

Article

Non-Invasive Individual Re-Identification of Water Monitors (*Varanus salvator*) Using Deep Learning

Chayatorn Thongsub¹, Chattraphas Pongcharoen¹, Warong Suksavate¹, Kornorn Srikulnath^{1,2} 
and Prateep Duengkae^{1,2,*}

¹ Special Research Unit for Wildlife Genomics (SRUWG), Department of Forest Biology, Faculty of Forestry, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand; chayatorn.tho@ku.th (C.T.); chat-traphas.p@ku.ac.th (C.P.); warong.su@ku.ac.th (W.S.); kornorn.s@ku.ac.th (K.S.)

² Animal Genomics and Bioresource Research Unit (AGB Research Unit), Faculty of Science, Kasetsart University, 50 Ngamwongwan, Chatuchak, Bangkok 10900, Thailand

* Correspondence: prateep.du@ku.ac.th

Abstract

Effective management of urban Asian water monitor (*Varanus salvator* (Laurenti, 1768)) populations requires precise individual identification, yet traditional physical-marking methods remain invasive and labor-intensive. This study developed a non-invasive, automated photographic re-identification (Re-ID) system using deep learning and computer vision to facilitate population monitoring in semi-urban environments. We evaluated seven deep learning configurations based on ResNet50 incorporating Squeeze-and-Excitation (SE), Convolutional Block Attention Module (CBAM), and Batch Normalization neck (BNNeck) optimizations and benchmarked them against traditional feature matching (HotSpotter) using an open-set evaluation dataset of 3311 images across 161 Side-IDs focusing on unique lateral head-scale patterns. HotSpotter demonstrated immediate field viability, achieving a Rank-1 accuracy of 99.88% and a mean Average Precision (mAP) of 90.40%. Among the deep learning architectures, the baseline ResNet50 achieved the highest Rank-1 accuracy of 75.39% and mAP of 55.97%. As a decision-support framework, the deep learning pipeline achieved over 87% Rank-5 accuracy, drastically reducing manual screening effort and cognitive load during capture–mark–recapture surveys. This non-invasive framework establishes a scalable, welfare-friendly protocol for long-term urban wildlife management and biodiversity monitoring.

Keywords: deep learning; *Varanus salvator*; re-identification (Re-ID); animal biometrics; urban wildlife management; non-invasive monitoring; ResNet50



Academic Editors: Xinhai Li and
Luc Legal

Received: 30 June 2026

Revised: 3 September 2026

Accepted: 4 September 2026

Published: 7 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Water monitors (*Varanus salvator*) serve as critical apex predators and bioindicators of ecosystem health in tropical ecosystems [1]. In rapidly urbanizing environments across Thailand, such as public parks and university campuses, water monitor populations have expanded significantly. A prominent example is the Bangkhen Campus of Kasetsart University in Bangkok, where an extensive network of interconnected ponds and green infrastructure provides favorable foraging habitats, resulting in high population densities [2]. Effective urban wildlife management, ecological monitoring, and human–wildlife conflict mitigation in these modified landscapes require reliable, long-term tracking of individual animals [3]. While traditional line-transect surveys provide broad population

density estimates, robust demographic parameters such as survival rates, movement dynamics, and site fidelity depend on capture–mark–recapture (CMR) frameworks. However, physical-marking techniques (e.g., toe-clipping, branding, or PIT-tagging) are invasive and labor-intensive, and can induce significant stress or alter natural behaviors [4].

To overcome these welfare and operational limitations, photographic individual re-identification (Re-ID) using natural biometric markings has emerged as a non-invasive alternative in wildlife research [3]. In Thailand, automated photo-identification has primarily been deployed for high-profile megafauna, such as pattern-recognition algorithms for Indochinese tigers (*Panthera tigris corbetti*) [5] or registration of captive Asian elephants (*Elephas maximus*). Although Asian water monitors possess intricate, individually unique scale patterns along their lateral head and neck regions [6], automated photographic identification has not yet been established for this species.

In computer vision for wildlife biometrics, automated individual re-identification generally follows two distinct computational paradigms: classical local feature matching and deep learning representation learning. Classical algorithms, most notably the HotSpotter framework [7] utilizing Scale-Invariant Feature Transform (SIFT) descriptors paired with Local Naive Bayes Nearest Neighbor (LNBNN) matching, excel at fine-grained keypoint matching. HotSpotter operates without requiring prior network training, making it exceptionally effective for datasets with limited individual instances [7]. Conversely, deep convolutional neural networks (CNNs), such as ResNet50 [8], augmented with attention modules like Squeeze-and-Excitation (SE) [9] or Convolutional Block Attention Module (CBAM) [10] learn abstract, end-to-end feature representations directly from raw images [11]. While deep learning offers superior scalability, fast inference via vector embeddings, and potential integration into automated edge devices [12,13], its performance heavily depends on the volume and diversity of training data [14].

Despite the growing application of both paradigms in ecological studies such as camera-trap mammal surveys [12] or spot-pattern recognition in monitor lizards [6], no study to date has benchmarked deep learning against established classical computer vision software like HotSpotter for *V. salvator*. This presents a critical question for conservationists and field ecologists: Does deep learning offer a performance advantage over established feature matching tools like HotSpotter when applied to realistic, data-constrained wildlife monitoring sets?

To address this gap, this study presents the first rigorous comparative benchmark between classical computer vision (HotSpotter) and deep learning architectures for non-invasive re-identification of Asian water monitors (*V. salvator*). Using a standardized dataset of lateral head-scale images collected at Kasetsart University, we evaluate seven deep learning variants based on a ResNet50 backbone incorporating SE, CBAM, and Batch Normalization neck (BNNeck) optimizations against the traditional SIFT/LNBNN HotSpotter baseline. Specifically, we evaluate retrieval performance using cumulative matching characteristics (CMC) and mean Average Precision (mAP), while employing gradient-weighted class activation mapping (Grad-CAM) to assess the spatial attention integrity of learned deep features. By explicitly evaluating deep learning under realistic dataset constraints, this study provides a realistic evaluation of current algorithmic limitations, highlights the immediate field viability of classical feature matchers, and establishes a decision-support baseline for future scalable, welfare-friendly urban wildlife management.

2. Materials and Methods

2.1. Study Site and Data Acquisition

We conducted the study at the Bangkhen Campus of Kasetsart University. This area features a dense population of water monitors (*Varanus salvator*) that live in a semi-urban

environment. Field data collection followed systematic line transects along campus ponds and water channels (Figure 1). Each sighting was recorded with its metadata and UTM coordinates to ensure that a reliable ground-truth dataset could be built.

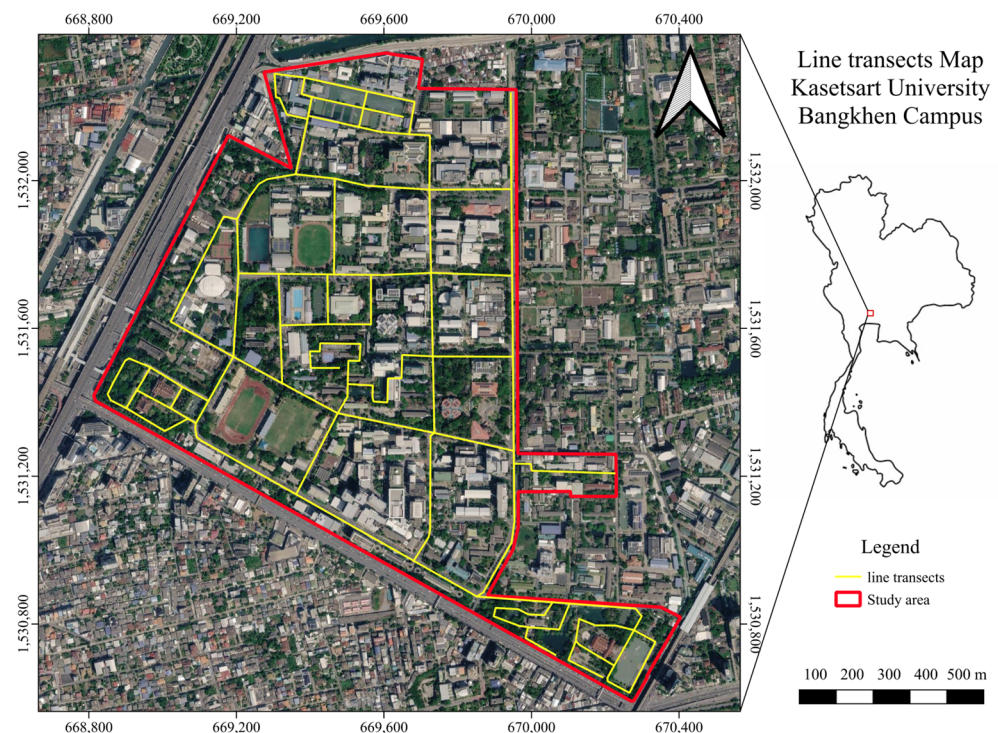


Figure 1. Map of the study area (red outline) and line transects (yellow lines) that researchers used to survey water monitors at the Bangkok Campus of Kasetsart University. The arrow indicates north.

Our photographic data targeted the left and right lateral profiles of the head region (Figure 2). We selected this target because the head displays highly distinct and stable markings. Furthermore, the head remains accessible when the lizard is semi-submerged, enabling non-invasive tracking. To avoid disturbing the animals' natural behavior, we maintained a distance of at least 3 m and photographed them using a Nikon D5 digital SLR camera with a 200–500 mm (Nikon Corporation, Tokyo, Japan) lens (Nikon Corporation, China).



Figure 2. Photographic methods: Photographs showing the (a) left and (b) right sides of the head.

2.2. Materials and Equipment

We performed data management, traditional feature matching inference, and initial inference on an ASUS TUF Gaming A15 laptop equipped with an AMD Ryzen 7 4800H

processor, 8 GB of RAM, and an NVIDIA GeForce GTX 1650 Ti GPU. By contrast, model training and ablation experiments were run on a powerful workstation featuring an Intel Core i7-13700K CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4070 SUPER GPU.

For software, we selected the Anaconda distribution, which ran Python 3.8.x and PyTorch 2.4.1 with CUDA 12.1 acceleration. Image processing and numerical tasks relied on core libraries, including Torchvision 0.19.1+cu121, OpenCV 4.11.0, and NumPy 1.24.4.

2.3. Data Preparation and Preprocessing

Prior to model training and evaluation, we manually screened the image collection to remove blurry, out-of-focus, or redundant photographs. To isolate diagnostic biometric markings and eliminate background clutter, all remaining images were tightly cropped around the head region (Figure 3). Images were horizontally flipped where necessary to maintain a uniform orientation.

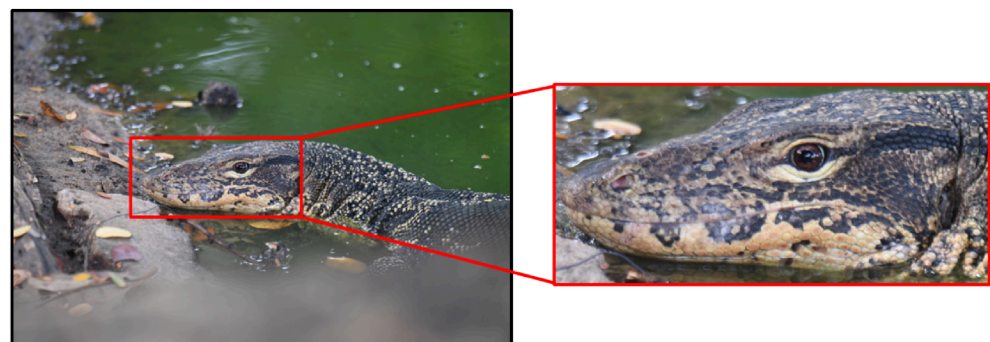


Figure 3. Example of the cropping process that isolates the head region of the animal.

To establish a high-rigor ground-truth dataset and mitigate labeling bias, all individual identities were independently verified by expert herpetologists with over three years of field experience studying *Varanus salvator*. Verification relied on multi-point spatial alignment of head-scale geometries and relative scale ratios. To prevent observer fatigue and ensure consistent accuracy, validation sessions were strictly capped at two hours per sitting.

Because bilateral scale patterns in *V. salvator* are asymmetric, left and right lateral profiles were treated as distinct identities (Side-IDs) during model feature extraction. To standardize spatial orientation and eliminate directional variance, all images were horizontally flipped to a uniform canonical direction prior to model input. Consequently, all model training and performance evaluations in this study were conducted strictly at the Side-ID level. In the proposed deployment framework, individual-level continuity across both sides is managed at the application layer via a relational database metadata table that maps paired left and right Side-IDs back to a single master individual identity.

Data Preprocessing and Augmentation: All cropped input images were resized to 384×384 pixels. For deep learning training, we implemented a data augmentation pipeline to help the model robustly handle varying outdoor conditions. These techniques included Random Affine transformations (rotating images by up to $\pm 15^\circ$ to 20° and translating them by up to 10%), Random Horizontal Flip, and Color Jitter (adjusting brightness and contrast by 0.1–0.2 or 10–20%). This was followed by standard ImageNet normalization. For the HotSpotter evaluation, the exact same set of cropped images was used without any geometric or color augmentations to maintain direct comparative integrity.

Batch Sampling Strategy: To construct effective mini-batches for the triplet margin loss, a PxK batch sampling strategy was employed. Specifically, each batch was formed by randomly selecting $P = 8$ identities (classes), with $K = 4$ instances sampled per identity. This configuration resulted in a total batch size of 32 images per iteration.

Network Architecture and Feature Embedding: ResNet50, initialized with default pre-trained weights, was utilized as the backbone. Across all experimental setups, the final feature embedding dimension was consistently projected to 256 dimensions. L2 normalization was applied to the output embeddings prior to loss computation.

Optimization Details: The models were optimized using the Adam optimizer with a fixed learning rate of 1×10^{-4} . The objective function employed was the triplet margin loss with a margin parameter set to 0.5.

Training Protocol and Early Stopping: The maximum number of training epochs was set to 120. To mitigate overfitting, an early stopping mechanism was implemented with a patience of 10 epochs. The “Best Model” was determined and saved based on the lowest observed training loss during the process.

2.4. Proposed Re-Identification Framework

To thoroughly evaluate lizard skin pattern re-identification, our methodology establishes a dual-paradigm framework comprising advanced Deep Learning (DL) architectures as the primary proposed approach, evaluated alongside HotSpotter [7] as the primary handcrafted baseline framework.

2.4.1. Deep Learning Architectures (Primary Proposed Approach)

ResNet-50 [8] was selected as the foundational backbone for deep feature extraction. To effectively capture fine-grained textural variations in lizard skin patterns, two attention mechanisms were integrated into the backbone:

- Squeeze-and-Excitation Block (SEBlock) [9]: Performs channel-wise recalibration to dynamically emphasize critical identity-specific pattern features;
- Convolutional Block Attention Module (CBAM) [10]: Sequentially applies channel and spatial attention to simultaneously identify “what” features are present and “where” informative skin regions are localized.

Additionally, a Batch Normalization neck (BNNeck) structure [11] was incorporated to reconcile the optimization imbalance between triplet loss and cross-entropy loss. Feature extraction performance was systematically evaluated at two distinct embedding stages: Pre-BN (representing raw backbone feature distributions) and Post-BN (representing the normalized feature embedding space). This comparative setup enabled us to determine the optimal configuration for maximum retrieval precision.

2.4.2. Primary Baseline Framework (HotSpotter)

To establish a rigorous comparative benchmark against traditional non-deep learning approaches, we implemented HotSpotter [7] as the primary baseline framework. HotSpotter utilizes the Local Naive Bayes Nearest Neighbor (LNBNN) algorithm to perform instance identification based on local invariant keypoints. The system detects invariant keypoints via Hessian-Affine detectors and extracts Scale-Invariant Feature Transform (SIFT) descriptors from lizard skin textures. Keypoint matching is accelerated using Fast Library for Approximate Nearest Neighbors (FLANN) kd-trees (configured with four randomized trees for memory-efficient indexing), and match scoring is executed via the LNBNN formulation to measure relative background distance without requiring offline training.

2.5. Training Procedure and Experimental Design

Our training routine relied on a joint learning approach that combined cross-entropy loss for classification with triplet margin loss (margin = 0.5) for embedding-space optimization [15]. Training was conducted using the Adam optimizer (LR = 1×10^{-4}) with early stopping (patience = 10). Each model was trained with a batch size of 32 for up to

120 epochs. The early stopping mechanism terminated training if the validation metric did not improve for ten consecutive epochs. We designed seven experimental configurations to evaluate the contribution of each architectural component. These configurations ranged from the standard ResNet50 baseline to hybrid CBAM + BNNeck models (Table 1).

Table 1. Experimental configurations used to evaluate the contribution of each architectural component.

Experiment ID	Backbone Architecture	Attention	Neck
1	ResNet50	-	Standard (Baseline)
2	ResNet50	SEBlock	Standard
3	ResNet50	SEBlock	BNNeck (Pre-BN)
4	ResNet50	SEBlock	BNNeck (Post-BN)
5	ResNet50	CBAM	Standard
6	ResNet50	CBAM	BNNeck (Pre-BN)
7	ResNet50	CBAM	BNNeck (Post-BN)

To enable a rigorous head-to-head comparison, HotSpotter was evaluated on the exact same dataset of 3311 cropped images under an All-to-All querying protocol. Each image served as a query against the entire database, excluding self-matches. Performance across all eight configurations (seven deep learning variants plus HotSpotter) was benchmarked using mean Average Precision (mAP) and cumulative matching characteristics (CMC) curves up to Rank-10.

2.6. Performance Evaluation and Model Interpretability

We evaluated model performance using cumulative matching characteristics (CMC) for Rank-1 and Rank-5 accuracy, as well as mean Average Precision (mAP). To ensure model transparency, we used gradient-weighted class activation mapping (Grad-CAM) to visualize attention heatmaps [16]. We further defined a spatial attention integrity score that quantifies how well the model’s attention aligned with the diagnostic head region. For each test image, the Grad-CAM activation map was thresholded. We then compared this map with the manually annotated head-scale region. The system expressed the overlap as an intersection-over-union ratio between the high-activation area and the annotated biometric markers on the water monitor’s head. The reported value (e.g., 91.48%) is the mean overlap ratio, which we averaged across all evaluated test images for a given configuration.

2.7. Statistical Significance Testing

To rigorously validate the performance differences observed among the seven deep learning architectures, we conducted a matched-pair statistical analysis on the Rank-1 binary identification outcomes (correct vs. incorrect) generated by the models across the entire open-set evaluation test set. Cochran’s Q test [17] was first employed to determine if there were statistically significant variations in identification performance across all architectures. Following a significant Cochran’s Q result, we performed post hoc pairwise comparisons between the primary baseline (ResNet-50) and each augmented configuration to isolate specific performance deviations. To maintain statistical rigor and mitigate Type I error inflation arising from multiple comparisons, we utilized McNemar’s exact test [18] coupled with a Bonferroni correction, establishing a conservative significance threshold of $\alpha_{adj} \approx 0.0083$ (0.05 divided by 6 comparisons) [19].

3. Results

3.1. Dataset Composition and Screening

The initial data collection yielded 15,469 images representing 291 individual water monitors. From this dataset, we established 510 Side-IDs to account for bilateral asymme-

try. We noted that only a single side was captured for 72 individuals. To ensure model robustness, we applied a strict image-per-ID threshold for the experimental sets:

- Training Set (Closed-set Evaluation): This set comprised 50 Side-IDs featuring a high data density of 30–40 images per ID, and contained a total of 1825 images, which provided sufficient intra-class variation;
- Test Set (Open-set Evaluation): This set comprised 161 Side-IDs, including the 50 known training IDs and 111 novel IDs that the model did not encounter during training, and included 3311 images with a density of 10–40 images per ID.

We excluded the remaining 349 Side-IDs from the primary experiments owing to insufficient image counts. This exact 3311-image test set was evaluated across all deep learning configurations and the HotSpotter benchmark.

3.2. Comparative Model Performance Benchmark

To evaluate non-invasive photographic re-identification for *Varanus salvator*, we benchmarked the primary handcrafted baseline framework, HotSpotter [7], alongside seven deep learning configurations based on a ResNet-50 backbone [8]. Performance was evaluated using Rank-1 and Rank-5 retrieval accuracy, mean Average Precision (mAP), and spatial attention integrity scores (for deep learning architectures). Table 2 summarizes these comparative results, with the traditional baseline positioned prominently to establish the benchmark performance. Furthermore, to offer practical insight into the computational cost of the proposed approach, training time was recorded. For each of the seven deep learning model configurations, completing a single full training run required approximately 23–24 h on the designated workstation.

Table 2. Performance metrics and spatial attention integrity scores comparing the eight configurations.

Experiment ID	Experiments	Rank-1 (%)	Rank-5 (%)	mAP (%)	Spatial Attention Integrity (%)
1	Baseline (ResNet50)	75.39	87.71	55.97	89.34
2	ResNet50 + SEBlock	74.54	88.25	55.49	88.26
3	ResNet50 + SEBlock + BNNeck (Pre)	60.92	80.4	40.06	88.09
4	ResNet50 + SEBlock + BNNeck (Post)	67.14	85.47	45.73	90.35
5	ResNet50 + CBAM	71.19	87.32	51.69	90.66
6	ResNet50 + CBAM + BNNeck (Pre)	72.64	87.13	53.57	91.48
7	ResNet50 + CBAM + BNNeck (Post)	74.81	88.1	55.73	90.32
8	HotSpotter (LNBNN + SIFT)	99.88	100	90.4	-

Figure 4 illustrates the retrieval performance across various ranking depths using CMC curves. The curves demonstrate that HotSpotter achieved near-perfect retrieval accuracy across all rank depths, outperforming all deep learning configurations by a substantial margin. Among the deep learning models, accuracy values converged rapidly toward higher accuracy by Rank-5, with Experiments 1, 2, 5, 6, and 7 approaching or exceeding 87%.

To rigorously evaluate performance differences among the seven deep learning architectures ($N = 3311$ test queries), a Cochran's Q test [17] was conducted on matched binary Rank-1 outcomes. The test confirmed statistically significant overall performance variation across models ($Q = 373.88$, $df = 6$, $p < 0.001$).

Post hoc pairwise McNemar's tests [18] with Bonferroni correction ($\alpha_{adj} = 0.0083$) [19] revealed that the standard Baseline ResNet-50 (75.40%) performed statistically equivalently to ResNet50 + SEBlock (74.29%, $p = 0.1982$) and ResNet50 + CBAM + BNNeck (Post) (74.63%, $p = 0.3741$). Conversely, all other modifications resulted in a statistically significant degradation in Rank-1 accuracy ($p < 0.0083$), with the largest performance drop observed in SEBlock + BNNeck (Pre) (61.32%, $p < 0.001$) (Table 3). Adhering to model parsimony, Baseline ResNet-50 remains the optimal architecture for this task.

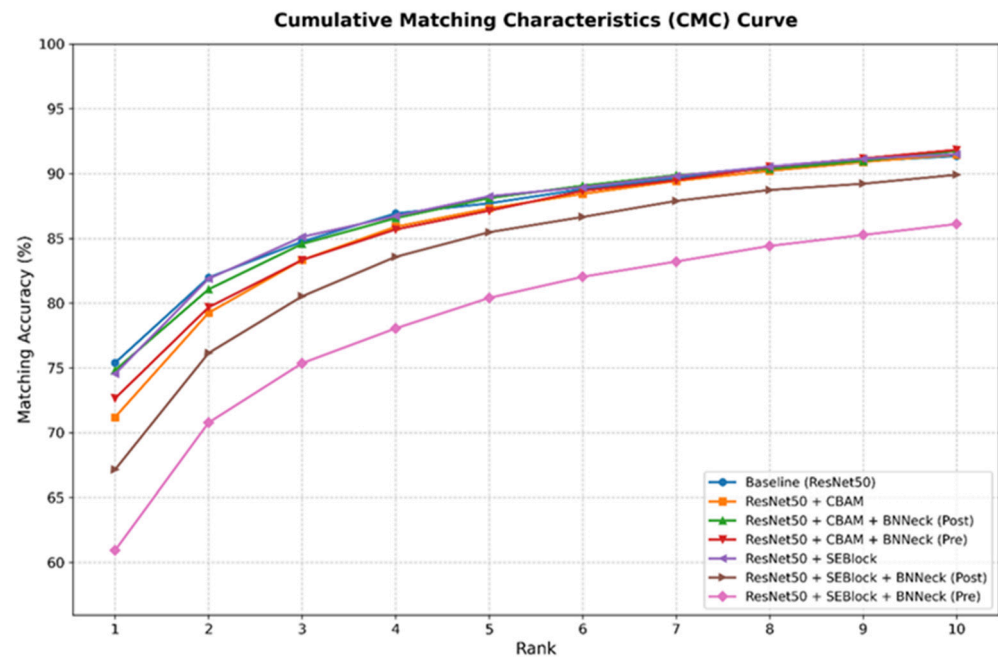


Figure 4. Comparison of cumulative matching characteristics (CMC) curves for the eight evaluated configurations. The graph illustrates the probability of correct individual identification within the top 1–10 candidates that the system returned.

Table 3. Pairwise McNemar’s test results comparing Baseline ResNet-50 with augmented variants.

Model Comparison	Baseline Rank-1 (%)	Variant Rank-1 (%)	<i>p</i> -Value	Significant
Baseline vs. SEBlock	75.4	74.29	0.1982	No
Baseline vs. SEBlock + BNNeck (Pre)	75.4	61.32	1.50×10^{-49}	Yes
Baseline vs. SEBlock + BNNeck (Post)	75.4	67.37	6.01×10^{-19}	Yes
Baseline vs. CBAM	75.4	71.53	1.05×10^{-5}	Yes
Baseline vs. CBAM + BNNeck (Pre)	75.4	72.61	0.0013	Yes
Baseline vs. CBAM + BNNeck (Post)	75.4	74.63	0.3741	No

3.3. Key Experimental Findings and Algorithmic Comparison

A systematic analysis of the empirical data reveals several key insights regarding the trade-offs between handcrafted feature matching and learned deep representations:

- **Superiority of Handcrafted Feature Matching (HotSpotter Benchmark):** HotSpotter [7] achieved the highest overall re-identification performance by a significant margin, recording a Rank-1 accuracy of 99.88%, a Rank-5 accuracy of 100.00%, and an mAP of 90.40%. This dominant result demonstrates that traditional local feature matching utilizing Hessian Affine keypoint detectors paired with Scale-Invariant Feature Transform (SIFT) descriptors and Local Naive Bayes Nearest Neighbor (LNBNN) scoring is exceptionally robust for extracting distinctive scale pattern biometric identifiers on lizard heads without requiring offline model training.
- **Deep Learning Baseline Performance and Representation Constraints:** Among the seven deep learning variants, the unaugmented baseline ResNet 50 model [8] achieved the highest deep learning accuracy, reaching a Rank-1 accuracy of 75.39% and an mAP of 55.97%. This was closely matched by Experiment 7 (ResNet 50 + CBAM + BNNeck Post-BN) [10,11] at 74.81% Rank-1 accuracy and 55.73% mAP. These results indicate that while deep networks successfully extract discriminative features, their precision remains constrained when trained under limited sample sizes per individual.

- **Spatial Attention Integrity and Feature Localization:** Incorporating the Convolutional Block Attention Module (CBAM) [10] enhanced the network's ability to focus feature activation on diagnostic anatomical areas. Specifically, Experiment 6 (ResNet 50 + CBAM + BNNeck Pre-BN) achieved the highest spatial attention integrity score at 91.48%, confirming that sequential channel and spatial attention effectively guide neural features toward informative head-scale regions. However, despite achieving the highest spatial focus score due to its broad coverage of the designated diagnostic region, Experiment 6 was not the most accurate deep learning model (Rank-1: 72.64%). This discrepancy suggests that deep neural networks do not necessarily rely on holistic, head wide pattern areas in the same manner as human observers, but may instead extract re-identification cues from compact, highly discriminative sub-features. (Note: Spatial attention integrity metrics are not applicable to HotSpotter, as it relies on discrete keypoint descriptors rather than continuous activation heatmaps.)
- **Impact of Structural Complexity and Bottlenecks:** Conversely, configurations combining Squeeze-and-Excitation blocks (SEBlock) [9] with a BNNeck structure [11] exhibited performance degradation. The Pre-BN variant (Experiment 3) recorded the lowest performance across all evaluated models (Rank-1: 60.92%, mAP: 40.06%). This suggests that adding architectural complexity alongside loss optimization constraints can aggravate overfitting in data-constrained wildlife re-identification scenarios.

4. Discussion

4.1. Practical Suitability and Immediate Field Viability

The empirical results of this study establish that traditional local feature matching via HotSpotter [7] is the primary, immediately deployable solution for non-invasive re-identification of Asian water monitors (*Varanus salvator*). HotSpotter achieved near-perfect retrieval capability, recording a Rank-1 accuracy of 99.88% and a mean Average Precision (mAP) of 90.40%. It significantly outperformed all seven evaluated deep learning configurations, where the top-performing baseline ResNet-50 [8] reached a maximum Rank-1 accuracy of 75.39% and an mAP of 55.97%.

The observed performance disparity highlights distinct operational roles for classical feature matchers and deep learning representations in wildlife monitoring. For small-to-medium monitoring datasets, handcrafted feature matchers eliminate the necessity of extensive offline model training while delivering superior standalone retrieval precision. Conversely, although deep learning configurations currently yield lower top-rank precision, their Rank-5 accuracy exceeding 87% underscores their utility as robust candidate retrieval engines.

Crucially, our proposed framework operates strictly as a human-in-the-loop decision-support system rather than an autonomous decision-maker. As illustrated in Figure 5, the pipeline shortlists the most probable candidate identities (Top-1 to Top-5) alongside similarity distance metrics, streamlining gallery screening and substantially reducing researchers' cognitive workload during capture-mark-recapture (CMR) surveys [3]. Because final identity verification is performed by a human expert acting as a gatekeeper, algorithmic false positives and false negatives are filtered out prior to demographic modeling. This operational design effectively prevents misidentification bias from propagating into downstream population density and survival estimates.

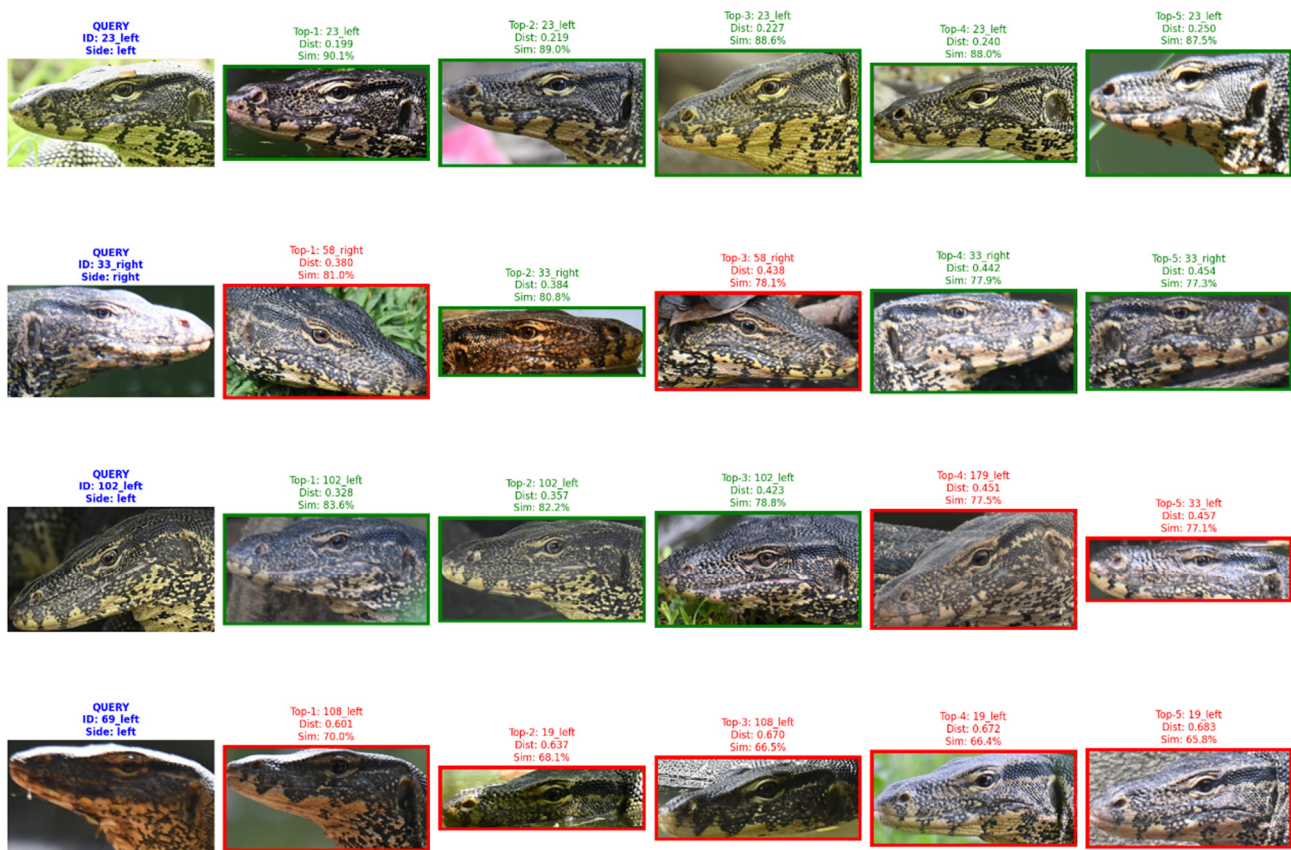


Figure 5. Visual examples from the retrieval system, including the query (ID) alongside the top 1–5 results that the system selected. The text marks the distance values and similarity percentages. These metrics demonstrate system performance across different conditions. A green frame indicates that the system guessed the identity correctly. By contrast, a red frame indicates that the system guessed partially or completely incorrectly.

4.2. Dataset Scale Constraints and Root Causes of Deep Learning Limitations

The primary constraints on deep learning performance in this study are dataset scale and sample density per identity. The training set was restricted to 50 individual Side-IDs with 30–40 images per ID, reflecting the real-world operational challenges of field data collection in natural and semi-urban habitats [13].

From an algorithmic perspective, deep metric learning architectures relying on triplet loss require extensive inter-class feature diversity and dense sample representation to optimize embedding spaces effectively [15]. In a restricted ID pool, the frequency of informative hard negative mining is diminished, leading to sub-optimal decision boundaries for open-set re-identification [14]. Furthermore, environmental noise such as varying lighting conditions, severe mud coverage, specular skin reflections, and dynamic body contortions requires higher sample density per individual for deep convolutional layers to learn invariant representations [20]. Conversely, HotSpotter’s SIFT/LNBNN formulation operates as a non-parametric, zero-shot keypoint matcher; it bypasses parametric representation learning altogether, making it inherently immune to small-sample overfitting [7].

4.3. Comparative Trade-Offs and Computational Scalability

Rather than asserting an unconditional superiority of one paradigm over another, practical deployment depends on specific operational constraints and dataset scales. Distinct trade-offs exist between handcrafted keypoint matching and learned deep representations.

Traditional SIFT-based matching via HotSpotter achieves superior matching accuracy on limited sample sizes because keypoint detection focuses directly on invariant local scale

patterns [7]. However, SIFT matching incurs a computational complexity of $O(N)$ per query against the descriptor database, resulting in exponential increases in query latency and memory overhead as image galleries expand [13].

In contrast, deep neural networks compress high-dimensional image inputs into compact, fixed-dimensional embedding vectors (e.g., 256 or 512 dimensions). This representation enables sub-linear or constant-time $O(1)$ similarity retrieval across massive image galleries when paired with approximate nearest neighbor vector indexing frameworks such as FAISS [13]. Additionally, deep learning backbones can be integrated into unified, end-to-end processing pipelines combining real-time object detection (e.g., YOLOv8) and feature extraction, enabling direct deployment onto edge-AI hardware or autonomous camera traps [20,21].

Finally, the empirical literature on computer vision scaling laws demonstrates that while handcrafted SIFT descriptors possess static representation capacity, deep learning models adhere to power-law performance gains relative to dataset volume [14]. Studies across diverse wildlife taxa confirm that when training datasets scale substantially in both individual count and sample density, deep learned representations progressively overcome local occlusion noise and surpass classical feature matchers in operational scalability [8,13,22].

4.4. Broader Conservation Implications and Future Directions

The successful deployment of photographic re-identification at Kasetsart University confirms the viability of computer vision as a welfare-friendly decision-support tool for urban *Varanus salvator* management. By replacing invasive physical tagging with photographic tracking, researchers can monitor population demographics while safeguarding animal welfare.

Future work should focus on advancing both algorithmic robustness and dataset comprehensiveness. To improve performance in open-world and data-scarce scenarios, future research should implement adaptive distance thresholding (θ) for explicit automated open-set rejection of novel individuals, alongside Few-Shot Metric Learning architectures to enhance generalization across low-sample regimes (e.g., 1–3 reference images per individual). Concurrently, expanding dataset diversity across varied lighting conditions and complex body contortions will improve feature representation invariance. From a practical deployment perspective, future efforts should target transferring unified detection and re-identification pipelines onto resource-constrained edge-AI hardware. Integrating these automated pipelines with municipal park surveillance infrastructure (CCTV) holds significant potential for establishing continuous, scalable wildlife monitoring networks within modified urban ecosystems.

4.5. Temporal Pattern Stability, Ontogenetic Shifts, and Localized Scarring Dynamics

A fundamental assumption underlying photographic re-identification frameworks is that individual skin patterns remain sufficiently stable over time to allow consistent matching. To date, no formal empirical studies have documented the temporal stability of skin markings in Asian water monitors (*V. salvator*). Based on our longitudinal field observations spanning over four years and structured dataset collection over three consecutive years, we identified two critical dynamic phases in pattern development:

- Ontogenetic Pattern Shifts in Juveniles (<1.3 m. total length): Juvenile individuals exhibit noticeable pattern metamorphosis and rapid shifts in spot density and contrast as they mature. Automated feature matching algorithms trained solely on adult patterns may encounter higher degradation when matching juvenile to adult transitions.

- **High Pattern Stability in Sub-Adults and Adults (>1.3 m. total length):** Once individuals attain sub-adult status, typically marked by a total length exceeding 1.3 m, their pattern configurations stabilize almost completely. Overall, a total of 90 individuals were encountered only once during the survey. To quantitatively validate this stability, we analyzed a longitudinal subset of 14 sub-adult and adult individuals tracked over intervals ranging from 12 to 36 months, with a maximum time gap of 36 months between the initial and final photographic encounters (Figure 6). Across this recaptured cohort, individuals were re-encountered an average of five times per ID (Figure 7). Longitudinal comparisons across these temporal intervals demonstrated near-zero structural drift in local skin markings (Figure 8a), validating that adult head-scale markings serve as highly reliable visual fingerprints for long-term tracking.

However, researchers must account for localized pattern alterations induced by physical trauma. In our dataset, a specific case study documented an adult individual experiencing skin injury across a light-pigmented pattern zone. Upon tissue healing, the regenerated epidermal area transformed into a dense, dark scar patch (Figure 8b).

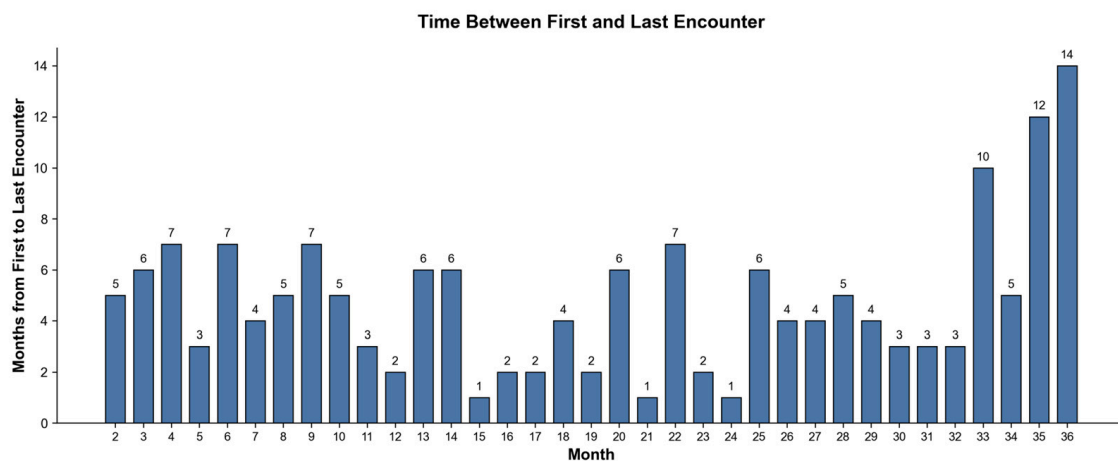


Figure 6. Duration between the first and last encounters. The bar chart illustrates the time span (in months) from the initial to the final observation of individuals.

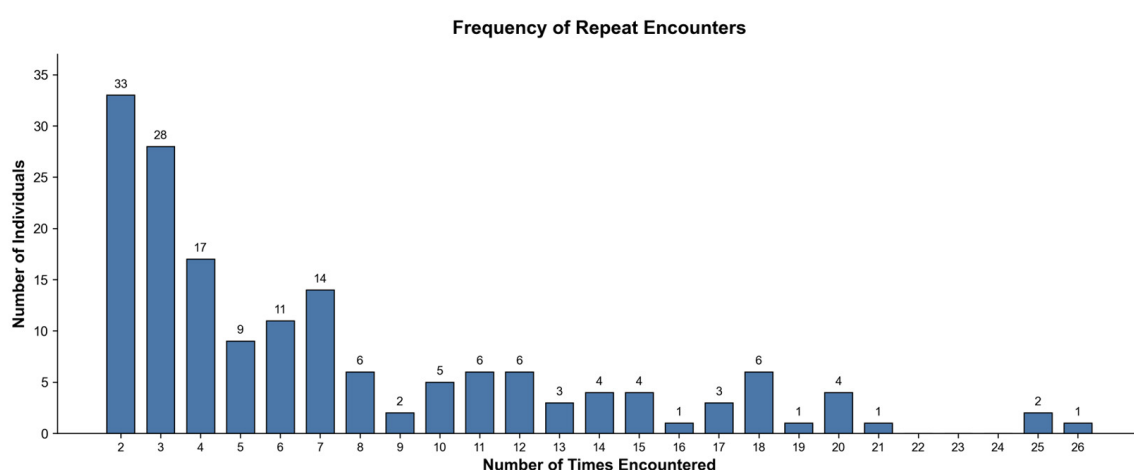


Figure 7. Frequency of repeat encounters. The bar chart displays the number of individuals recorded across varying numbers of encounters (ranging from 2 to 26 times).

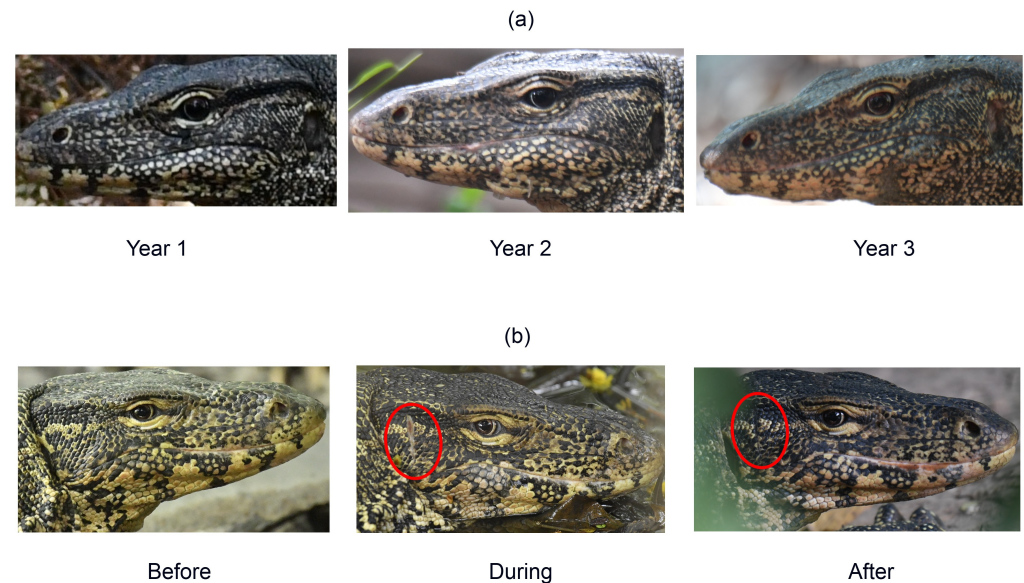


Figure 8. Longitudinal comparison of skin pattern stability in an adult *Varanus salvator* across three consecutive years (a) and a case study of localized pattern modification due to scarring (b), with red circles indicating the site of injury and subsequent pattern changes after healing.

5. Conclusions

This study evaluated non-invasive photographic re-identification (Re-ID) approaches for urban Asian water monitors (*Varanus salvator*) based on unique lateral head-scale patterns (Side-ID). The comparative evaluation demonstrated distinct trade-offs between traditional feature matching and deep learning architectures:

Feature Matching Baseline: Traditional SIFT/LNBNN matching via HotSpotter achieved the highest performance, with a Rank-1 accuracy of 99.88% and a mean Average Precision (mAP) of 90.40%. This confirms that local feature matching provides a highly accurate, immediately deployable solution for current dataset sizes without requiring model training.

Deep Learning Framework: Among the evaluated deep learning configurations, the baseline ResNet50 achieved the highest performance (75.39% Rank-1, 55.97% mAP). Across multiple configurations, Rank-5 accuracy exceeded 87%, indicating its utility as a decision-support tool to reduce manual screening effort in capture–mark–recapture (CMR) workflows. The primary constraints on deep learning performance in this study were dataset scale and the number of training identities. Future research should focus on expanding dataset volume and diversity, integrating end-to-end detection pipelines, and optimizing models for automated field operations. Overall, photographic Re-ID serves as a feasible, non-invasive alternative to traditional physical-marking methods for water monitor population monitoring.

Author Contributions: Conceptualization, C.T., C.P., W.S. and P.D.; data curation, C.T.; formal analysis, C.T.; funding acquisition, P.D.; investigation, C.T.; methodology, C.T.; project administration, C.T.; resources, P.D. and C.T.; software, C.T.; supervision, K.S., P.D., C.P. and W.S.; validation, C.T.; visualization, C.T.; writing—original draft preparation, C.T.; writing—review and editing, C.T., K.S. and P.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it relied entirely on non-invasive photographic monitoring of wild animals in their natural habitat, involving no physical capture, restraint, handling, or intentional disturbance to the species' natural behavior.

Data Availability Statement: Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to express their sincere gratitude to Pimchanok Buranapitak, Pattariya Sonaca, and Thaweesak Phinitichit for their valuable feedback and insightful suggestions throughout the study. Special thanks are also extended to the field data collection team, comprising Issarapong Bansongkij, Jedsada Noo wong, Piyapong Lapo, Janthiwa Phusomphong, Supansa Rerkdee, and Kanlaya Kaewklom, whose dedication and hard work were essential for data acquisition. Additionally, Chayatorn Thongsub would like to offer heartfelt thanks to his family, Suchat Thongsub, Sasithon Srisawat, Kamonpan Tongsub, and Mesiya Srisawat, for their unconditional love, patience, and constant encouragement throughout this academic journey. During the preparation of this manuscript, the authors used Gemini 3.1 Pro and Gemini Thinking Mode 3.1 to assist in the design of the BNNeck (Post-BN) structural comparison experiments, optimize the training scripts, and provide technical interpretation of the interaction between triplet and cross-entropy losses. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BN	Batch Normalization
BNNeck	Batch Normalization neck
CBAM	Convolutional Block Attention Module
CMC	Cumulative matching characteristics
CMR	Capture–mark–recapture
Grad CAM	Gradient-weighted class activation mapping
mAP	Mean Average Precision
Re-ID	Re-identification
SEBlock	Squeeze-and-Excitation block
Side-ID	Side specific identity
SRUWG	Special Research Unit for Wildlife Genomics

References

1. Suekamnurt, W. Ecological Studies of Water Monitor (*Varanus salvator*) and Impact of Utilization in the Laem Phak Bia Environmental Study, Research and Development Project Under Royal Initiatives. Master's Thesis, Silpakorn University, Bangkok, Thailand, 2007. (In Thai)
2. Thongsub, C.; Sonaca, P.; Buranapitak, P.; Duengkae, P. Monitoring on abundance and distribution of water monitors (*Varanus salvator*) in Kasetsart University Bangkok Campus: A case study at Kasetsart University, Bangkok, Thailand. *J. Wildl. Biodivers.* **2024**, *8*, 122–133.
3. Köhl, H.S.; Burghardt, T. Animal biometrics: Quantifying and detecting phenotypic appearance. *Trends Ecol. Evol.* **2013**, *28*, 432–441. [[CrossRef](#)]
4. McMahon, C.; Van Den Hoff, J.; Burton, H. Handling intensity and the short-and long-term survival of elephant seals: Addressing and quantifying research effects on wild animals. *AMBIO J. Hum. Environ.* **2005**, *34*, 426–429. [[CrossRef](#)]
5. Duangchantrasiri, S.; Sornsa, M.; Jathanna, D.; Jornburom, P.; Pattanavibool, A.; Simcharoen, S.; Kanishthajata, P.; Suebsen, P.; Klanprasert, S.; Kumar, N.S. Rigorous assessment of a unique tiger recovery in Southeast Asia based on photographic capture-recapture modeling of population dynamics. *Glob. Ecol. Conserv.* **2024**, *53*, e03016. [[CrossRef](#)]
6. Moore, H.A.; Champney, J.L.; Dunlop, J.A.; Valentine, L.E.; Nimmo, D.G. Spot on: Using camera traps to individually monitor one of the world's largest lizards. *Wildl. Res.* **2020**, *47*, 326–337. [[CrossRef](#)]
7. Crall, J.P.; Stewart, C.V.; Berger-Wolf, T.Y.; Rubenstein, D.I.; Sundaresan, S.R. Hotspotter—Patterned species instance recognition. In *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2013; pp. 230–237.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778.
9. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 7132–7141.

10. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19.
11. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Long Beach, CA, USA, 16–17 June 2019.
12. Norouzzadeh, M.S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.S.; Packer, C.; Clune, J. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E5716–E5725. [[CrossRef](#)]
13. Schneider, S.; Taylor, G.W.; Linquist, S.; Kremer, S.C. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods Ecol. Evol.* **2019**, *10*, 461–470. [[CrossRef](#)]
14. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 843–852.
15. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
16. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 618–626.
17. Cochran, W.G. The comparison of percentages in matched samples. *Biometrika* **1950**, *37*, 256–266. [[CrossRef](#)]
18. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [[CrossRef](#)]
19. Dunn, O.J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **1961**, *56*, 52–64. [[CrossRef](#)]
20. Parham, J.; Stewart, C.; Crall, J.; Rubenstein, D.; Holmberg, J.; Berger-Wolf, T. An animal detection pipeline for identification. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2018; pp. 1075–1083.
21. Varghese, R.; Sambath, M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
22. Ferreira, A.C.; Silva, L.R.; Renna, F.; Brandl, H.B.; Renoult, J.P.; Farine, D.R.; Covas, R.; Doutrelant, C. Deep learning-based methods for individual recognition in small birds. *Methods Ecol. Evol.* **2020**, *11*, 1072–1085. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.