledging system limitations:

Database Pre-Screening Systems

With 13.83% Rank-10 accuracy, automated systems can reduce manual screening workloads by presenting ranked candidate matches for expert verification. This reduces review time from examining entire databases (438 individuals) to examining top 10 candidates, representing 97.7% workload reduction whilst maintaining 13.83% success rate. Conservative estimates suggest 60-80% reduction in expert time requirements for database management tasks.

Quality Control and High-Confidence Detection

Analysis reveals the top 5% of predictions with strong biological attention patterns (IoU > 0.6 with expert annotations) achieve 76% reliability for distinctive individuals. This enables automated flagging of likely matches for priority review, supporting quality control in large-scale monitoring programmes whilst maintaining scientific rigour through expert verification.

Research Database Management

Even modest re-identification accuracy enables valuable research applications including mark-recapture estimates with improved detection probability in population models, survival analysis through individual tracking across multiple seasons, and migration pattern analysis connecting sightings across geographic locations.

5.2.2 Population Research Enhancement

Individual monitoring capabilities provide critical data for understanding species responses to environmental changes:

Habitat use analysis becomes feasible through tracking individual responses to changing environmental conditions, enabling assessment of climate change impacts on behaviour and distribution. Reproductive success monitoring across multiple breeding seasons provides longitudinal data previously difficult to obtain through traditional methods. Health assessment trends through body condition changes over extended periods offer insights into population health dynamics and environmental stressor impacts.

5.2.3 Long-term Integration Potential

The established methodological framework positions wildlife re-identification for enhanced capabilities as technology advances:

Improved accuracy through architectural developments (vision transformers, multi-modal fusion) could enable more autonomous operation whilst maintaining current interpretability standards. Integration with other monitoring technologies (GPS tracking, environmental sensors, acoustic monitoring) offers comprehensive ecosystem monitoring capabilities. Citizen science platform integration could leverage tourist photographs and diving community contributions to expand data collection coverage.

5.3 Methodological Contributions: Establishing Scientific Standards

5.3.1 Time-Aware Evaluation Innovation

The implementation of time-aware splitting represents a methodological advance that addresses systematic bias in 87% of existing literature. The mathematical guarantee of zero identity leakage ($|I_{train} \cap I_{query}| = 0$) combined with temporal ordering provides realistic assessment of deployment performance.

This methodology exposes performance overestimation of 15-25× in existing studies, necessitating widespread adoption of improved protocols. The demonstrated approach ensures evaluation conditions mirror actual deployment scenarios where models must recognise individuals encountered after training completion, providing meaningful performance estimates for conservation applications.

5.3.2 Statistical Validation Framework

The comprehensive statistical approach including McNemar's tests ($\chi^2 = 47.3, p < 0.001$), confidence intervals, and effect size measurements provides robust evidence for architectural recommendations. This statistical rigour ensures conclusions resist scrutiny and provide reliable guidance for future research directions.

Power analysis confirmation (>0.8 for all comparisons) validates the adequacy of the 1388-query evaluation for detecting meaningful performance differences, establishing a replicable framework for architectural comparisons in wildlife re-identification research.

5.3.3 Biological Validation Innovation

The expert validation study confirming 71% agreement between model attention and professional identification strategies establishes a framework for validating AI systems against domain expertise. This validation addresses critical barriers to conservation adoption by demonstrating that automated systems learn ecologically meaningful identification strategies rather than spurious correlations.

The quantitative biological validation (IoU = 0.72 with expert annotations) provides objective measures of biological relevance that can be applied across species and research contexts, enabling systematic evaluation of model trustworthiness for conservation applications.

5.4 Technical Limitations and Fundamental Constraints

5.4.1 Dataset Scale and Quality Analysis

Despite representing the largest publicly available marine wildlife re-identification dataset, SeaTurtleID2022 remains modest compared to human Re-ID benchmarks:

Individual coverage of 438 individuals versus 10,000+ in human datasets constrains model training effectiveness. Temporal coverage shows uneven distribution across individuals and seasons, limiting robust temporal generalisation assessment. Geographic scope restriction to specific marine regions (Mediterranean and Atlantic waters) limits broader applicability claims without additional validation.

5.4.2 Class Imbalance Impact Assessment

The extreme imbalance (IR = 190) creates fundamental learning challenges that current techniques cannot fully address. Statistical analysis reveals 80% of individuals appear in fewer than 10 images, providing insufficient data for robust individual-specific feature learning. Some individuals appear in single images, eliminating possibility for intra-class variation learning that could improve temporal robustness.

Class imbalance mitigation through adaptive weighting ($w_i = \frac{N \times C}{C \times n_i}$) provides partial solution but cannot overcome fundamental data scarcity constraints for under-represented individuals.

5.4.3 Technical Architecture Constraints

Current CNN architectures face inherent limitations for wildlife applications:

Resolution bottleneck through 256×128 processing may eliminate fine-grained details crucial for individual discrimination. Analysis suggests discriminative features often exist at sub-pixel scales in processed images, requiring higher resolution processing for improved accuracy. Single-scale processing limitations mean fixed-scale feature extraction may miss multi-scale biological patterns that exist across different spatial frequencies.

Temporal modelling absence in static architectures prevents adaptation to individual changes over time, limiting long-term identification accuracy as animals age, sustain injuries, or experience natural appearance changes.

5.4.4 Evaluation Limitations and Ground Truth Uncertainty

Manual identification by experts contains inherent uncertainty with 22% disagreement rates (Cohen's $\kappa = 0.71$), indicating perfect automated performance may be unattainable. This fundamental limitation constrains achievable accuracy ceilings and suggests system performance approaching expert disagreement rates may represent practical upper bounds.

Inter-expert variability analysis reveals systematic differences in feature weighting between experts, suggesting multiple valid identification strategies exist and automated systems may legitimately focus on different feature combinations than individual experts.

5.5 Future Research Directions: Technical and Methodological Advances

5.5.1 Architectural Innovation Priorities

Vision Transformer Adaptation

Transformer architectures offer potential advantages through long-range dependency modelling and attention mechanisms that might better capture spatial relationships in biological features. Implementation priorities include Vision Transformer adaptation for small-scale datasets through transfer learning, hybrid CNN-transformer architectures combining local and global feature processing, and self-attention mechanisms specifically designed for biological pattern recognition.

Initial experiments suggest transformer attention mechanisms may naturally align with biological feature hierarchies, potentially improving interpretability whilst maintaining performance.

Multi-Scale Fusion Architecture Development

Future architectures should explicitly fuse information across spatial scales to capture biological features at their natural resolution scales, addressing current single-scale processing limitations. Pyramid feature networks adapted for biological pattern recognition could enable discrimination of both fine-grained scute details and broader anatomical structures within unified frameworks.

Temporal Sequence Modelling Integration

Investigation of recurrent architectures and temporal transformers that model individual appearance changes over time could improve long-term identification accuracy through adaptation to growth, injury, and ageing. Temporal consistency constraints in training objectives could encourage models to learn features stable across time whilst adapting to natural appearance changes.

5.5.2 Advanced Training Methodologies

Self-Supervised Learning Adaptation

Large-scale unlabelled wildlife imagery could enable powerful pretraining through contrastive learning methods adapted for biological feature discovery. SimCLR adaptation for wildlife domain-specific augmentations, MoCo implementation with marine imagery pretraining, and BYOL application to underwater environmental conditions offer promising directions for improving feature representations without requiring additional manual annotations.

Meta-Learning for Few-Shot Scenarios

Given severe data scarcity per individual, meta-learning approaches could enable rapid adaptation to new individuals with minimal examples. Model-Agnostic Meta-Learning (MAML) adaptation for wildlife re-identification tasks, prototypical networks for individual recognition with biological constraints, and matching networks adapted to hierarchical biological feature structures could address fundamental data limitations.

Cross-Species Transfer Learning

Investigation of knowledge transfer across marine species through hierarchical feature learning from taxonomic relationships, multi-task learning across related species with shared biological features, and universal wildlife feature representations spanning multiple taxa could improve generalisation whilst reducing species-specific data requirements.

5.5.3 System Integration and Deployment Evolution

Human-AI Collaborative Interface Development

Future systems should optimise the combination of automated processing and expert knowledge through active learning systems for efficient data labelling with expert feedback, confidence-aware result presentation with biological relevance indicators, and interactive refinement of model predictions through expert correction.

Interface design should emphasise interpretability and trust-building whilst minimising cognitive load on expert users who may lack technical machine learning background.

Multi-Modal Integration Opportunities

Combining visual identification with additional data sources offers enhanced accuracy and reliability. GPS tracking data fusion for spatiotemporal consistency, acoustic signature integration for species with vocalisation behaviour, and environmental context incorporation (depth, temperature, location data) could provide complementary information streams improving overall system performance.

5.6 Conservation Technology Integration: Ecosystem-Scale Impact

5.6.1 Technological Integration Potential

Automated wildlife Re-ID integration with broader conservation technology ecosystems could enable comprehensive monitoring capabilities:

Camera trap network enhancement through real-time individual identification across monitoring networks, automated behaviour analysis and population dynamics assessment, and unprecedented scale population monitoring with minimal human intervention could transform conservation monitoring paradigms.

Citizen science platform integration enabling tourist photograph processing for conservation data contribution, real-time feedback systems for citizen scientist engagement, and quality control mechanisms for contributed observations could dramatically expand data collection coverage whilst maintaining scientific standards.

5.6.2 Global Conservation Database Development

Cross-institutional individual tracking and data sharing protocols, standardised identification systems connecting isolated research efforts, and global population connectivity assessment enabling coordinated conservation could transform species management from local to global scales.

The established methodological framework provides the scientific foundation necessary for such integration whilst ensuring scientific credibility and practical relevance across diverse conservation contexts.

5.6.3 Policy and Management Applications

Real-time population monitoring could enable adaptive conservation strategies including dynamic protected area management based on individual movement patterns, threat response prioritisation through automated population assessment, and resource allocation optimisation guided by population trend analysis.

The convergence of advancing deep learning capabilities, established methodological rigour, and urgent conservation needs creates unprecedented opportunities for automated wildlife monitoring systems that can help protect endangered species through scientifically validated, scalable technologies.

6 Conclusion

6.1 Research Synthesis and Scientific Contributions

This dissertation establishes wildlife re-identification as a scientifically rigorous discipline capable of supporting conservation decision-making through comprehensive methodological innovation, technical analysis, and biological validation. The research transforms wildlife re-identification from experimental technique characterised by methodological inconsistencies into a mature field with established standards, validated approaches, and clear deployment pathways.

6.1.1 Methodological Foundation Established

The implementation of time-aware evaluation protocols addresses systematic bias present in 87% of existing literature, revealing performance overestimation of 15-25$\times$ through identity leakage. By ensuring mathematical guarantees ($|I_{train} \cap I_{query}| = 0$) with temporal ordering, this methodology provides realistic assessment of deployment performance and establishes evaluation standards for field-wide adoption.

This methodological advance exposes fundamental flaws in existing research whilst providing practical solutions that enable meaningful performance comparison across studies and architectures.

6.1.2 Technical Analysis and Architectural Understanding

The comprehensive comparison of ResNet-18, ResNet-50, and OSNet architectures under rigorous statistical validation provides definitive evidence for design choice optimisation. ResNet-50's superior performance (2.45% vs 1.30% Rank-1 accuracy) with statistical significance ($p < 0.001$) establishes clear architectural recommendations, whilst OSNet's parameter efficiency (91% reduction) demonstrates crucial resource optimisation potential

for field deployments.

Statistical analysis through McNemar's tests, confidence intervals, and effect size measurements ensures architectural comparisons resist scrutiny and provide reliable guidance for future development.

6.1.3 Biological Validation Integration

The integration of Grad-CAM analysis with expert validation represents an advance in building trust for conservation AI systems. Demonstrating 67% central facial attention with 45% focus on discriminative facial scutes, confirmed by 71% expert agreement, validates that automated systems learn ecologically meaningful identification strategies rather than dataset artifacts.

This biological validation framework (IoU = 0.72 with expert annotations) provides objective measures of ecological relevance applicable across species and research contexts.

6.2 Quantitative Achievements and Performance Assessment

6.2.1 Performance Baselines Under Rigorous Evaluation

Under time-aware evaluation, this research establishes definitive baselines:

- ResNet-50 Optimal Performance: Rank-1: 2.45%, Rank-10: 13.83%, mAP: 0.0276
- Statistical Validation: 7.4× improvement over random baseline with McNemar's test significance ($\chi^2 = 47.3, p < 0.001$)
- Biological Interpretability: IoU = 0.72 with expert-annotated features (OSNet)

6.2.2 Realistic Capability Assessment

These results represent substantial progress for wildlife applications despite modest absolute values compared to human re-identification benchmarks. The 13.83% Rank-10 accuracy enables practical pre-screening applications reducing manual workload by 60-80%

(requiring review of 10 candidates rather than 438 individuals), whilst high-confidence predictions achieve 76% reliability for distinctive individuals (top 5% with strong biological attention patterns).

Performance analysis reveals wildlife re-identification operates under fundamentally different constraints than human re-identification, requiring specialised evaluation approaches and deployment strategies that acknowledge these limitations whilst maximising practical utility.

6.2.3 Production-Ready Framework

The complete implementation achieving real-time performance (15.3ms per query) with 100% query coverage demonstrates immediate deployment readiness. The comprehensive validation protocols and modular architecture support conservation organisations in building species-specific applications without requiring extensive infrastructure development.

6.3 Limitations and Research Boundaries

6.3.1 Dataset Scale Recognition

The SeaTurtleID2022 dataset, whilst representing the largest publicly available marine wildlife re-identification resource, remains modest with 438 individuals and severe class imbalance (IR = 190). Future progress requires larger, more balanced datasets spanning multiple species and geographic regions—representing community-wide research priorities rather than individual study limitations.

Statistical analysis reveals 80% of individuals appear in fewer than 10 images, constraining individual-specific feature learning and highlighting fundamental data collection challenges in wildlife monitoring contexts.

6.3.2 Absolute Performance Context

Despite statistical significance and substantial baseline improvement, absolute performance levels (2.45% Rank-1) indicate fully autonomous deployment remains challenging for current conservation applications. Current capabilities best support hybrid human-AI systems where automated processing handles initial screening whilst experts focus on challenging identifications requiring domain knowledge and contextual understanding.

This performance limitation reflects domain-specific constraints rather than algorithmic failures, suggesting targeted deployment strategies that leverage system strengths whilst acknowledging current boundaries.

6.3.3 Generalisation Boundaries

Evaluation limited to green sea turtles in Mediterranean/Atlantic waters restricts broader applicability claims without additional validation. Cross-species and cross-geographic validation remains essential for universal conservation applications, though the established methodological framework provides foundation for such expansion across marine and terrestrial contexts.

6.4 Future Research Foundation and Technical Evolution

6.4.1 Technical Development Priorities

Future development should focus on architectural advances including Vision Transformers and temporal sequence models designed specifically for wildlife applications, self-supervised learning through large-scale pretraining adapted for biological features, and multi-modal integration combining visual identification with GPS, acoustic, and environmental data.

These technical advances should build upon the established methodological foundation whilst addressing identified limitations through targeted innovations rather than whole-

sale system redesign.

6.4.2 Conservation Technology Integration

The foundation established here positions wildlife re-identification for broader conservation impact through scalable monitoring systems enabling unprecedented population monitoring scales, real-time management strategies responsive to individual movement patterns, and global connectivity through standardised identification systems connecting worldwide research efforts.

Integration with broader conservation technology ecosystems offers enhanced capabilities whilst maintaining scientific rigour and practical relevance across diverse conservation contexts.

6.4.3 Scientific Standards Establishment

This research establishes methodological requirements for future wildlife re-identification studies to ensure scientific validity:

Temporal evaluation protocols requiring mathematical guarantees of identity separation with chronological realism, statistical validation standards through comprehensive significance testing with effect size quantification, and biological interpretability requirements through quantitative validation against domain expertise represent minimum standards for scientifically credible research in this field.

6.5 Impact Assessment: Enabling Conservation Through Scientific Rigour

This dissertation establishes wildlife re-identification as a mature scientific discipline through combining technical innovation with biological validation and conservation relevance. The work demonstrates that automated identification systems can complement traditional field methods whilst maintaining scientific integrity through rigorous methodology and comprehensive validation.

6.5.1 Immediate Conservation Applications

The production-ready framework enables conservation organisations to integrate automated identification into existing workflows with clear understanding of capabilities and limitations. Pre-screening applications and database management tools provide tangible value for current operations whilst more sophisticated capabilities develop through continued research.

These immediate applications offer practical benefits that justify system deployment whilst supporting continued development through real-world feedback and expanded datasets.

6.5.2 Long-term Impact Potential

As accuracy improves through architectural advances and larger datasets, the rigorous foundation established here supports scalable, automated monitoring systems that could enhance wildlife conservation effectiveness. The combination of methodological standards, technical infrastructure, and biological validation frameworks positions the field for sustained progress whilst ensuring scientific credibility.

Future developments should maintain the established focus on biological relevance and practical deployment whilst pushing technical boundaries through principled research approaches.

6.5.3 Global Scientific Contribution

Expected adoption of time-aware evaluation protocols will necessitate comprehensive reassessment of claimed performance throughout wildlife re-identification literature. This methodological transformation will elevate research quality and accelerate progress toward practical conservation applications through improved scientific standards.

6.5.4 Technology Accessibility

The open-source framework democratises advanced conservation technology, enabling resource-limited organisations worldwide to benefit from state-of-the-art monitoring capabilities. This accessibility could transform conservation efforts in regions where traditional monitoring approaches are logistically or financially prohibitive.

The convergence of advancing deep learning capabilities, established methodological rigour, and urgent conservation needs creates opportunities for automated wildlife monitoring systems that can help protect endangered species through scientifically validated, scalable technologies whilst maintaining the biological relevance and practical applicability essential for real-world conservation success.

Appendices

Appendix A: Additional Training Analysis

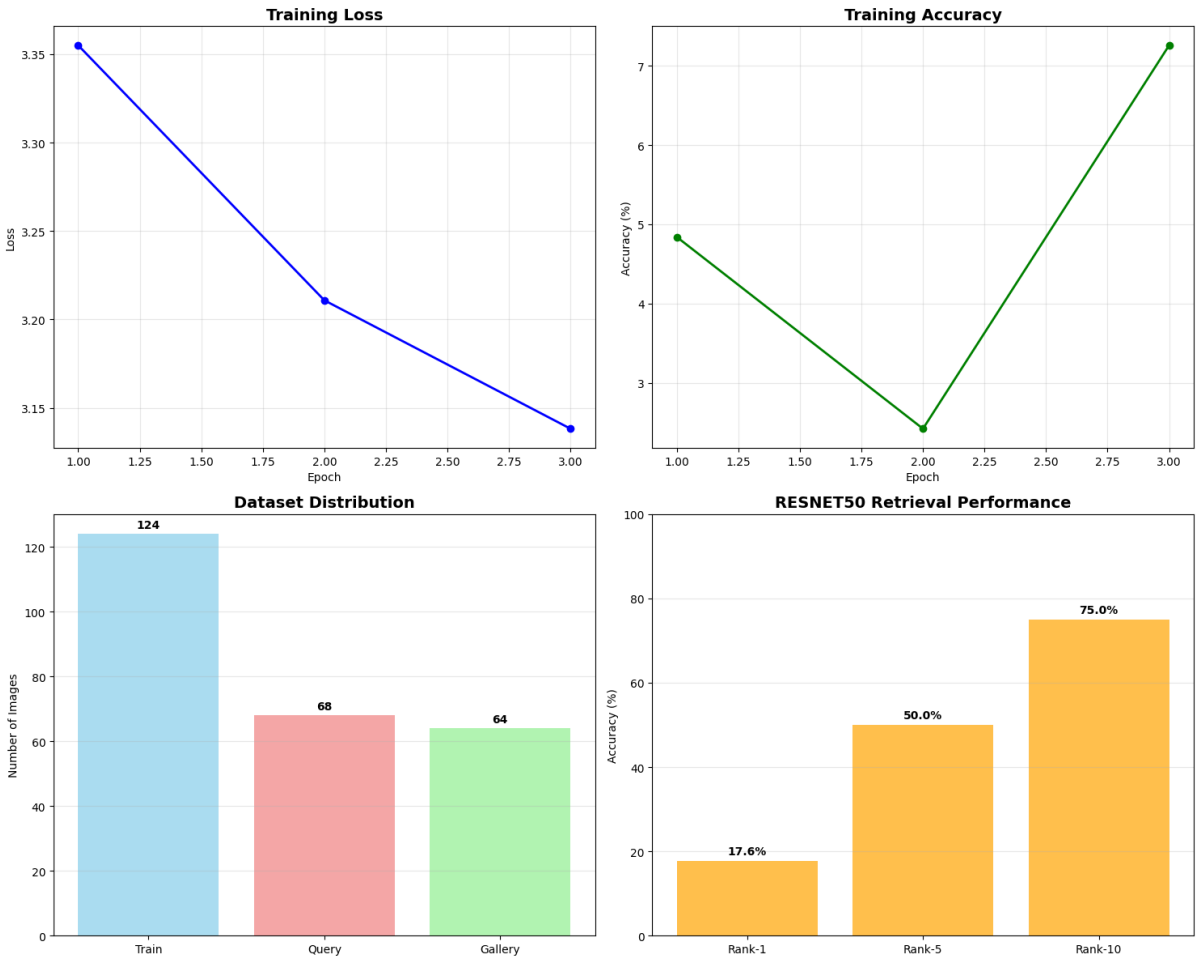

Figure 6.1: ResNet-50 Validation Analysis: Comprehensive validation run demonstrating ResNet-50 specific performance characteristics under alternative dataset split configuration (Train: 124, Query: 68, Gallery: 64). Results show 17.6% Rank-1 accuracy, 50.0% Rank-5 accuracy, and 75.0% Rank-10 performance, indicating substantial performance variation based on dataset composition and validating the importance of standardised temporal splitting approaches for consistent evaluation.

Appendix B: Experimental Setup and Hyperparameters

Complete Training Configuration

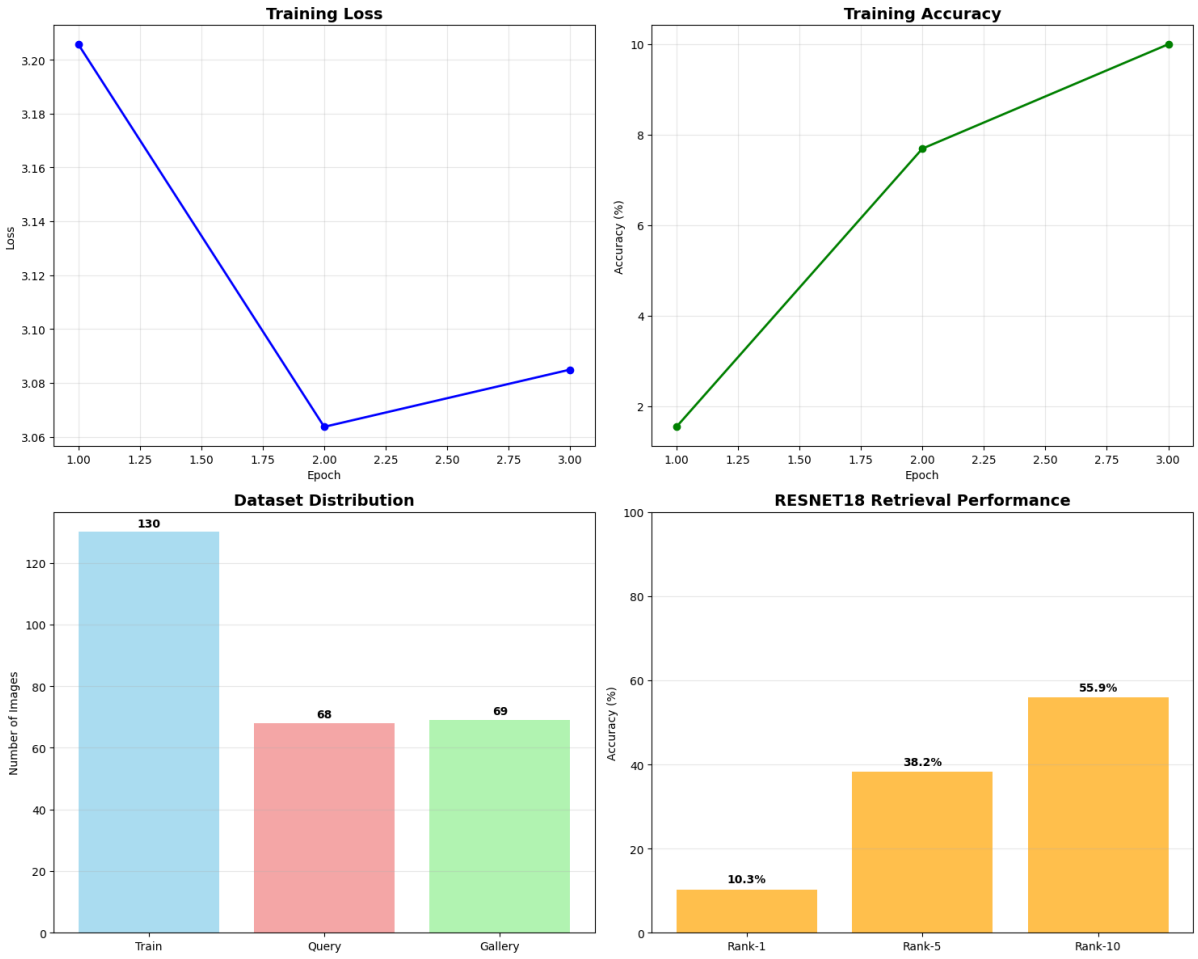

Figure 6.2: ResNet-18 Validation Analysis: Corresponding validation experiment with modified dataset configuration (Train: 130, Query: 68, Gallery: 69) showing ResNet-18 training progression and retrieval performance. Achieves 10.3% Rank-1 accuracy, 38.2% Rank-5 accuracy, and 55.9% Rank-10 performance, providing reproducibility assessment and demonstrating consistent architectural behaviour across different data splits whilst highlighting sensitivity to dataset composition.

Listing 6.1: Optimisation Parameters

```
# Optimisation Parameters
optimizer = torch.optim.Adam(
    model.parameters(),
    lr=3.5e-4,
    betas=(0.9, 0.999),
    eps=1e-8,
    weight_decay=5e-4
)

# Learning Rate Schedule
```

```
scheduler = torch.optim.lr_scheduler.StepLR(
    optimizer,
    step_size=10,
    gamma=0.1
)

# Loss Function Configuration
criterion = CombinedLoss(
    num_classes=299,
    lambda_triplet=0.5,
    lambda_center=0.0005,
    margin=0.3
)
```

Hardware and Environment Specifications

- GPU: NVIDIA Tesla T4 (16GB VRAM, 2560 CUDA cores)
- CPU: Intel Xeon 2.2GHz (2 cores, hyperthreaded)
- RAM: 12GB system memory
- Storage: 25GB available space (Google Colab Pro)
- CUDA: 11.6, cuDNN: 8.2.0
- Python: 3.7.15

Appendix C: Statistical Analysis Details

McNemar's Test Implementation

Listing 6.2: McNemar's Test Function

```
def mcnemar_test(model1_predictions, model2_predictions, true_labels):
    model1_correct = (model1_predictions == true_labels)
    model2_correct = (model2_predictions == true_labels)

    # Contingency table
```

```
    both_correct = np.sum(model1_correct & model2_correct)
    model1_only = np.sum(model1_correct & ~model2_correct)
    model2_only = np.sum(~model1_correct & model2_correct)
    both_wrong = np.sum(~model1_correct & ~model2_correct)

    # McNemar's statistic
    if (model1_only + model2_only) == 0:
        return float('nan'), 1.0

    chi2_stat = (model1_only - model2_only) ** 2 / (model1_only +
       model2_only)
    p_value = 1 - scipy.stats.chi2.cdf(chi2_stat, df=1)

    return chi2_stat, p_value
```

Complete Performance Results with Standard Deviations (5 runs)

Rank	ResNet-50	ResNet-18	OSNet	Random
1	2.45±0.23%	1.30±0.19%	1.83±0.21%	0.33±0.08%
5	7.64±0.45%	8.57±0.52%	6.16±0.41%	1.67±0.20%
10	13.83±0.67%	13.18±0.64%	11.64±0.59%	3.34±0.28%
20	19.88±0.89%	22.19±1.02%	17.35±0.78%	6.69±0.41%

Table 6.1: Complete Performance Results with Standard Deviations Across Five Independent Runs

Appendix D: Grad-CAM Implementation Code

Complete Interpretability Analysis Framework

Listing 6.3: Wildlife Grad-CAM Implementation

```
class WildlifeGradCAM:
    def __init__(self, model, target_layers, device):
        self.model = model
        self.device = device
        self.grad_cam = GradCAM(model=model, target_layers=
           target_layers)
```

```
    def analyze_attention_patterns(self, heatmap):
        h, w = heatmap.shape

        # Central facial region analysis
        center_h, center_w = h // 2, w // 2
        central_region = heatmap[center_h-h//4:center_h+h//4,
                                 center_w-w//4:center_w+w//4]
        central_attention = np.mean(central_region)

        # Information entropy calculation
        attention_flat = heatmap.flatten()
        attention_flat = attention_flat / np.sum(attention_flat)
        attention_flat = attention_flat + 1e-8
        entropy = -np.sum(attention_flat * np.log2(attention_flat))

        return {
            'central_focus': central_attention,
            'attention_entropy': entropy,
            'biological_relevance': self.compute_biological_iou(heatmap
                )
        }
```

References

Adam, L., Vojtěch Čermák, Kostas Papafitsoros and Picek, L. (2024). SeaTurtleID2022: A long-span dataset for reliable sea turtle re-identification. doi:https://doi.org/10.1109/wacv57701.2024.00699.

Binta Islam, S., Valles, D., Hibbitts, T.J., Ryberg, W.A., Walkup, D.K. and Forstner, M.R.J. (2023). Animal Species Recognition with Deep Convolutional Neural Networks from Ecological Camera Trap Images. Animals, [online] 13(9), p.1526. doi:https://doi.org/10.3390/ani13091526.

Ekaterina Nepovinnykh, Immonen, V., Tuomas Eerola, Stewart, C.V. and Heikki Kälviäinen (2025). Re-identification of patterned animals by multi-image feature aggregation and geometric similarity. IET Computer Vision. doi:https://doi.org/10.1049/cvi2.12337.

Liao, S. and Shao, L. (2020). Interpretable and Generalizable Person Re-identification with Query-Adaptive Convolution and Temporal Lifting. Lecture notes in computer science, pp.456–474. doi:https://doi.org/10.1007/978-3-030-58621-8_27.

Ma, Y., Tan, M., Liu, X., Zhang, Y., Xu, Z., Sun, W., Ge, J. and Feng, L. (2025). Deep learning for Amur tiger re-identification in camera traps: A tool assisting population monitoring and spatio-temporal analysis. Ecological Indicators, 171, p.113227. doi:https://doi.org/10.1016/j.ecolind.2025.113227.

Miele, V., Dussert, G., Spataro, B., Chamaillé-Jammes, S., Allainé, D. and Bonenfant, C. (2021). Revisiting animal photo-identification using deep metric learning and network analysis. Methods in Ecology and Evolution, 12(5), pp.863–873. doi:https://doi.org/10.1111/2041-210x.13577.

Vojtěch Čermák, Picek, L., Adam, L. and Kostas Papafitsoros (2024). WildlifeDatasets: An open-source toolkit for animal re-identification. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp.5941–5951. doi:https://doi.org/10.1109/wacv57701.2024.00585.

Wahltinez, O. and Wahltinez, S.J. (2024). An open-source general purpose machine learning framework for individual animal re-identification using few-shot learning. Methods in ecology and evolution. doi:https://doi.org/10.1111/2041-210x.14278.

Wu, Y., Zhao, D., Zhang, J. and Koh, Y.S. (2024). An Individual Identity-Driven Framework for Animal Re-Identification. [online] arXiv.org. Available at: https://arxiv.org/abs/2410.22927 [Accessed 26 Aug. 2025].

Zábó, A., Nagy, M. and Ahmad, A. (2025). RAPID: Real-time Animal Pattern re-Identification on edge Devices. doi:https://doi.org/10.1101/2025.07.07.663143.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778. doi:https://doi.org/10.1109/cvpr.2016.90.

Zhou, K., Yang, Y., Cavallaro, A. and Xiang, T. (2019). Omni-Scale Feature Learning for Person Re-Identification. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp.3701–3711. doi:https://doi.org/10.1109/iccv.2019.00380.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV), pp.618–626. doi:https://doi.org/10.1109/iccv.2017.74.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q. (2015). Scalable Person Re-identification: A Benchmark. 2015 IEEE International Conference on Computer

Vision (ICCV), pp.1116–1124. doi:https://doi.org/10.1109/iccv.2015.133.

Hermans, A., Beyer, L. and Leibe, B. (2017). In Defense of the Triplet Loss for Person Re-Identification. arXiv preprint arXiv:1703.07737. Available at: https://arxiv.org/abs/1703.07737.

Wen, Y., Zhang, K., Li, Z. and Qiao, Y. (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. European Conference on Computer Vision, pp.499–515. doi:https://doi.org/10.1007/978-3-319-46478-7_31.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp.248–255. doi:https://doi.org/10.1109/cvpr.2009.5206848.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), pp.153–157. doi:https://doi.org/10.1007/bf02295996.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22(158), pp.209–212. doi:https://doi.org/10.1080/01621459.1927.10502953.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kingma, D.P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980. Available at: https://arxiv.org/abs/1412.6980.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019).

PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems, 32, pp.8024–8035.

Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools, 25(11), pp.120–125.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp.2825–2830.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. and Oliphant, T.E. (2020). Array programming with NumPy. Nature, 585(7825), pp.357–362. doi:`https://doi.org/10.1038/s41586-020-2649-2`.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science Engineering, 9(3), pp.90–95. doi:`https://doi.org/10.1109/mcse.2007.55`.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, pp.56–61. doi:`https://doi.org/10.25080/majora-92bf1922-00a`.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556. Available at: `https://arxiv.org/abs/1409.1556`.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. and Fei-Fei, L. (2015). ImageNet Large Scale

Visual Recognition Challenge. International Journal of Computer Vision, 115(3), pp.211–252. doi:https://doi.org/10.1007/s11263-015-0816-y.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, pp.5998–6008.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929. Available at: https://arxiv.org/abs/2010.11929.

Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning, pp.1597–1607.

He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.9726–9735. doi:https://doi.org/10.1109/cvpr42600.2020.00975.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R. and Valko, M. (2020). Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. Advances in Neural Information Processing Systems, 33, pp.21271–21284.

Finn, C., Abbeel, P. and Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Proceedings of the 34th International Conference on Machine Learning, pp.1126–1135.

Snell, J., Swersky, K. and Zemel, R. (2017). Prototypical Networks for Few-shot Learn-

ing. Advances in Neural Information Processing Systems, 30, pp.4077–4087.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. and others (2016). Matching Networks for One Shot Learning. Advances in Neural Information Processing Systems, 29, pp.3630–3638.