Supporting Amphibian Migration Tracking with State-of-the-Art Image Processing

Marie Ceccaldi[1], Alexis Deseure[1], Akhil Nampally[1], Nils Ragot[1], Laura Tomokiyo[1], Lena Collet[2], Emilie Busson[3], Alain Morand[3], and Cédric Pradalier[1,4]

[1] Georgia Institute of Technology, USA (in alphabetical order)
[2] Ecosphere, France, https://ecosphere.fr/
[3] Cerema, France, https://www.cerema.fr/en
[4] CNRS IRL2958 GeorgiaTech-CNRS

Abstract. Amphibians are the most imperiled group of vertebrates, with habitat loss and fragmentation identified as the primary causes of their population declines. Existing assessments of mitigation structures are scarce and often superficial, creating an urgent need for a rigorous framework to measure their effectiveness and to refine underpass design. Camera traps are well established as an effective tool for wildlife monitoring, and within the broader push to standardize protocols for underpass assessment, deep learning models for automated image analysis constitute a significant step forward for amphibian conservation. Here, we investigate how well foundation models perform at zero-shot detection of animal presence in camera-trap images. In addition to simple presence–absence filtering, we examine whether these models can also facilitate more detailed downstream tasks, such as species identification and body pose estimation.

Keywords: Natural image processing · Ecological monitoring.

1 Introduction

Amphibians are the most threatened vertebrate taxa, with habitat loss and fragmentation proven to be the main drivers of population decline. The complex biphasic life cycle of most European species requires, for example, migrating between hibernation forests and reproductive water ponds, resulting in massive road mortality. Authors of [4] reveals an estimation of 25 to 50 million adult amphibians killed by roads in one reproduction season. To mitigate road impact and restore landscape connectivity, over- and underpasses are installed [5]. However, these mitigation measures remain largely under-evaluated, particularly for amphibians, even though they have been implemented for several decades (40 years in France) as concluded by Testud et al. [14]. The few existing studies of mitigation structure evaluation lack scientific rigor, impeding the generation of pertinent results, repeatability across sites, and the implementation of conservation actions [7]. There is an urgent need to develop a robust methodology to evaluate effectiveness and to improve underpass design.

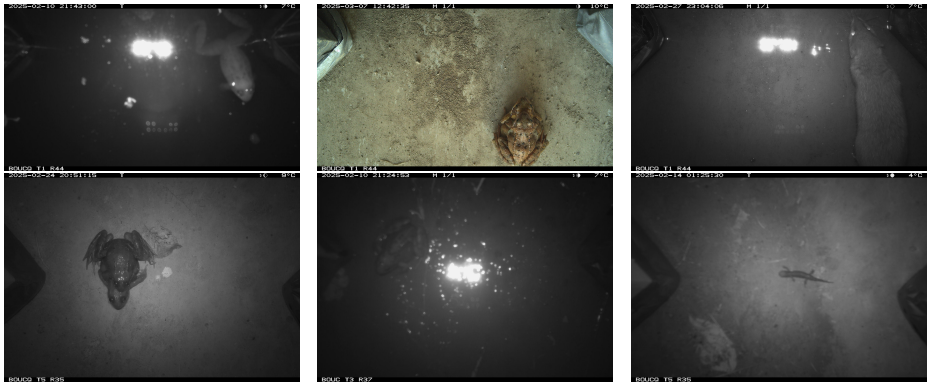

Fig. 1. Sample images from Dataset 2025

Camera traps have proven efficient for animal monitoring. Compared to mammals, monitoring amphibians requires an automatic preset trigger that generates a large number of images in a short period of time [2]. For one season of migration and eight monitored tunnels, processing and labeling 1.2 million images was estimated to take more than 17 days. In the global ecological effort to develop a standardized protocol for underpass evaluation, deep learning models for camera-trap imagery represent a major advance in amphibian conservation.

Our primary objective is to evaluate recently released (end of 2025) foundation models' effectiveness for zero-shot *animal presence detection* in camera-trap imagery, and to leverage it to generate pseudo-annotations to train a lightweight model without requiring manual pixel-level labels. Beyond binary filtering, we assess whether these model capabilities can be utilized to support more detailed downstream analysis, including species identification and body pose detection.

2 State of the Art

Camera traps provide a non-invasive way to document animals in their natural habitats, making them attractive for ecological monitoring [9]. Manual filtering and annotation of imagery is time-consuming, and machine learning techniques offer great potential for automation [8].

The field of **object detection** was revolutionized by the unified regression approach of YOLO, which accelerates detection by performing it in a single stage compared to two-stage approaches. For camera-trap images, the classic MegaDetector pipeline integrates "class-agnostic" filtering, which isolates generic animal presence to eliminate false positives before any specific classification via YOLO [2]. Recently, Vision-Language Models (VLMs) have paved the way for zero-shot classification. CLIP aligns visual and textual representations, an approach that BioCLIP [13] specializes in capturing biological taxonomic nuances without requiring supervised training.

In contrast to bounding box–based approaches, segmentation methods provide pixel-level representations of objects. Semantic segmentation models such as SegNet and U-Net assign class labels at the pixel level, while instance segmentation further distinguishes between individual objects. However, they require datasets annotated with pixel-level masks, which are considerably more expensive to obtain than bounding box annotations.

The introduction of foundation models for segmentation, such as **Segment Anything Model (SAM)** [3] and self-supervised visual models like DINO [12], which reduce reliance on fully annotated datasets while enabling high-quality mask generation, has triggered a paradigm shift. SAM supports zero-shot learning via "promptable" segmentation. Trained on the massive SA-1B dataset, it can generate high-quality masks from points, boxes, or text. The subsequent iteration, SAM 2 [11], extended these capabilities to video. The ability of SAM 3 [3] to segment all instances of a concept in text enables its use in a fully automated pipeline. While SAM enables high-quality segmentation through prompting, it still relies on external inputs; in contrast, **the DINO family of models** learns general visual representations that can support downstream tasks like clustering and classification. DINOv2 [10] demonstrates that self-supervised learning produces high-performance visual features when trained on a massive, automatically curated dataset. DINOv3 [12] provides updates, including a larger patch size, constant scheduling, and a much larger teacher model, and outperforms DINOv2 on challenging visual tasks.

Binary classification in ecology often involves distinguishing a target species from environmental "noise," such as shadows, vegetation, or infrastructure. While many models rely on ImageNet-1k for pretraining, that dataset has a significant taxonomic bias; for example, it contains over 100 dog breeds but fewer than five amphibian-related classes. This lack of diversity means that high performance on standard benchmarks does not always translate to success in scientific or field-based vision tasks. Despite these limitations, DINOv2 has shown strong zero-shot results on specialized ecological data, including iWildCam [1]. Recent work has demonstrated that vision foundation models such as DINOv2 provide strong representations for **image segmentation** [1, 6]. We extend this line of research to ecological imagery of toads, moving from foreground–background segmentation toward anatomical segmentation (e.g., limb and head regions).

Species classification is a fundamental task in ecological computer vision, enabling automated biodiversity monitoring from visual data collected in unconstrained natural environments. Given the strong transferability reported for vision foundation models such as DINOv2, it is relevant to reproduce and evaluate existing classification-related results on our own challenging toad dataset. Huo et al. suggest that DINOv2 embeddings are robust to ecological variability and could support species classification on our noisy dataset [6]. Finally, zero-shot wildlife sorting using vision transformers shows that self-supervised embeddings can organize wildlife images with high semantic coherence without supervised training [1].

Table 1. Summary of datasets used in this study.

Dataset	Total images	Positive	Negative
Dataset 2025	16975	8479	8480
Subset 436	436	218	218
Dataset 2026	1461	307	1154
Dataset 2026-extended	30700	307	30393

Pose estimation aims to localize anatomical keypoints of organisms, enabling fine-grained analysis of posture and behavior. In ecological contexts, this task is particularly challenging due to occlusions, low image quality, and pose and scale variability. Transformer-based approaches such as ViTPose [15] have demonstrated strong performance by leveraging global context and scalable architectures. These models outperform convolutional baselines across several benchmarks. However, they still depend on large annotated datasets, motivating the exploration of foundation models such as DINO.

Overall, recent advances in ecological computer vision reflect a shift from task-specific, fully supervised pipelines toward flexible foundation models (DINO, SAM) that can operate under limited annotation and environmental variability, enabling zero-shot segmentation and transferable visual representations across diverse domains. However, effectively adapting these models to ecological tasks such as handling multi-animal scenes, domain-specific variation, and rare species remains an open challenge.

3 Methodology

3.1 Datasets

Our *Toad* dataset comprises 1.25 million wildlife images captured as part of amphibian conservation efforts in and around Metz, France. In a collaboration with Cerema, wildlife capture devices were installed in 8 culverts. Image capture occurs between February and April (aligning with the toad breeding season) at 15s intervals. Cameras are Reconyx Hyperfire devices fitted with lenses to focus at 40cm, capturing color RGB images during daylight and infrared-illuminated grayscale images at night (2048x1536 pixels). Images have been labeled by experts for features including animal presence, species, sex, and direction of travel, with 11320 of 1246329 images containing at least one individual. Figure 1 gives an example of images in the dataset.

Our dataset presents challenges common to other wildlife capture tasks, including sparsity (most images contain no individuals), environmental conditions (flooding, condensation, lighting), and partial/multiple/overlapping individuals in the frame.

Several image subsets were annotated and used for the experiments reported in Table 2. **Dataset 2025** contains 16,975 manually classified images (8,479 positive, 8,480 negative). It was split into 95% training and 5% validation. For

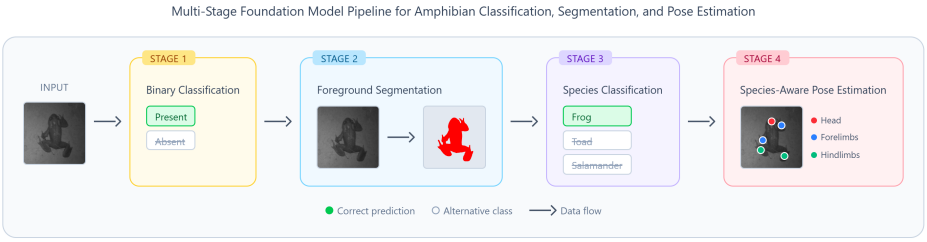

Fig. 2. Data flow of the proposed pipeline, demonstrating how foundation models are applied to dataset images to achieve key amphibian tracking tasks.

training segmentation models without manual masks at scale, pseudo-ground-truth masks were generated for the Dataset 2025 using SAM3.

Subset 436 was curated by selecting 218 positive images with clearly identifiable subjects, often near the image center, and 218 random negative samples. This Dataset is used for evaluation only. Binary masks were manually produced using the Segment Anything UI.

To evaluate year-to-year generalization, we introduce **Dataset 2026**, which contains 1,461 images (307 positive, 1,154 negative) sampled randomly from unseen 2026 imagery. Dataset 2026 is fully manually annotated: binary labels are provided for all images, and a single binary segmentation mask is produced manually for each positive image.

Finally, to assess performance under realistic class imbalance where false positives are costly, we introduce **Dataset 2026-extended**. This dataset reuses the 307 positive images (and corresponding manual masks) from Dataset 2026 and adds 30,393 automatically selected negative images to approximate a 1% positive rate.

3.2 Pipeline overview

Our overarching objective is to investigate how state-of-the-art foundation models can be leveraged for key ecological tasks of amphibian migration tracking. We specifically address the following tasks:

Presence/Absence Binary Classification: detecting whether an individual is in the image.

Foreground Segmentation: detecting the specific area of the image that contains an individual.

Species Classification: identifying the individual at a given taxonomic level (e.g., frog vs. toad, or amphibian vs. mammal).

Pose Estimation: determining the location of key body parts (e.g., limbs, head, tail) once an individual has been detected and identified at the species level.

The data flow from the dataset images to the specific products of our pipeline is illustrated in Fig. 2.

3.3 Binary mask extraction

The mask extraction stage is a binary segmentation task using Intersection over Union (IoU) as primary metric. To assess practical viability, we record inference runtime, measured as the average time required for a model to process a single image. Our binary segmentation stage compares three pipelines: SAM3 with text prompts, lightweight segmentation models finetuned on the masks generated by SAM3, and a binary segmentation network using DINO features. SAM3 was prompted with a predefined set of text prompts – ['animal', 'creature', 'amphibian', 'batrachian', 'reptile', 'bird', 'rat', 'mouse', 'frog', 'toad', 'salamander', 'newt', 'lizard'] – leading to redundant mask predictions fused together using a logical OR operation. From the masks generated by SAM3, we fine-tuned several lightweight semantic segmentation models: SegFormer (mit-b0, mit-b2, mit-b3), and DeepLabV3 with ResNet-50 and ResNet-101 backbones.

Beyond supervised models, we implement an unsupervised segmentation pipeline using the foundation models DINOv2 and DINOv3. These models produce a high-dimensional feature vector per image patch ($6,144$ dimensions for DINOv2; $4,096$ for DINOv3). These high-dimensional vectors are compressed via IPCA and clustered using K-Means. After identifying the "amphibian" cluster, patches with features too far from it are assigned to the image background.

3.4 Detection

The main objective of the detection stage is to determine whether an image contains an object of interest (primarily amphibians, though other animals use the tunnels as well). As a result, the detection stage classifies images as either containing at least one animal or being empty. The performances of this stage are evaluated using standard binary classification metrics: **sensitivity** (assesses what percentage of animals were detected) and **specificity** (assesses what percentage of empty images were filtered out). With 1.2 million images per season, processing time is also an important metric. Our detection pipeline compares several algorithms. YOLO26, as an image-level classifier, is used as a baseline. Our segmentation models (SAM3, finetuned models and DINO) are used as detectors, treating an image as non-empty if the binary mask is sufficiently large. The specific threshold is evaluated on the validation set to maximize the detection F1-score. Furthermore, we evaluate BioCLIP [13] to assess the capabilities of foundation models specifically tailored for ecological data. BioCLIP is applied in a zero-shot classification setting, utilizing the identical set of taxonomic text prompts previously established for SAM3. Finally, all models were evaluated on Dataset 2025 (held-out test split), Dataset 2026, and Dataset 2026-extended.

3.5 Species identification

The goal of species identification experiments is to investigate how well DINO features capture taxonomically meaningful characteristics. Classifying images as amphibians or not, and classifying amphibians by type (tailed vs. tailless), could

simplify labeling and enable order-level investigation without additional manual annotation. Preliminary experiments found that a 3-way KNN classifier representing tailed/tailless/other classes, trained on the Dataset 2025 (test) dataset and tested on a small, curated dataset from Dataset 2026, achieved very high accuracy. Experiments are therefore extended to test the classifier's robustness on the larger Dataset 2025 (train). Images are first cropped to the binary mask. Features are generated from the smallest DINO models. The classifier assigns each test image the class most common among its five closest training embeddings in feature space.

3.6 Body part identification

The objective of this stage is to evaluate whether DINO embeddings can be used to localize body parts without supervised training, in line with the segmentation and clustering experiments. Our approach does not rely on training a dedicated pose estimation model. Instead, we formulate body part localization as a similarity matching problem in the DINO feature space. The pipeline is composed of two stages. First, a small reference set of annotated images is used. We manually annotate approximately 50 images containing a single fully visible toad, each with 5 keypoints corresponding to predefined anatomical landmarks. Each image is processed using a DINO backbone (ViT-based), producing an average reference feature for each landmark. Second, inference is performed on a set of multi-toad images. For each test image, DINO features are extracted, and the closest reference features (below a threshold) are used to identify the predicted landmarks.

Performance is evaluated using the normalized mean pixel error (NME) and the percentage of correct keypoints for a given threshold (PCK@t).

4 Results

4.1 Binary Classification

Table 2 shows the binary classification results of all models on the three evaluation datasets. Performance on the less challenging Subset 436 was near-optimal across most architectures, except DeepLabV3 and BioCLIP failed to transfer their pre-training to this specific domain. When evaluated on the unseen 2026 imagery, YOLO26s exhibited poor generalization. While it maintained near-perfect sensitivity, its specificity plummeted to 0.56. In contrast, segmentation-based models and vision transformers proved much more robust. The zero-shot SAM3 baseline performed remarkably well (0.86 specificity, 0.90 sensitivity), validating its use as a reliable pseudo-label generator. DINOv2 (zero-shot) excels in specificity (0.99), effectively minimizing false positives, and outperforms its successor, DINOv3, on this task. Ultimately, given the extreme rarity of animal occurrences in real-world deployment, sensitivity remains the absolute priority to prevent data loss. Because SegFormer mit-b3 captures nearly all animal presences at a

lower processing cost while still successfully eliminating 93% of negative images on the extended dataset, it stands out as the most viable model for drastically reducing the manual review burden for biologists.

Table 2. Binary Classification and Segmentation Performance Across Subset 436, Dataset 2026, and Dataset 2026-Extended. (SPEC.: Specificity, SENS.: Sensitivity)

Model	Subset 436			2026			2026-extended		Time
	SPEC.	SENS.	IoU	SPEC.	SENS.	IoU	SPEC.	SENS.	[ms]
YOLO26s	**1.00**	**1.00**	-	0.56	**0.99**	-	0.87	**0.99**	66.8
SAM3 (zero-shot)	0.98	0.99	**0.91**	0.86	0.90	0.81	0.91	0.90	239.1
SegFormer mit-b0	**1.00**	0.99	0.89	0.97	0.90	0.78	0.92	0.90	47.1
SegFormer mit-b2	0.99	**1.00**	**0.91**	0.95	0.98	0.86	0.92	0.98	55.0
SegFormer mit-b3	0.98	**1.00**	**0.91**	0.88	**0.99**	**0.87**	0.93	**0.99**	58.4
DeepLabV3 R50	0.34	0.98	0.61	0.41	0.94	0.40	0.66	0.94	50.9
DeepLabV3 R101	0.57	0.93	0.60	0.48	0.75	0.33	0.65	0.75	59.2
BioCLIP (zero-shot)	0.53	0.69	-	0.10	0.88	-	0.13	0.88	**40.5**
DINOv2 (zero-shot)	0.98	1.00	0.37	**0.99**	0.86	0.42	**0.99**	0.86	298.0
DINOv3 (zero-shot)	**1.00**	0.79	0.62	0.98	0.85	0.57	0.89	0.85	72.0

4.2 Segmentation

In terms of zero-shot segmentation (Table 2, IoU), SAM3 provides a far superior mask precision than DINO models that are inherently penalized by the "blocky" nature of the 14×14 patch architecture. However, a qualitative analysis revealed that even if the masks are too large, they still serve the same purpose: facilitating the downstream manual review process. Therefore, the DINOv3 approach, 3 times faster at inference than SAM3 (Table 2, Time) and not requiring manual prompt selection, should be prioritized in a large-scale zero-shot setting. However, if more time can be allocated to train a specialized model, then using SAM3 to supervise the training of a Segformer-mit B3 yields the best results, both on mask precision (0.91 IoU on 2025 and 0.87 IoU on 2026), binary classification, and processing time. Figure 3 illustrates the difference in prediction of the different models.

(a) Original (b) GT (c) SAM3 (d) SegFormer (e) DINOv3

Fig. 3. Qualitative comparison of segmentation performance. The panels display (a) the original input image, (b) the ground truth mask, and the predicted masks generated by (c) SAM3, (d) SegFormer-B3, and (e) DINOv3.

4.3 Species Identification

This section evaluates whether DINO embeddings can support reliable species classification, specifically a 3-way classification of tailed amphibian/tailless amphibian/other animal. We compare the performance of the DINOv2 and DINOv3 models on both a curated single-animal dataset and a larger, uncurated dataset. The uncurated set is separated into wet and dry environment conditions (night images only). Results are shown in Table 3

Table 3. Species classification performance using segmented-mask crops and DINO embeddings. Accuracy is computed only for successfully masked images. Mask Success indicates the proportion of original images for which a usable mask was generated.

Dataset	Model	Masked Images	Mask Success	Accuracy	Macro F1	Weighted F1
Curated	v2	154	**83.2%**	0.97	0.89	0.97
Curated	v3	154	**83.2%**	0.96	0.83	0.96
2025 dry	v2	3783	82.0%	0.95	0.90	0.95
2025 dry	v3	3783	82.0%	0.95	0.91	0.95
2025 wet	v2	2745	82.2%	**0.98**	0.91	**0.98**
2025 wet	v3	2745	82.2%	**0.98**	**0.93**	**0.98**

Both models achieved a satisfactory performance. Although DINOv3 is the newer model, the experimental results do not show a clear advantage, except for the Macro F1 score on the smaller dataset. Some of the DINOv3 innovations were motivated by model collapse, which was not a problem in our case. With lower-resolution images and already high accuracy, our system may not be able to leverage the v3 architecture. Overall, species classification experiments demonstrate that DINO embeddings can be used for order-level identification with high accuracy across environmental conditions, without the need to train a specific encoder.

4.4 Pose Estimation

We evaluate our approach in a zero-shot setting ($n = 51$, no fine-tuning) under a multi-instance protocol on 35 annotated frames (79 toad instances). The pipeline achieves encouraging performances, with a low NME (0.071) and high PCK@0.10 and PCK@0.20 scores (0.771 and 0.971). This indicates that body-part identification is possible, which would help determine whether an amphibian is heading towards the reproduction pond or returning.

5 Conclusion

Ecology and artificial intelligence are two disciplines that largely coalesce in the recognition of animal patterns. Still, the extent of image recognition concerns

almost exclusively mammals, while amphibians remain understudied, partly because of the complexity of using camera traps on this taxon. This investigation relies on testing various detection models in camera-trap imagery, yielding conclusive results and filling a gap in knowledge of underpass evaluation after construction. This tremendous step forward in amphibians' conservation will allow us to enhance knowledge of their utilization by amphibians and hopefully improve conception and take conservation actions. In the future, this deep learning tool can be integrated into an ongoing standardized protocol for evaluating underpass effectiveness.

Acknowledgments. We would like to thank the department of Meurthe-et-Moselle for the annual financial support that has enabled us to conduct monitoring for 2 months over the last 3 years. We are also grateful to Justine Colin from the same department and to Johan Claus from Lorraine Regional Natural Park for their regular field support in changing SD cards and batteries.

References

1. Beery, S., Agarwal, A., Cole, E., Birodkar, V.: The iwildcam 2021 competition dataset. arXiv preprint arXiv:2105.03494 (2021)
2. Beery, S., Morris, D., Yang, S.: Efficient pipeline for camera trap image review. arXiv preprint arXiv:1907.06772 (2019)
3. Carion, N., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
4. Cerema: Amphibiens et dispositifs de franchissement des infrastructures de transport terrestre. In: Collection: Connaissances (2019), sous la dir. de Morand A. & Carsignol J
5. Cerema: Les passages à faune. préserver et restaurer les continuités écologiques avec les infrastructures linéaires de transport. In: Collection: Références (2021), sous la dir. de Nowicki, F
6. Huo, J., Muthivhi, M., van Zyl, T.L., Gustafsson, F.: Nearest-class mean and logits agreement for wildlife open-set recognition. ArXiv **abs/2510.17338** (2025)
7. Lesbarreres, D., Fahrig, L.: Measures to reduce population fragmentation by roads: what has worked and how do we know? Trends in ecology & evolution **27**(7) (2012)
8. Norouzzadeh, M.S., et al.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences **115**(25) (2018)
9. O'Connell, A.F., Nichols, J.D., Karanth, K.U.: Camera traps in animal ecology: methods and analyses, vol. 271. Springer (2011)
10. Oquab, M., et al.: Dinov2: Learning robust visual features without supervision (2024)
11. Ravi, N., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
12. Siméoni, O., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
13. Stevens, S., et al.: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2024)
14. Testud, G., Miaud, C.: From effects of linear transport infrastructures on amphibians to mitigation measures. In: Reptiles and amphibians. IntechOpen (2018)
15. Xu, Y., et al.: Vitpose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)