Part-Based Re-Identification of Mediterranean Spur-Thighed Tortoises via Anatomical Scute Decomposition

Muhammad Mahamid[1], Efrat Yagur[2,3], and Ilan Shimshoni[1][0000-0002-5276-0242]

[1] University of Haifa, Haifa 3498838, Israel ishimshoni@is.haifa.ac.il
[2] Society for the Protection of Nature in Israel (SPNI)
[3] Jack, Joseph and Morton Mandel Gazelle Valley Urban Nature Park, Operated by the Municipality of Jerusalem and the SPNI

Abstract. We present a non-invasive re-identification system for Mediterranean spur-thighed tortoises (*Testudo graeca*) that exploits the anatomically conserved arrangement of 13 carapace scutes as a structural prior, decomposing identity matching into semantically meaningful part-level comparisons. The pipeline integrates YOLOv11 instance segmentation (96.0% mAP@50), Hungarian algorithm–based geometric alignment to a canonical template (96.7% labeling accuracy), a fine-tuned Vision Transformer with SubCenter ArcFace and Multi-Similarity losses ($d' = 2.26$), and a weighted geometric-mean aggregation of per-scute cosine similarities. On a curated dataset of 150 tortoises collected at the Gazelle Valley Urban Nature Park, Jerusalem, closed-set evaluation achieves 66.7% Top-1 (95% CI: 54.0–77.8%), 87.3% Top-5, and 98.4% Top-10, improving over a whole-image baseline by +40.1 pp Top-1 (McNemar $\chi^2 = 21.3$, $p < 10^{-3}$; Cohen's $h = 0.88$). Inference-time k-reciprocal re-ranking lifts Top-1 to 69.8% without retraining. We additionally evaluate against 138 *genuinely unseen* tortoises from a separate field archive and find that a leave-N-out simulation overestimates open-set AUC by up to 55% (0.930 $\rightarrow$ 0.600, DeLong $p < 10^{-10}$), revealing that standard simulated protocols mis-state deployment performance whenever the metric-learning loss sees every identity at training time. The deployable operating point is thus a Top-5 human-in-the-loop workflow (87.3%); we also analyse stage-wise error propagation and field robustness. The full annotated dataset and code will be released.

Keywords: Wildlife re-identification · Tortoise · Part-based matching · Vision Transformer · Hungarian algorithm · Open-set recognition.

1 Introduction

Individual identification of wildlife underpins modern conservation biology, enabling researchers to monitor population dynamics, estimate survival, and evaluate management interventions [30]. For long-lived species such as tortoises, which

may survive several decades, reliable individual recognition is essential for longitudinal mark–recapture studies. Mediterranean spur-thighed tortoises (*Testudo graeca*) are classified Vulnerable on the IUCN Red List [35] and face habitat loss, road mortality, and illegal collection [25,6,27].

Traditional marking methods carry substantial welfare costs: shell notching damages tissue [26], paint markers obstruct UV radiation [38], and PIT tags require invasive implantation [3], all under stressful handling [5]. Photo-identification from natural markings offers a non-invasive alternative that has succeeded across whales [7,10], zebras [11], tigers [22], seals [24], elephants [32], and sea turtles [8,9,16]. The biological foundation for tortoise photo-identification is strong: scutes form through reaction–diffusion dynamics during embryogenesis [23], and terrestrial chelonians do not undergo ecdysis, so the original pattern geometry is preserved throughout life [1]. Plastron morphometry achieves 97.5–99.8% identification on *T. graeca ibera* [17], but requires manual landmark annotation incompatible with automated, non-invasive pipelines.

Application context. This work addresses a concrete conservation need at the Gazelle Valley Urban Nature Park, Jerusalem, operating a rehabilitation program for spur-thighed tortoises confiscated from illegal captivity. Efrat Yagur, who is responsible for wildlife at the park and a co-author of this paper, monitors a resident population exceeding 200 individuals; current practice relies on manual visual comparison of carapace photographs, scaling linearly with population size and demanding 15–30 minutes per identification.[4] The deployment target is therefore not autonomous Top-1 recognition but a Top-5 shortlist that staff confirm, replacing exhaustive pairwise comparison with a check of five candidates (Table 1).

Contributions. We make four contributions:

- A *scute-level* re-identification pipeline that leverages the standardized 13-scute anatomical template (Figure 1), decomposing whole-image matching into semantically meaningful part comparisons with anatomically consistent type matching ($V_1 \leftrightarrow V_1$, etc.).
- We show that the conserved 13-scute geometry alone—without any learned localisation head—suffices to assign anatomical labels at 96.7% accuracy under viewpoint and orientation variation, via Hungarian matching to a fixed Mahalanobis template with a Smart Rotation Search fallback for ambiguous orientations.
- Empirical demonstration that part-based decomposition lifts Top-1 accuracy by +40.1 pp over a whole-image baseline (McNemar $\chi^2 = 21.3$, $p < 10^{-3}$; Cohen's $h = 0.88$), accompanied by a stage-wise error-propagation analysis through the full segmentation → alignment → embedding cascade.
- A *real* open-set evaluation against 138 genuinely unseen tortoises, paired with a leave-N-out simulation that serves as an upper-bound reference. The simulation overestimates open-set AUC by up to 55%, revealing a systematic bias of leave-N-out protocols when the metric-learning loss trains on every identity.

[4] Estimated by park staff during data collection.

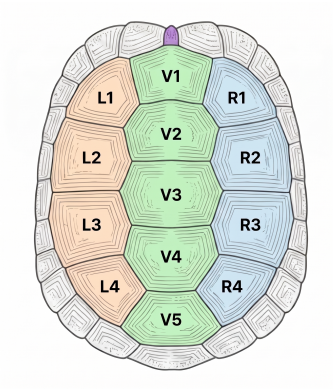


Fig. 1: Carapace scute anatomy of *Testudo graeca*. The 13 carapace scutes are arranged in a conserved pattern: five vertebral (V_1–V_5, green) along the midline, four costal scutes on each flank (L_1–L_4, orange; R_1–R_4, blue), and the unpaired *nuchal* (purple) at the anterior midline immediately above the neck. The nuchal, together with the detected head, fixes the forward direction of the shell and serves as the structural prior for the proposed pipeline.

2 Related Work

Wildlife re-identification. The field has migrated from hand-crafted descriptors (e.g. the SIFT-based HotSpotter [11] and astronomical pattern-matching for whale sharks [2]) toward deep metric learning. MegaDescriptor [29] pretrains a Vision Transformer [15] on millions of wildlife images spanning hundreds of species and provides powerful transfer features even for low-shot regimes. Modern re-identification systems combine triplet loss [31,19] with angular-margin losses [14]; recent variants such as Multi-Similarity loss [37], SubCenter ArcFace [13], and center loss [39] address pair mining, intra-class variance, and feature compactness, respectively. At the post-retrieval stage, k-reciprocal reranking [40] exploits gallery structure to sharpen initial similarities. Domain-specific systems include ATRW for Amur tigers [22], SealID for Saimaa ringed seals [24], and fully automated photo-ID for humpback whales [10].
Chelonian photo-identification. Sea turtle facial-scale identification reaches over 90% Top-1 [8,9,16], leveraging high-contrast facial scale patterns. For terrestrial tortoises, plastron morphometry achieves 97.5–99.8% accuracy [17] but requires manual landmark annotation and physical handling. Citizen-science photo-ID has been validated for endangered reptiles [18]. To our knowledge, no prior work has delivered a *fully automated* deep-learning system for re-identifying terrestrial tortoises from carapace scute patterns; the present work fills this gap.
Part-based representations. Part-based methods decompose images into semantic regions, providing robustness to occlusion and viewpoint. PCB demonstrated the benefit for person re-identification [34]; similar approaches succeeded for tigers [22] and seals [24]. Where parts have known semantic identity, the

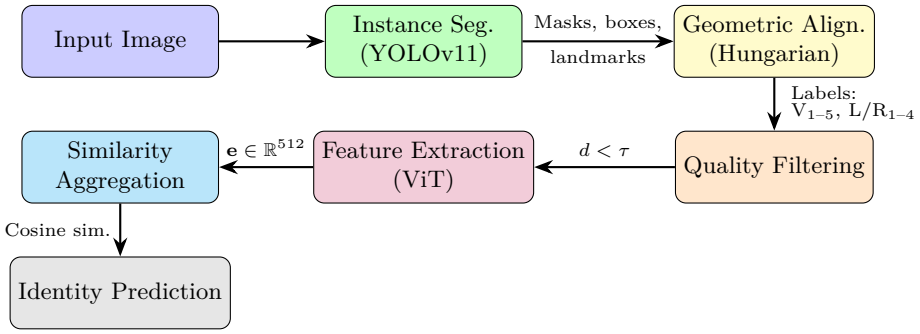


Fig. 2: End-to-end re-identification pipeline. Input image is processed by YOLOv11 instance segmentation, anatomical labels are assigned by Hungarian matching to a canonical template, uncertain assignments are quality-filtered, each scute crop is embedded by a fine-tuned Vision Transformer, type-matched cosine similarities are aggregated by weighted geometric mean, and the Top-k identity is returned. Arrow annotations indicate inter-stage data signatures.

matching problem reduces to optimal assignment via the Hungarian algorithm [21]. We exploit the anatomically conserved 13-scute carapace as natural part boundaries with type-matched comparisons.

Open-set recognition. Standard closed-set protocols overstate deployment readiness because real systems must reject queries from unseen identities. Extreme value theory [28] and OpenMax [4] provide principled frameworks for novelty detection on top of trained recognition systems. We empirically quantify a more subtle bias: when the metric-learning loss *trains* on all K enrolled identities (as is standard with ArcFace), leave-N-out simulations leak identity information through the learned embedding, inflating reported open-set AUC.

3 Method

We formulate tortoise re-identification as a hierarchical matching problem exploiting the anatomically conserved carapace structure. The pipeline (Figure 2) decomposes the shell into its 13 constituent scutes and performs identity inference through aggregated part-level comparisons.

3.1 Scute Detection and Instance Segmentation

The first stage employs YOLOv11 [20] with a segmentation head, trained to detect four anatomically distinct classes: `scute`, `tortoise_shell`, `head`, and `nuchal`. The latter two serve as anatomical landmarks for subsequent geometric alignment. We trained YOLOv11x-seg on 856 manually annotated images, with standard augmentations (rotation $\pm 15°$, scaling 0.8–1.2×, horizontal flips, brightness/contrast/saturation), AdamW ($\eta = 10^{-3}$, cosine schedule), 300 epochs

with early stopping, initialized from COCO weights. At inference we keep detections with confidence ≥ 0.5, yielding 12.5 ± 1.6 scutes per image.

Crop extraction. Each scute instance is extracted by computing the minimum enclosing rectangle of its mask, applying a 15% margin expansion to preserve boundary context, and resizing to 384×384 pixels using Lanczos interpolation. We retain the unmasked crop: the Vision Transformer's global self-attention can exploit inter-scute boundary topology—identity-bearing geometric information that background masking would discard. A mask-area filter (< 500 px) and shell-overlap verification (centroid inside the detected shell) reject spurious detections.

3.2 Geometric Alignment and Scute Labeling

Each detected scute receives an anatomical label $\{\mathrm{V}_{1\text{–}5}, \mathrm{L}_{1\text{–}4}, \mathrm{R}_{1\text{–}4}\}$ via Hungarian matching [21] to a canonical template.

Coordinate normalization. Detected scute centers are transformed into a normalized shell frame in which the lateral axis spans $[-0.5, 0.5]$ and the forward axis spans $[0, 1]$. The frame is established by a priority cascade: (i) head + nuchal landmarks (forward vector points from shell centroid to their midpoint); (ii) shell ellipse fit (narrow end $\Rightarrow$ anterior); (iii) PCA fallback on scute centers. Only the head and nuchal are used as orientation landmarks; the tail is deliberately not detected, as it is often retracted or occluded and the head–nuchal axis already fixes the anterior–posterior direction.

Hungarian assignment. For n detected scutes the cost matrix $\mathbf{C} \in \mathbb{R}^{n \times 13}$ has entries

$$C_{ij} = \sqrt{\left(\frac{\mathrm{lat}_i - \mu_j^{\mathrm{lat}}}{\sigma_j^{\mathrm{lat}}}\right)^2 + \left(\frac{\mathrm{fwd}_i - \mu_j^{\mathrm{fwd}}}{\sigma_j^{\mathrm{fwd}}}\right)^2} + \lambda_{ij}, \tag{1}$$

where (μ_j, σ_j) are learned position statistics for scute type j, and λ_{ij} is a position-adaptive penalty ($\lambda = 8$–15 for cross-side, 5–10 for vertebral displacements). Assignments exceeding $\tau = 4.0$ Mahalanobis units (under our diagonal-Σ assumption, equivalent to standardized Euclidean distance) are rejected. The Hungarian algorithm solves $\pi^\star = \arg\min_\pi \sum_i C_{i,\pi(i)}$ in $O(n^3)$. Because the minimum penalty exceeds τ, anatomically implausible assignments are unconditionally rejected.

Smart Rotation Search. When the initial labeling confidence falls below $\gamma_{\mathrm{thresh}} = 0.75$ or fewer than 5 inliers exist, we test the four cardinal rotations and select $\theta^\star = \arg\max_\theta \left[10 \cdot n_{\mathrm{in}}(\theta) + \gamma(\theta)\right]$ with $\gamma \in [0, 0.95]$; the $10\times$ multiplier ensures inlier count dominates with confidence as tiebreaker.

3.3 Feature Extraction

We build on MegaDescriptor-L-384 [29], a Vision Transformer [15,36] that embeds each 384×384 scute crop into a 1536-d feature vector. Per crop, a two-layer projection head with LayerNorm, GELU, and dual dropout maps this vector to a compact 512-d embedding $\mathbf{e}$—the per-scute representation used in all subsequent matching:

$$\mathbf{z} = \mathrm{DO}_2\big(\mathrm{LN}_2(W_2\,\mathrm{DO}_1(\mathrm{GELU}(\mathrm{LN}_1(W_1\mathbf{f} + b_1))) + b_2)\big), \quad \mathbf{e} = \mathbf{z}/\|\mathbf{z}\|_2, \tag{2}$$

with $W_1 \in \mathbb{R}^{1024\times1536}$, $W_2 \in \mathbb{R}^{512\times1024}$, $\mathrm{DO}_1 = 0.4$, $\mathrm{DO}_2 = 0.2$.

Composite loss.

$$\mathcal{L} = \lambda_{\mathrm{ms}}\mathcal{L}_{\mathrm{MS}} + \lambda_{\mathrm{tri}}\mathcal{L}_{\mathrm{triplet}} + \lambda_{\mathrm{cen}}\mathcal{L}_{\mathrm{center}} + \lambda_{\mathrm{arc}}\mathcal{L}_{\mathrm{SubArc}} + \lambda_{\mathrm{scute}}\mathcal{L}_{\mathrm{scute}}, \tag{3}$$

combining Multi-Similarity [37], batch-hard triplet [19], center [39], SubCenter ArcFace [13] (K=2 sub-centers/identity, accommodating intra-class variance from wet/dry shell, mud, lighting), and a 13-way scute-type classification head as auxiliary regularizer (λ_{scute}=0.1). At inference the classification heads are discarded; only $\mathbf{e}$ is used for cosine retrieval.

Chirality preservation. Horizontal and vertical flips are strictly disabled: mirroring an L_1 scute creates a synthetic geometry identical to R_1, directly violating the structural prior. Rotations are restricted to $\pm20°$.

Three-stage training. The training schedule is the principal methodological contribution that enabled successful fine-tuning of the large backbone in our extreme few-shot regime ($\sim$2.4 images/identity). The loss weights $\lambda_\bullet$ refer to the composite loss in Equation (3). **Phase 1** (50 epochs): backbone frozen, projection head only ($\eta = 5\times10^{-4}$, OneCycleLR [33] with 10% warmup); **Phase 2** (*ArcFace-dominant*, 100 epochs): full network with differential learning rates (5×10^{-6} backbone, 5×10^{-5} head/loss parameters), λ_{arc}=1.0, λ_{ms}=1.0, λ_{tri}=0, ArcFace scale s=64, margin m=0.25; **Phase 3** (*pairwise recalibration*, 50 epochs): $\eta = 10^{-5}$, λ_{ms}=1.5, λ_{tri}=1.0, λ_{arc}=0.15, s=20, m=0.4 (λ_{cen}=0.2 throughout). Removing triplet loss during Phase 2 eliminates the gradient conflict between pairwise metric and classification objectives, allowing ArcFace's angular margin to dominate the optimization landscape; Section 4.5 confirms that the original single-stage configuration consistently degraded performance.

3.4 Similarity Matching and Inference Toolchain

For each scute type $s \in \mathcal{S}$ present in both query q and gallery entry t, we compute cosine similarity $\mathrm{sim}_s(q,t) = \mathbf{e}_q^s \cdot \mathbf{e}_t^s$ and aggregate via a weighted geometric mean:

$$S(q,t) = \exp\Big(\frac{\sum_{s\in\mathcal{S}_{q,t}} w_s \log \mathrm{sim}_s(q,t)}{\sum_{s\in\mathcal{S}_{q,t}} w_s}\Big). \tag{4}$$

Acting as a normalised product t-norm, the geometric mean is *hyper-sensitive* to severe local mismatches—a confident mismatch on any anatomical part substantially lowers identification confidence. We clamp similarities to $\max(\epsilon, \mathrm{sim}_s)$ with ϵ=10^{-6} before the logarithm.

Inference toolchain. Three components stack on top of Equation (4): (i) *Max-similarity gallery construction* retains all gallery exemplars per identity and keeps the maximum cosine similarity per scute type (robust to a single low-quality gallery photograph); (ii) *Query-side test-time augmentation* (Q-TTA) averages embeddings over 9 deterministic transforms of the query crop; (iii) *k-Reciprocal re-ranking* [40] (k_1=5, k_2=3, λ=0.9) is applied to non-augmented embeddings, since TTA-smoothed embeddings attenuate the local neighborhood structure that re-ranking exploits.

3.5 Open-Set Evaluation Protocols

Deployment requires handling truly unseen tortoises. We design *two complementary protocols*:

Primary — real open-set. 145 adult tortoise photographs from a separate field archive of the reserve, never seen by any component of the training pipeline. Hash-based duplicate detection and manual verification of ≥ 0.85-similarity matches removed 7 misfiled images, yielding 138 clean unknowns. The full 150-tortoise gallery is used; 63 known validation queries serve as positive controls. AUC is computed per score-derived feature.

Complementary — leave-N-out simulation. For each holdout $N \in \{2, 4, 6, 8\}$, N identities are randomly removed from the 34 evaluation tortoises, treated as "unknown" probes against the remaining gallery. Repeated 100× per setting. *Caveat:* all 34 identities were in ArcFace training, so this protocol upper-bounds, rather than matches, deployment performance.

4 Experiments and Results

4.1 Dataset and Evaluation Protocol

Photographs of *T. graeca* were collected at the Gazelle Valley Urban Nature Park, Jerusalem, over several months using consumer smartphones under natural lighting, during routine veterinary care visits requiring no additional handling beyond standard husbandry. Efrat Yagur, who is responsible for wildlife at the park, coordinated data collection. After merging duplicate identifiers and excluding one identity whose images failed quality filtering, the dataset contains 150 tortoises, 363 images and 4,560 scute crops (mean 2.42 images per tortoise; 12.6 scutes per image). We adopt a *smart split*: the 116 tortoises with ≤ 2 images contribute only to training, while the 34 tortoises with 3+ images split between training (68 images) and validation (63 valid queries). We report Top-k accuracy and Mean Reciprocal Rank with 95% identity-stratified percentile bootstrap CIs ($n = 1000$).

Closed-set protocol. The gallery contains the 34 evaluation tortoises, aggregated by the max-similarity rule of Section 3.4 unless otherwise stated. This reflects deployment in which the system identifies individuals from a known, enrolled population. The correct identity is always present.

4.2 Detection and Labeling

YOLOv11x-seg achieves 96.0% mAP@50 for box detection and 95.5% mAP@50 for masks across 1,878 instances on a held-out validation set (98.6% precision, 91.6% recall). Per-class mAP@50: scute 99.4%, shell 99.0%, head 96.4%, nuchal 94.9%. Hungarian alignment correctly labels 4,152/4,293 = 96.7% of scutes evaluated on 348 images, with 81.6% of images perfect; lateral scutes attain 98.1–98.2%, vertebral 95.5%. Importantly, evaluation exposed *no systematic confusions*—no left/right swaps and no adjacent-vertebral confusions—confirming that the Smart Rotation Search robustly resolves orientation ambiguity.

Table 1: Closed-set re-identification (63 queries, 34-tortoise gallery). McNemar's test computed on the 59 queries common to both methods: $\chi^2 = 21.3$, $p < 10^{-3}$; Cohen's $h = 0.88$ (large). 95% CIs in brackets.

Method	**Top-1 (%)**	**Top-1 macro**	**Top-3**	**Top-5**	**Top-10**	**MRR**
Whole-image	26.6	—	51.6	67.2	87.5	0.445
Scute-based	**66.7** [54.0–77.8]	**61.3**	**76.2**	**87.3**	**98.4**	**0.755**
Δ	+40.1	—	+24.6	+20.1	+10.9	+0.310

4.3 Re-Identification: Closed-Set Results

Table 1 contrasts the proposed scute-based method with a whole-image baseline trained with the same backbone, training protocol, and data splits.

The scute-based pipeline reaches 66.7% Top-1 and 98.4% Top-10 (62/63; the lone failure ranks 15); all 63 queries shared ≥ 8 scute types with their gold entry. **Embedding-space quality.** The learned embedding achieves $d' = 2.26$ (with $d' = (\mu_{\text{diff}} - \mu_{\text{same}}) / \sqrt{\frac{1}{2}(\sigma^2_{\text{diff}} + \sigma^2_{\text{same}})}$ on cosine-distance distributions), well above the $d' > 1.0$ biometric threshold [12]; same-identity pairs ($d_{\cos} = 0.213 \pm 0.137$) are well-separated from different-identity pairs (0.755 ± 0.310). All 13 scute types attain $d' \geq 1.9$ (range 1.90–2.40), explaining why uniform weighting matches optimized weights ($p = 0.48$). Figure 3 visualises identity clustering on the evaluation set.

Comparison with broader wildlife re-ID literature. Cross-taxa comparison is confounded by species distinctiveness, gallery size, and images-per-identity, so Table 2 is contextual, not a benchmark: established systems exploit more distinctive markings and far richer data (sea-turtle facial scales [9,8], ATRW tigers [22], humpback flukes [10]). We instead tackle subtler variation in *T. graeca* at 2.4 images/identity, where the Top-5 workflow (87.3%) is operationally comparable under far harder conditions.

4.4 Open-Set Novelty Detection

To address the realistic deployment scenario in which previously unseen tortoises arrive at the reserve, we evaluate against the 138 *genuinely unseen* adult tortoises described in Section 3.5. We test three score-derived novelty features summarising the top-k retrieval distribution: the *top-1 similarity*, the *score gap* (top-1 minus top-3), and the *top-5 score std* (std. of the five highest gallery similarities). Table 3 reports per-feature AUC: the headline *top-5 score std* attains real AUC = 0.600 (95% CI $[0.510, 0.672]$), while the leave-N-out simulation overestimates across all three. The gap arises because the simulation's smaller gallery has every identity seen by the ArcFace loss at training time, leaking identity information into the embedding. The external archive also carries richer session-, lighting-, and substrate-variation than the enrolled gallery, so the measured gap

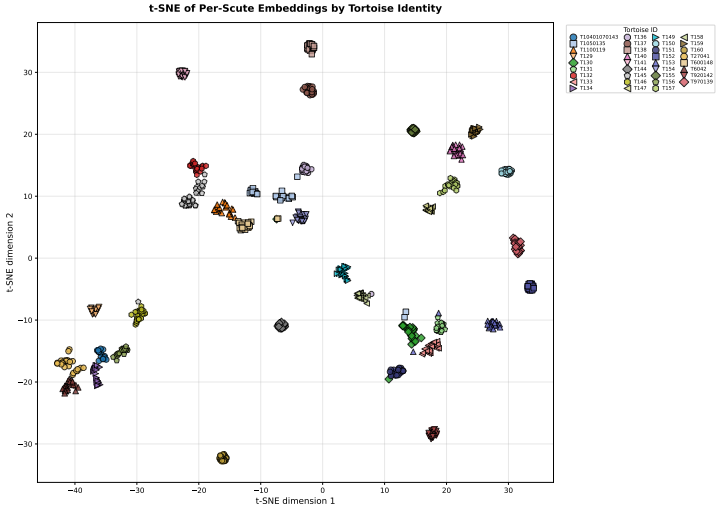


Fig. 3: t-SNE projection of per-scute embeddings for the 34 evaluation tortoises. Each tortoise is represented by a unique (color, marker) combination, with markers cycling through 10 shapes to remain distinguishable under colorblind conditions and grayscale printing. Tight intra-identity clusters confirm $d' = 2.26$; adjacent clusters correspond to the doppelganger pairs identified in Section 4.6. PCA pre-reduction to 50 dim, perplexity 30, cosine metric.

reflects both this protocol bias and genuinely harder field appearance. Three gallery identities account for 60% of false acceptances at threshold 0.85, mirroring the *confuser-hub* structure observed in closed-set errors. Although above chance and statistically robust under Bonferroni correction, the real-AUC values are too low for autonomous classification: the deployed system therefore returns the top-5 candidates together with the *top*-5 *score std* as an open-set indicator (a low std flags the query as likely unknown), and a human operator validates the match or enrols a new identity.

4.5 Consolidated Ablation

Table 4 consolidates the principal ablation factors (backbone fine-tuning, data availability) and the inference-time toolchain components (max-similarity gallery, Q-TTA, k-reciprocal re-ranking) into a single page-budget-friendly view. Fine-tuning under the three-stage protocol provides the largest single improvement (+27.0 pp; McNemar $\chi^2 = 13.5$, $p = 2.4\times10^{-4}$, Cohen's $h = 0.55$). Performance scales consistently with images-per-tortoise (+23.2 pp from 3 to $\geq$ 5 images), directly informing data-collection priorities. On the inference side, alternative scoring strategies (V_5 exclusion, costal-heavy weights, drop-worst aggregation, frozen+fine-tuned ensembles) all match—but never exceed—the 66.7% baseline, demonstrating that the residual accuracy ceiling resides in the learned represen-

Table 2: Contextual comparison with established wildlife re-ID systems. Numbers are author-reported under each system's own evaluation protocol; differences in gallery construction, query selection, and species-level visual distinctiveness make these figures **contextual rather than directly comparable**. The Top-5 workflow row reflects our recommended human-in-the-loop deployment.

System	**Species**	**|Gallery|**	**Imgs/ID**	**Top-1**
HotSpotter [11]	Plains zebra	20	$\sim$10	94%
Carter et al. [9]	Green sea turtle	180	$\sim$5	$>$90%
ATRW [22]	Amur tiger	182	$\sim$100	89.4%
SealID [24]	Saimaa ringed seal	57	$\sim$20	$\sim$46–77%
Cheeseman et al. [10]	Humpback whale	42,000	—	97–99%
Ours (Top-1)	*T. graeca*	34	2.4	**66.7**%
Ours (Top-5 workflow)	*T. graeca*	34	2.4	**87.3**%

Table 3: Open-set discrimination by score-derived feature: real-archive AUC (138 unseen tortoises) against the leave-N-out simulation. The simulation overestimates because every gallery identity was in ArcFace training (DeLong one-sided $p < 10^{-7}$ for score gap and $p < 10^{-10}$ for the headline top-5 score std; the +11% top-1 similarity gap is not significant, $p = 0.16$).

Feature	**Real AUC**	**Sim AUC**	**Overest.**
Top-1 similarity	0.584	0.649	+11%
Score gap (top-1 − top-3)	0.591	0.881	+49%
Top-5 score std	**0.600**	0.930	+55%

tations rather than the scoring stage. The lone scoring-stage component that yields a positive (though under-powered) effect is k-reciprocal re-ranking, which captures gallery-level structure beyond pairwise cosine.

4.6 Stage-Wise Error Propagation

The pipeline is sequential, prompting the legitimate concern that errors compound across stages. We characterise the cascade outputs directly: YOLOv11 localises 12.5 ± 1.6 scutes per image (range 3–18), with 95.1% of images yielding ≥ 10 detections; the Hungarian step labels 96.7% of detected scutes correctly, with no systematic left/right or adjacent-vertebral confusions (Section 4.2); and all 63 evaluation queries reach the matching stage with ≥ 8 common scute types versus their gallery entry. Per-query inspection of a representative near-miss (true identity ranked second, behind a visually similar gallery individual) shows all 13 scutes successfully detected and correctly labeled—the misidentification originates in the embedding/matching stage. The aggregator's ≥ 2 common-types requirement and the quality filter (mask area > 500 px, shell-overlap verification)

Table 4: Consolidated ablation. Top block: training factors. Middle block: data availability (per-bucket n in parentheses); **the ≥ 4 and ≥ 5 buckets are small (n=15 and n=7) and should be read as directional rather than precise rates**. Bottom block: inference-time scoring components (additive on top of *max-sim + Q-TTA* unless stated). All numbers on the same 63 queries / 34-gallery closed-set protocol. Best reachable Top-1: 69.8% with re-ranking.

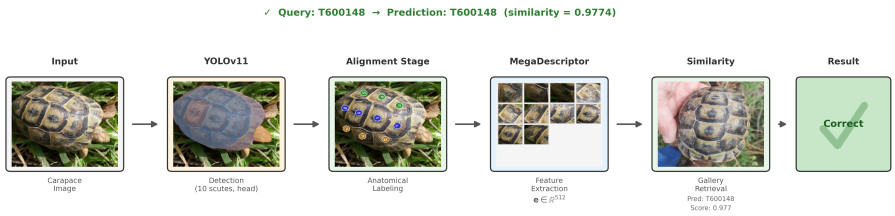
Configuration		**Top-1 (%)**	**Top-5 (%)**	**MRR**
Backbone fine-tuning				
Frozen (Phase 1 only)		39.7	76.2	0.552
Fine-tuned (Phases 1–3, **ours**)		**66.7**	**87.3**	**0.755**
Data availability (images/tortoise)				
$=3$	(n=48)	62.5	75.0	0.716
≥ 4	(n=15)	66.7	93.3	0.778
≥ 5	(n=7)	85.7	100.0	0.929
Inference toolchain (additive on max-sim + Q-TTA baseline)				
Baseline (max-sim, Q-TTA, uniform w)		66.7	87.3	0.755
+ Weights / drop-worst / frozen+FT ensemble		66.7	87.3	0.755
+ k-Reciprocal re-ranking (Q-noTTA)		**69.8**	84.1	**0.770**

form a structural firewall against upstream noise, so detection and alignment contribute only weakly to the residual error budget.

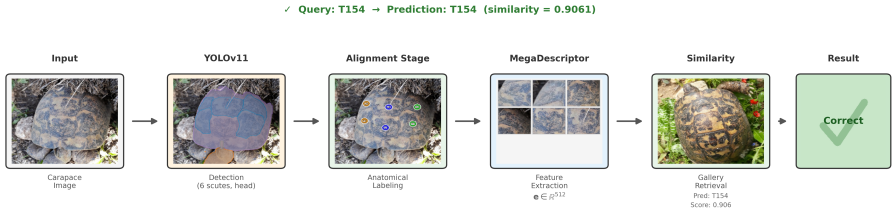
Confusion structure. Of the 21 misidentifications, 6 are *near misses* (rank 2–3, mean error rank 5.05). Two *confuser hubs* attract 52% of errors from multiple unrelated queries; k-reciprocal re-ranking rescues 4 queries from ranks 2–6 while losing 1, a net +2 over the max-sim + Q-TTA baseline (Top-1 66.7% $\rightarrow$ 69.8%), but the hubs persist, indicating a perceptual rather than modeling limit. The same hub structure surfaces *independently* in the real open-set experiment (Section 4.4)—a cross-protocol agreement motivating the deployment recommendation to monitor hub identities.

4.7 Field-Condition Robustness

Robustness evidence comes from the 145-image field-archive of the real open-set protocol (Section 3.5): consumer-smartphone images from surveys under uncontrolled lighting and substrate, in sessions separate from training. All 145 candidates traversed the detection $\rightarrow$ alignment $\rightarrow$ embedding $\rightarrow$ scoring pipeline and 138 produced valid scores; the 7 removals were duplicates or misfiled knowns, *none* a detection or alignment failure. Edge cases ($> 60°$ oblique views, heavy occlusion, ambiguous orientation; 0.9% of images) are caught by the quality gate ($\gamma < 0.5$, < 6 scutes, or shell area $< 20\%$) before matching. Three mechanisms drive this: the anatomical consistency check rejects implausible labelings; Sub-Center ArcFace's $K = 2$ sub-centers model wet/dry/lit appearance; and the geometric mean absorbs partial occlusion via common-subset matching (Figure 4b:



(a) High-confidence Top-1 (S=0.977).



(b) Partial-visibility Top-1, 6 scutes (S=0.906).

Fig. 4: Correct identifications across image quality and partial-visibility; the geometric-mean aggregator (Equation (4)) degrades gracefully to fewer scutes.

correct Top-1, 6 scutes). Partial visibility thus degrades the score gracefully, not catastrophically; a scute-count ablation is the next probe.

4.8 Gallery-Size Scalability

To project toward the full reserve population, we sampled $K \in \{5, 8, 10, 15, 20, 25, 30, 34\}$ identities from the 34 tortoises (100 repetitions each). Top-1 degrades logarithmically, from 87.6±12.8% at K=5 to 66.7% at K=34 (≈ -6.7 pp per doubling), projecting $\approx$55% at K=200 (directional). Top-5 stays $\geq$87.3% throughout, supporting the human-in-the-loop workflow at the $\approx$200-individual scale.

5 Discussion and Conclusion

What worked. A modest, anatomically-grounded prior—the 13-scute template—lifts Top-1 accuracy by +40.1 pp over a whole-image baseline at the same backbone and training budget. Among architectural choices, the three-stage ArcFace-dominant training schedule is the single most consequential ingredient (+27.0 pp from frozen embeddings), enabling fine-tuning at $\sim$2.4 images per identity.
What did not. Alternative scoring strategies (weight optimisation, drop-worst aggregation, ensemble construction) all match the 66.7% baseline, localising the residual ceiling to the embedding stage; the +3.1 pp gain from k-reciprocal reranking is promising at our scale but awaits validation on a larger gallery. Future

work therefore prioritises representation-side avenues (synthetic augmentation, micro-class formulation at Tortoise × Scute-Type) over scoring-side refinements. Further reviewer-suggested directions include stronger feature-stage augmentation, a canonical 3D shell model that normalises viewpoint and recovers occluded scutes, and cross-dataset validation on green-sea-turtle scales [9].

Open-set deployment. The headline finding—that leave-N-out overestimates real open-set AUC by up to 55% whenever the metric-learning loss trains on every identity—recurs across the wildlife re-ID literature. We therefore report a real-archive AUC alongside the simulation (Section 3.5) and treat the gap as an honest measure of deployment readiness.

Limitations and conclusion. The dataset is small (150 identities, mean 2.4 images) but uniquely curated under realistic field conditions and rigorously deduplicated against an external archive—a strength for ecological realism that bounds the absolute accuracy figures reported here. External validation at a second reserve and longer temporal spans remain key to establishing generality, and the fixed constants (loss weights, λ, γ, τ) await a systematic sensitivity analysis. The anatomically-decomposed pipeline turns *T. graeca* re-identification from a 26.6% Top-1 problem into a deployable Top-5 workflow at 87.3% recall, while honestly characterising both the closed-set ceiling and the open-set gap. Code, weights, and the annotated dataset will be released.

References

1. Alibardi, L.: Proliferation in the epidermis of chelonians and growth of the horny scutes. Journal of Morphology **265**(1), 52–69 (2005)
2. Arzoumanian, Z., Holmberg, J., Norman, B.: An astronomical pattern-matching algorithm for computer-aided identification of whale sharks *Rhincodon typus*. Journal of Applied Ecology **42**(6), 999–1011 (2005)
3. Beausoleil, N.J., Mellor, D.J., Stafford, K.J.: Methods for marking New Zealand wildlife: Amphibians, reptiles and marine mammals. Tech. rep., Department of Conservation, Wellington, New Zealand (2004)
4. Bendale, A., Boult, T.E.: Towards open set deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1563–1572 (2016)
5. Benn, A.L., McLelland, D.J., Whittaker, A.L.: A review of welfare assessment methods in reptiles, and preliminary application of the Welfare Quality® protocol to the pygmy blue-tongue skink, *Tiliqua adelaidensis*, using animal-based measures. Animals **9**(1), 27 (2019)
6. Bergin, D., Nijman, V.: Trade in spur-thighed tortoises *Testudo graeca* in Morocco: Volumes, value and variation between markets. Oryx **53**(3), 403–406 (2019)
7. Bogucki, R., Cygan, M., Khan, C.B., Klimek, M., Milczek, J.K., Mucha, M.: Applying deep learning to right whale photo identification. Conservation Biology **33**(3), 676–684 (2019)
8. Carpentier, A.S., Jean, C., Barret, M., Chassagneux, A., Ciccione, S.: Stability of facial scale patterns on green sea turtles *Chelonia mydas* over time: A validation for the use of a photo-identification method. Journal of Experimental Marine Biology and Ecology **476**, 15–21 (2016)

9. Carter, S.J.B., Bell, I.P., Miller, J.J., Gash, P.P.: Automated marine turtle photograph identification using artificial neural networks, with application to green turtles. Journal of Experimental Marine Biology and Ecology **452**, 105–110 (2014)
10. Cheeseman, T., Southerland, K., Park, J., Olio, M., Flynn, K., Calambokidis, J., et al.: Advanced image recognition: A fully automated, high-accuracy photo-identification matching system for humpback whales. Mammalian Biology **102**, 915–929 (2022)
11. Crall, J.P., Stewart, C.V., Berger-Wolf, T.Y., Rubenstein, D.I., Sundaresan, S.R.: HotSpotter—patterned species instance recognition. In: IEEE Workshop on Applications of Computer Vision (WACV). pp. 230–237 (2013)
12. Daugman, J.: Biometric decision landscapes. Technical Report TR482, University of Cambridge Computer Laboratory (2000), establishes $d' > 1.0$ as adequate discriminability threshold for biometric systems
13. Deng, J., Guo, J., Liu, T., Gong, M., Zafeiriou, S.: Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In: European Conference on Computer Vision (ECCV). pp. 741–757 (2020)
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
16. Dunbar, S.G., Ito, H.E., Bahjri, K., Dehom, S., Salinas, L.: Recognition of juvenile hawksbill sea turtles (*Eretmochelys imbricata*) through face scale digitization and automated searching. Endangered Species Research **44**, 65–75 (2021)
17. Faber, M., Strubbe, D., Lens, L.: Specimen identification from time-series photographs using plastron morphometry in *Testudo graeca ibera*. Herpetological Journal **23**, 185–189 (2013)
18. Gould, J., Beranek, C., Madani, G.: Dragon detectives: Citizen science confirms photo-ID as an effective tool for monitoring an endangered reptile. Wildlife Research **51**(1), WR23036 (2024)
19. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
20. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLO11. `https://github.com/ultralytics/ultralytics` (2024), version 11.0
21. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955)
22. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: ATRW: A benchmark for Amur tiger re-identification in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2590–2598 (2020)
23. Moustakas-Verho, J.E., Zimm, R., Cebra-Thomas, J., Lempiäinen, N.K., Kallonen, A., Könönen, K., Lehti, M.S., Werneburg, I., Gilbert, S.F.: The origin and loss of periodic patterning in the turtle shell. Development **141**(15), 3033–3039 (2014)
24. Nepovinnykh, E., Chelak, I., Eerola, T., Kälviäinen, H., Hautala, M., Kunnasranta, M., et al.: SealID: Saimaa ringed seal re-identification dataset. Sensors **22**(19), 7602 (2022)
25. Pérez, I., Giménez, A., Anadón, J.D., Martínez, M., Esteve, M.A.: Non-commercial collection of spur-thighed tortoises (*Testudo graeca graeca*): A cultural problem in southeast Spain. Biological Conservation **118**(2), 149–160 (2004)

26. Plummer, M.V., Ferner, J.W.: Marking reptiles. In: McDiarmid, R.W., Foster, M.S., Guyer, C., Gibbons, J.W., Chernoff, N. (eds.) Reptile Biodiversity: Standard Methods for Inventory and Monitoring, pp. 143–150. University of California Press, Berkeley, CA (2012)
27. Rodríguez-Caro, R.C., Graciá, E., Anadón, J.D., Giménez, A.: Hidden threats to persistence: Changes in population structure can affect well-preserved spur-thighed tortoise populations. Herpetologica **81**(1) (2025)
28. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(7), 1757–1772 (2013)
29. Schneider, S., Saravanan, M., Roth, H.: MegaDescriptor: A billion-parameter descriptor for wildlife re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024)
30. Schneider, S., Taylor, G.W., Kremer, S.C.: Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods in Ecology and Evolution **10**(4), 461–470 (2019)
31. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015)
32. de Silva, N.I., Parham, J., Holmberg, J., Halloran, K.M., Khan, K.A., Schaaf, A.V., Wijewickrama, S., Fernando, S.: Feasibility of using convolutional neural networks for individual-identification of wild Asian elephants. Mammalian Biology **102**, 1–14 (2022)
33. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. vol. 11006, p. 1100612 (2019)
34. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling. In: European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
35. Tortoise & Freshwater Turtle Specialist Group: *Testudo graeca*. The IUCN Red List of Threatened Species (1996), e.T21646A9305693
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
37. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5022–5030 (2019)
38. Warwick, C., Arena, P., Lindley, S., Jessop, M., Steedman, C.: Assessing reptile welfare using behavioural criteria. In Practice **35**(3), 123–131 (2013)
39. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision (ECCV). pp. 499–515 (2016)
40. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1318–1327 (2017)