

Research Article

M2V-SnakeNet: Multi-View Learning with Frozen DINOv2 for Snake Species and Venom Classification

Hamza Maqsood Khan ^{*}, Wakeel Ahmad , and Syed M. Adnan 

Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan;
e-mail : 24-ms-cs-02@students.uettaxila.edu.pk; wakeel.ahmad@uettaxila.edu.pk; syed.adnan@uettaxila.edu.pk

* Corresponding Author : Hamza Maqsood Khan

Abstract: Accurate identification of snake species, particularly venom status, is important for emergency medical response and biodiversity monitoring. In practice, this task is challenging because visually similar species may differ in venom status, photographs may capture only part of the animal, and curated datasets are often limited in size. Existing image-based approaches commonly rely on fine-tuning deep networks on small regional datasets, while partial observation, probability calibration, and the relationship between species identity and venom status remain insufficiently addressed. This study proposes M2V-SnakeNet, which couples a frozen self-supervised vision foundation model, DINOv2, with presence-aware multi-view decomposition, cross-view attention fusion, and a prototypical metric-based head, together with a hierarchical venom head linked to the species distribution through a consistency constraint. The framework is designed to represent partial observations without fabricating absent views, while restricting task-specific training to 9.5 of 95.2 million parameters. On a balanced and leakage-audited dataset of seven Indian snake species comprising 1,776 images from a single geographic region, M2V-SnakeNet achieves $93.4\% \pm 1.5\%$ species accuracy and $96.9\% \pm 1.4\%$ venom accuracy under five-fold cross-validation. On the fixed held-out split, a five-seed ensemble reaches 97.4% species accuracy and 98.5% venom accuracy, with 99.3% recall for the venomous class. Ablation results identify the frozen DINOv2 encoder as the largest performance contributor, improving species accuracy by 17.4 percentage points over an equal-sized ImageNet-supervised backbone, while learned fusion and multi-view decomposition provide additional gains. External meta-pretraining on a large snake corpus does not improve performance in this setting. These findings indicate that frozen foundation features combined with lightweight task-specific adaptation can provide a competitive low-data approach to joint snake species and venom-status classification, while the limited species and geographic coverage remains an important constraint on generalization.

Keywords: Calibration; Computer vision; Deep learning; Metric-based meta-learning; Multi-view learning; Snake species classification; Venom-status classification; Vision foundation models.

Received: August, 1st 2026

Revised: August, 24th 2026

Accepted: August, 27th 2026

Published: September, 5th 2026

Curr. Ver.: September, 5th 2026



Copyright: © 2026 by the authors.
Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

The World Health Organization recognizes snakebite envenoming as a neglected tropical disease. It remains a serious and under-resourced public-health problem in many tropical and rural regions, where the timely distinction between venomous and non-venomous species can influence clinical decisions [1]. Expert morphological identification, based principally on head and body scalation, morphometry, and coloration [2], can be accurate but is time-consuming, subjective, and dependent on trained herpetologists who may be unavailable in remote clinics or field settings. Automated image-based recognition could therefore provide useful support for first-response assessment and biodiversity surveillance.

Deep learning has substantially advanced image-based snake recognition. Early studies have fine-tuned convolutional neural networks or compared them with classical machine-learning pipelines on regional species sets [3]–[5]. A related line of work directly addresses the

binary venomous/non-venomous decision using transfer learning [6], [7], while larger regional studies extend species classification to dozens of classes [8], [9]. At larger scale, the SnakeCLEF challenges have motivated fine-grained and long-tailed recognition approaches based on residual or transformer backbones with auxiliary location metadata [10]–[12], gradient-boosting ensembles [13], [14], and additional transfer-learning strategies [15], [16]. Other studies address detection in cluttered field imagery [17] or explore self-supervised and frozen-feature representations under label scarcity [18], [19], whereas cross-continental benchmarks emphasize geographic coverage and report lower performance under broader distributional variation [20]–[22].

Despite these advances, several limitations remain relevant to low-data snake recognition. Fine-tuned convolutional backbones are relatively simple and well established, but updating tens of millions of parameters on small regional datasets increases the risk of overfitting and provides limited insight into run-to-run variability when evaluation relies on a single split. Vision transformers and self-supervised foundation models offer stronger transferable representations, yet many existing approaches still fine-tune these models end-to-end, potentially reducing the benefit of their pretrained representations when labeled data are scarce. Detection and part-based methods can localize snakes effectively, but localization alone does not resolve venom status, while conventional cropping does not explicitly account for whether a particular body region is actually visible. Metric-based meta-learning is well suited to few-shot classification and requires relatively few task-specific parameters, but its application to snake recognition remains limited and has not, to our knowledge, been combined with a frozen vision foundation encoder.

These limitations become more important under realistic image conditions. Snake photographs may contain visually similar venomous species and harmless mimics, substantial variation in pose and background, and partial observations in which only the head, a coil, or another body region is visible. At the same time, datasets that are sufficiently large, balanced, and clean for robust model development are difficult to obtain. Three issues are therefore particularly relevant. First, partial observation is rarely modeled explicitly, even though conventional whole-image classification implicitly assumes that the available image provides sufficient evidence for the entire animal. Second, predictive calibration is seldom reported, despite its importance when model outputs are used to support uncertainty-aware decisions. Third, species identity and venom status are commonly treated as separate prediction tasks, leaving an opportunity to exploit their natural relationship in a joint model.

To address these issues, this study proposes M2V-SnakeNet, a low-data framework that combines a frozen self-supervised vision foundation model with a presence-aware representation of complementary image regions. Rather than adapting the backbone to the small target dataset, the framework learns a compact task-specific representation on top of fixed pretrained features. Multiple views are used to capture complementary evidence when different parts of the snake are visible, while their fusion allows the model to account for differences in the information provided by each view. A prototypical classification head is adopted to support the low-data setting, and a hierarchical venom head links the venom decision to the predicted species distribution through a consistency constraint. This design is intended to address the limitations identified above while keeping the number of trainable parameters substantially smaller than that of a fully fine-tuned backbone. The main contributions of this work are as follows:

- A presence-aware multi-view decomposition that represents each image using a foreground view and automatically discovered salient-region views, together with a presence mask that accommodates partial observations without fabricating unavailable crops.
- A low-data recognition framework, M2V-SnakeNet, that combines a frozen DINOv2 encoder with cross-view attention fusion and a prototypical meta-learning head, while training only 9.5 of the 95.2 million model parameters.
- A hierarchical species–venom prediction head that jointly addresses fine-grained species identification and the venomous/non-venomous decision through a consistency constraint between the direct venom prediction and the venom probability implied by the species distribution.
- A systematic evaluation incorporating five-fold cross-validation, five-seed ensembling, calibration analysis, per-class error analysis, and component ablation. The experiments also examine external SnakeCLEF meta-pretraining and report a negative result, showing

that additional meta-pretraining does not improve performance in the present low-data setting.

The remainder of this paper is organized as follows. Section 2 reviews related work on image-based snake recognition, venom-status classification, foundation models, multi-view learning, and metric-based meta-learning. Section 3 describes the proposed M2V-SnakeNet framework, including the dataset, multi-view decomposition, frozen DINOv2 encoder, cross-view fusion, classification heads, training objective, and evaluation protocol. Section 4 presents the experimental results, including training behavior, main performance, ablation studies, error analysis, calibration, and cross-validation. Section 5 discusses the findings in relation to representative prior studies, computational considerations, and generalization limitations. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related Work

2.1. Image-Based Snake Identification

Early image-based snake identification combined handcrafted descriptors with classical classifiers. Rajabizadeh and Rezghi [5] compared holistic pipelines based on k-nearest neighbors, support vector machines, and logistic regression using PCA- or LDA-derived features with convolutional networks on six Iranian species, reporting 84% accuracy for LDA with SVM and 93.16% for fine-tuned MobileNetV2. Deep transfer learning subsequently became a dominant approach for regional snake recognition. Patel et al. [3] applied deep recognition to Galápagos species and reported 75% accuracy, while Vasmatkar et al. [4] developed an early CNN-based identification system. For binary venomous/non-venomous classification, Proga et al. [6] combined convolutional models with augmentation and ensembling, achieving approximately 91.3% accuracy under five-fold cross-validation on 1,766 images, while Joshi et al. [7] applied MobileNetV2 transfer learning to the same binary task.

These studies demonstrate the effectiveness of deep visual representations, but they also highlight the sensitivity of reported performance to the underlying dataset and evaluation setting. For example, Firnando et al. [23] compared several pretrained architectures across three rice-leaf disease datasets and reported 99.53%, 58.94%, and 90.00% accuracy for their best model across the respective datasets. Although this study concerns a different application domain, the large variation illustrates that performance cannot be attributed to architecture alone. Accordingly, the comparisons presented later in Section 5 are intended to provide contextual positioning rather than a direct ranking across heterogeneous datasets and protocols. More recent regional snake studies report higher performance on larger species sets: Ahmed et al. [8] achieved 97.09% with VGG16 across 45 species, whereas Iguernane et al. [9] reported 92.23% accuracy and 93.67% F1 using EfficientNetB0 for snakebite-management applications.

2.2. Detection, Self-Supervision, and Large-Scale Recognition

A parallel research direction focuses on detection and self-supervised representation learning rather than whole-image classification. Wang et al. [17] proposed Snake-DETR, a lightweight detector for fine-grained snake detection in complex natural environments, achieving 95.23% mAP@0.5. Although detection can facilitate localization in cluttered scenes, it does not directly address species or venom-status classification. Li et al. [18] adapted SimCLR for small-scale snake image classification and reported 97.48% accuracy, demonstrating the potential of self-supervised objectives when labeled data are limited. However, the resulting model still relies on fine-tuning after pretraining.

At larger scale, the SnakeCLEF challenges have focused on fine-grained and long-tailed recognition across hundreds to thousands of species [10], [22]. On the 772-species SnakeCLEF 2021 edition, Chamidullin et al. [12] fine-tuned an ensemble of ResNeSt, ResNeXt, and ResNet backbones with country-based prediction filtering, placing third in the challenge with 91.6% accuracy and an F1 score of 0.830. Bloch and Friedrich [11] combined EfficientNet and Vision Transformer backbones with a location-based prior and achieved 82.88% macro country-F1. On the same benchmark, Desingu et al. [15] used a frozen CNN feature extractor with a gradient-boosting classifier, while Kalinathan et al. [16] employed a fine-tuned ResNet, reporting 42.96% and approximately 75% test accuracy, respectively. The substantial variation across these approaches further suggests that performance depends not

only on backbone choice but also on training strategy, auxiliary information, and the difficulty of the evaluation setting.

The larger SnakeCLEF 2022 benchmark, comprising 1,572 species, further illustrates the difficulty of long-tailed generalization. Palaniappan et al. [13] combined ResNeXt101 and EfficientNet-B6 features with an XGBoost ensemble and obtained 51.39% macro-F1 on the held-out test set despite 87.13% validation accuracy. On SnakeCLEF 2023, Hu et al. [14], combined seesaw loss with a CLIP-derived metadata prior and achieved 91.31% on the combined challenge metric. Beyond the SnakeCLEF series, Durso et al. [20] established a supervised benchmark across 45 species with accuracies ranging from 68% to 79%, while Bolon et al. [21] evaluated a worldwide model and reported mean accuracy above 90% across 88% of countries, with weaker performance in several under-represented regions.

Most directly related to the present study, Miyaguchi et al. [19] used frozen DINOv2 features with a linear classifier for SnakeCLEF 2024, covering 182,261 images. Their results demonstrate that self-supervised foundation features can transfer effectively to snake recognition at large scale. However, a frozen foundation encoder with a linear probe leaves the representation largely under-exploited for a low-data regional setting: it does not explicitly model complementary views, adapt the classifier to few-shot structure, or couple species prediction with venom status. M2V-SnakeNet builds on this direction by retaining the frozen encoder while introducing a prototypical head, presence-aware multi-view decomposition with attention-based fusion, and a hierarchical species–venom prediction mechanism.

2.3. Synthesis of the Research Gap

The literature reviewed above reveals several complementary research directions in image-based snake recognition. To consolidate their methodological strengths, limitations, and remaining opportunities, the representative approaches are synthesized in Table 1.

Table 1. Synthesis of representative approaches for image-based snake recognition.

Approach	Studies	Strengths	Limitations	Gap left open
CNN transfer learning on curated regional datasets	[3]–[5], [8], [9]	Simple, established, and effective on small regional datasets.	Often requires full backbone fine-tuning; limited analysis of variance and calibration.	Low-data adaptation and species–venom coupling remain insufficiently addressed.
Binary venom-status classification	[6], [7]	Directly targets the clinically relevant venomous/non-venomous decision; lightweight models can be efficient.	Discards fine-grained species information; limited dataset scale and evaluation analysis.	Species information is not used to constrain the venom decision.
Self-supervised pretraining followed by fine-tuning	[18]	Reduces dependence on labeled data and provides strong training; evaluation is based on a relatively small dataset.	Backbone is still fine-tuned after pretraining; evaluation is based on a relatively small dataset.	The potential of frozen self-supervised features with a task-specific low-data head remains underexplored.
Frozen foundation features with a shallow classifier	[15], [19]	Avoids backbone training and reduces task-specific computational cost.	Relies on shallow classifiers and does not explicitly model complementary views or task relationships.	Frozen foundation features can potentially be exploited more effectively through adaptive fusion and metric-based classification.
Large-scale fine-grained challenge systems	[10]–[14], [16], [22]	Address large species spaces and long-tailed recognition; ensembles and metadata can improve performance.	Often computationally intensive and may rely on location metadata; performance varies substantially across the long tail.	A compact low-data framework without inference-time geographic metadata remains desirable for fixed regional species sets.
Detection and localization	[17]	Improves localization in cluttered field imagery.	Addresses detection rather than species or venom classification; evaluated using detection metrics.	Partial observation is not explicitly modeled as missing or complementary visual evidence.
Global benchmarking and deployment studies	[20], [21]	Provide broader geographic coverage and evidence across diverse regions.	Require large training corpora and reveal uneven performance across under-represented regions.	Calibration and safety-relevant venomous-class recall remain insufficiently characterized.

As summarized in Table 1, prior studies address individual aspects of the problem effectively, but these capabilities remain largely fragmented. Regional CNN-based approaches provide strong

recognition accuracy but generally rely on backbone adaptation, whereas frozen foundation-model approaches reduce the training burden but typically use shallow classifiers. Self-supervised learning has shown promise under limited labels, yet its combination with frozen representations and task-specific metric learning remains insufficiently explored in snake recognition. Similarly, detection and localization methods can address visual complexity but do not directly model partial observations as complementary evidence for classification.

More importantly, species recognition and venom-status prediction are generally treated as separate tasks, despite their biological relationship and the asymmetric consequences of venom-class errors. The literature also provides limited analysis of calibration and safety-relevant class-specific performance. These observations define the specific gap addressed in this study: whether a frozen foundation representation, explicit multi-view handling of partial observations, adaptive view fusion, and joint species–venom prediction can be integrated effectively within a low-data framework. M2V-SnakeNet is therefore designed to examine this combination under a controlled regional setting. The objective is not to claim direct superiority over studies evaluated on different datasets or protocols, but to determine whether these complementary design choices provide measurable benefits while limiting task-specific adaptation of the foundation encoder.

3. Proposed Method

This section describes the proposed M2V-SnakeNet framework for low-data snake species and venom-status classification. The framework is designed around four methodological considerations identified from the limitations of existing approaches: retaining transferable visual representations without fine-tuning a large backbone, representing complementary evidence from partially observed images, adaptively integrating views with different information content, and jointly modeling species identity and venom status. The overall architecture is illustrated in Figure 1.

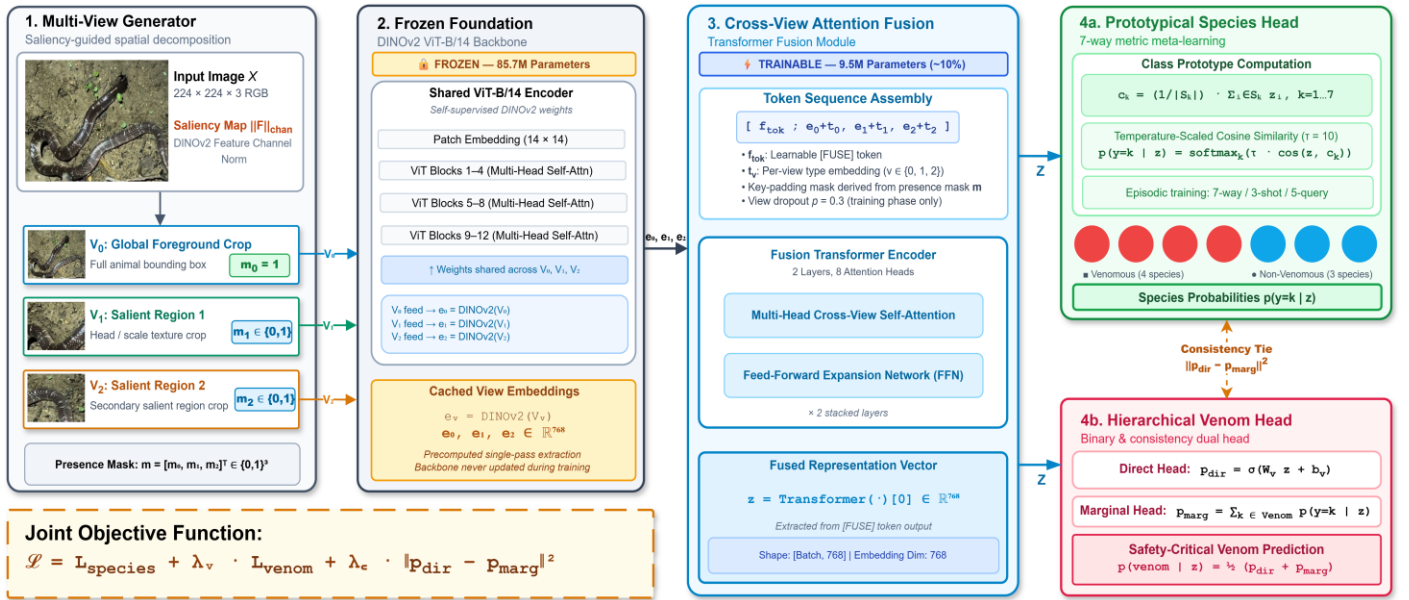


Figure 1. Overview of the M2V-SnakeNet architecture.

3.1. Framework Overview

M2V-SnakeNet processes each input image through four main stages. First, a presence-aware multi-view generator produces an always-present foreground view and up to two salient-region views, together with a binary presence mask indicating which region views are available. Second, a frozen DINOv2 Vision Transformer encodes each available view into a fixed-dimensional visual representation. Third, a cross-view attention-fusion module integrates the available view embeddings while accounting for their view types and presence. Finally, the fused representation is passed to a prototypical species head and a hierarchical venom head to obtain the species and venom-status predictions. Formally, for an input image I , the multi-view generator produces

$$V(I) = \{v_0, v_1, v_2\}, \quad \mathbf{m} \in \{0,1\}^3, \quad (1)$$

where v_0 denotes the always-present global foreground view, v_1 and v_2 denote optional salient-region views, and $\mathbf{m} = [m_0, m_1, m_2]$ records their availability, with $m_0 = 1$. Each available view is encoded by the frozen DINOv2 encoder:

$$\mathbf{e}_i = f_\theta(v_i) \in \mathbb{R}^d, \quad d = 768, \quad m_i = 1, \quad (2)$$

where f_θ is the frozen DINOv2 ViT-B/14 encoder. The available embeddings are then combined by the cross-view fusion module:

$$\mathbf{z} = g_\psi(\{\mathbf{e}_i: m_i = 1\}), \quad (3)$$

where g_ψ denotes the trainable attention-fusion module. The resulting representation is used to estimate the species distribution and venom status:

$$p(y|\mathbf{z}), \quad p_{\text{venom}}(\text{venom}|\mathbf{z}) = h_\omega(\mathbf{z}), \quad (4)$$

where h_ω represents the combined species and venom prediction heads described in Section 3.6.

Only the fusion parameters ψ and prediction-head parameters ω are updated during training; the DINOv2 parameters θ remain frozen throughout. This design substantially reduces the number of task-specific parameters while retaining the pretrained visual representation. The subsequent subsections describe the dataset, multi-view generation, frozen encoder, cross-view fusion, prediction heads, and training procedure in detail.

3.2. Dataset

The dataset contains 1,776 images of seven snake species found in the Indian subcontinent, curated from two publicly available sources: research-grade observations from iNaturalist [24] and selected images from the publicly released SnakeCLEF dataset [22]. For the iNaturalist images, research-grade status was used as an initial species-label check before manual curation. The final dataset comprises four venomous species—common krait, Indian cobra, king cobra, and Russell's viper—and three non-venomous species—chequered keelback, Indian rock python, and oriental rat snake.

The dataset is balanced at the venom-status level, with 888 venomous and 888 non-venomous images, and is divided into training, validation, and test sets containing 1,241, 264, and 271 images, respectively. The per-species distribution across these splits is reported in Table 2.

Table 2. Dataset composition per species and split.

Species	Venom status	Train	Validation	Test	Total
Chequered keelback	Non-venomous	207	44	45	296
Common krait	Venomous	155	33	34	222
Indian cobra	Venomous	155	33	34	222
Indian rock python	Non-venomous	207	44	45	296
King cobra	Venomous	155	33	34	222
Oriental rat snake	Non-venomous	207	44	45	296
Russell's viper	Venomous	155	33	34	222
Total	—	1,241	264	271	1,776

The complete curated dataset, including the 1,776 images, species and venom-status labels, and the exact training, validation, and test partitions, has been released publicly through Kaggle. The corresponding dataset description and documentation are provided with the release. Making the partition explicit is important because it allows the reported experiments to be reproduced using the same image assignments rather than relying only on the total dataset size.

Figure 2 provides representative examples of the dataset and illustrates the variation in appearance, pose, and visible body regions across the seven species. Because image duplication across splits can lead to overly optimistic performance estimates, the dataset was additionally audited for near-duplicate images. A 64-bit perceptual hash based on the discrete cosine transform was computed for each image. Two images were flagged as near-duplicates when their Hamming distance satisfied $d_H(h_i, h_j) \leq \tau$, $\tau = 8$.

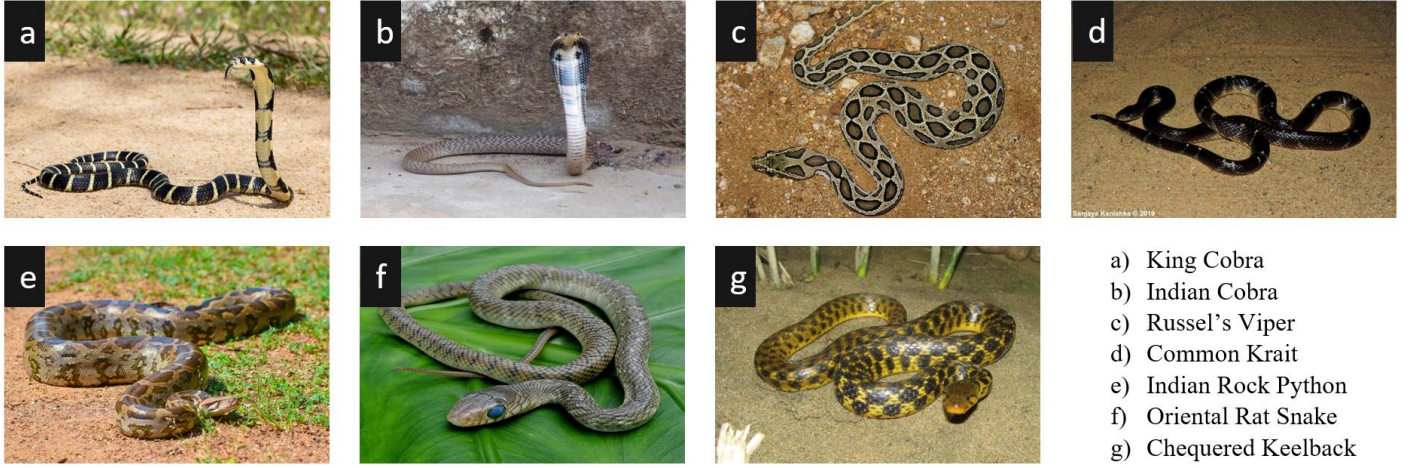


Figure 2. Representative images from the seven species in the curated dataset.

The criterion was validated by construction: a re-encoded and resized copy of the same image produced a Hamming distance of zero, whereas unrelated images produced substantially larger distances. No cross-split near-duplicate pairs were detected under this criterion, indicating that no detectable image-level duplication was present across the predefined splits. Before encoding, all images were converted to RGB and resized using aspect-preserving padding. The resulting dataset provides a controlled low-data setting for evaluating the proposed framework while preserving a clear distinction between training, validation, and held-out test data. At the same time, its restriction to seven species from a single geographic region should be considered when interpreting the generalization of the reported results; broader external validation is addressed as a limitation in Section 5.

3.3. Presence-Aware Multi-View Generation

To accommodate partial observations, each input image is represented using a global foreground view and up to two additional salient-region views. Rather than assuming that predefined anatomical regions are always visible, the region views are identified from the visual evidence available in each image. This allows the representation to retain complementary local information while explicitly distinguishing between present and unavailable views. For each image, a saliency map is computed from the channel-wise norm of a lightweight convolutional feature map:

$$S(x, y) = \|\phi(x, y)\|_2, \quad \phi(x, y) \in \mathbb{R}^C \quad (5)$$

A foreground bounding box is obtained by thresholding the saliency map, $B_0 = \text{bbox}\{(x, y): S(x, y) \geq \tau_0\}$, and the corresponding crop is used as the always-present global view v_0 . Two additional region views are then selected iteratively by locating the highest-saliency point outside the regions already covered:

$$(x^*, y^*) = \arg \max_{(x, y) \notin \text{covered}} S(x, y) \quad (6)$$

A fixed-size window centered at this location defines each candidate region view $v_i, i \in \{1, 2\}$. A binary presence flag m_i is assigned to each region view. The flag is set to zero when no sufficiently salient region remains in the uncovered area, i.e., when $\max_{(x, y) \notin \text{covered}} S(x, y) < \tau_0$. Thus, the model does not assume that all region views exist for every image. An image containing only a visible head, coil, or other partial portion of the snake can therefore be

represented using the evidence that is actually available. Absent views are subsequently excluded from the attention computation through the presence mask described in Section 3.5.

This design differs from conventional fixed-region cropping because the additional views are determined by image content rather than predefined anatomy. It also avoids fabricating crops when insufficient visual evidence is available. Partial observation is a recognized challenge in other fine-grained recognition domains; for example, Muhamada et al. [25] reviewed machine- and deep-learning approaches for occluded face recognition, where only part of the subject may be visible. In the present setting, the missing information is often anatomical rather than caused by an external occluder, motivating an explicit representation of view presence.

3.4. Frozen DINOv2 Encoder

Each available view is encoded using the frozen DINOv2 Vision Transformer (ViT-B/14) [26]: $\mathbf{e}_i = f_{\theta}(v_i) \in \mathbb{R}^d$, $d = 768$. The encoder parameters θ remain fixed throughout training. This design retains the pretrained visual representation while avoiding task-specific updates to the large backbone on the relatively small target dataset. In this setting, freezing the encoder also limits the number of parameters that must be optimized and reduces the risk of over-adaptation to the target training set. The experimental evidence for the relative contribution of the frozen DINOv2 representation is presented in Section 4.6.

Because the encoder is frozen, the embedding produced for a given view is deterministic under a fixed preprocessing pipeline. The resulting view embeddings can therefore be computed once and cached before training the task-specific modules, substantially reducing repeated encoder computation during training. The total model contains approximately 95.2 million parameters, of which only 9.5 million belong to the trainable fusion and prediction modules. Thus, approximately 10% of the overall parameters receive gradient updates, while the remaining parameters of the DINOv2 encoder remain frozen. The parameter allocation is summarized in Table 3.

Table 3. Model configuration and computational characteristics of M2V-SnakeNet.

Property	Value
Backbone	DINOv2 ViT-B/14 (frozen)
Frozen backbone parameters	85.72 M
Trainable fusion and head parameters	9.46 M ($\approx 10\%$)
Views per image	3
Input resolution per view	224×224
Fusion Transformer	2 layers, 8 attention heads
Region-view dropout	0.3
Episodic configuration	7-way, 3-shot, 5-query
Ensemble seeds	5
Head training time	261 s for all five seeds (single T4 GPU)

This separation between a fixed foundation representation and a compact trainable component is particularly relevant to the low-data setting considered here. Rather than relying on extensive backbone adaptation, the subsequent modules learn how to combine and classify the pretrained representations for the target species and venom-status tasks.

3.5. Cross-View Attention Fusion

The available view embeddings are integrated using a shallow Transformer encoder [27]. A learnable fusion token $\mathbf{f}_{\text{tok}}$ is prepended to the sequence, while a learnable type embedding distinguishes the global foreground view from the salient-region views: $\mathbf{e}'_i = \mathbf{e}_i + \mathbf{t}_{\text{type}(i)}$. This allows the fusion module to distinguish the role of each view while still learning interactions among their visual representations. The input sequence is formed as $[\mathbf{f}_{\text{tok}}; \mathbf{e}'_0; \mathbf{e}'_1; \mathbf{e}'_2]$. A key-padding mask derived from the presence vector prevents absent region views from contributing to the attention computation. The attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V, \quad (7)$$

where the additive mask is

$$M_{:,i} = \begin{cases} 0, & m_i = 1 \\ -\infty, & m_i = 0 \end{cases} \quad (8)$$

The output corresponding to the fusion-token position is used as the integrated representation:

$$\mathbf{z} = \text{Transformer}([\mathbf{f}_{\text{tok}}; \mathbf{e}'_0; \mathbf{e}'_1; \mathbf{e}'_2])[0] \quad (9)$$

Consequently, the fusion module can integrate only the views that are actually present while retaining information about their respective roles. This is preferable to treating all views as equally informative, particularly when one region contains limited or ambiguous evidence.

During training, additional view dropout is applied to the two region views. Each region view is independently masked with probability $p_{\text{drop}} = 0.3$, whereas the global foreground view is always retained. This regularizes the fusion module and exposes it to configurations in which only a subset of the available views is observed, complementing the explicit presence modeling introduced in Section 3.3. The resulting fused representation $\mathbf{z}$ is subsequently provided to the prototypical species head and hierarchical venom head described in Section 3.6. Importantly, the attention module is not intended to replace the pretrained representation; rather, it provides a lightweight mechanism for adapting and integrating complementary view-level evidence while keeping the DINOv2 encoder fixed.

3.6. Prototypical Species Head and Hierarchical Venom Head

Species classification uses a prototypical metric-learning head [28] to adapt the frozen visual representation to the low-data setting. During episodic training, a support set and a query set are sampled for each episode. For each species k , its prototype is computed as the mean embedding of the corresponding support examples:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{i \in S_k} \mathbf{z}_i \quad (10)$$

where S_k denotes the support examples belonging to species k . A query embedding $\mathbf{z}$ is classified according to its temperature-scaled cosine similarity to each class prototype:

$$p(y = k|\mathbf{z}) = \text{softmax}_k(\tau \cos(\mathbf{z}, \mathbf{c}_k)) \quad (11)$$

where,

$$\cos(\mathbf{z}, \mathbf{c}_k) = \frac{\mathbf{z}^\top \mathbf{c}_k}{\|\mathbf{z}\| \|\mathbf{c}_k\|} \quad (12)$$

Here, τ is a logit scaling factor; its value is set to 10 in the experiments. At inference, class prototypes are computed from the training embeddings, and test samples are classified using the same prototype-based decision rule.

Venom status is modeled jointly with species identity through two complementary estimates. The first is a direct binary prediction from a trainable venom head, $p_{\text{dir}}(\text{venom}|\mathbf{z})$, while the second is obtained by marginalizing the species distribution over the set V of venomous species:

$$p_{\text{marg}}(\text{venom}|\mathbf{z}) = \sum_{k \in V} p(y = k|\mathbf{z}) \quad (13)$$

The final venom probability is defined as the average of these two estimates:

$$p_{\text{venom}} = \frac{1}{2}(p_{\text{dir}} + p_{\text{marg}}) \quad (14)$$

A consistency term further encourages the direct and species-derived venom probabilities to agree. This coupling allows species information to contribute to the venom-status decision while retaining a dedicated direct prediction pathway.

The overall training objective combines the species-classification loss, venom-status loss, and consistency regularization. For $K = 7$ species and label-smoothing coefficient $\epsilon = 0.1$, the species loss is a label-smoothed cross-entropy over the prototypical class probabilities:

$$\mathcal{L}_{\text{species}} = - \sum_{k=1}^K \left[(1 - \epsilon)y_k + \frac{\epsilon}{K} \right] \log p(y = k|\mathbf{z}) \quad (15)$$

The binary venom loss is

$$\mathcal{L}_{\text{venom}} = -[v \log p_{\text{venom}} + (1 - v) \log(1 - p_{\text{venom}})] \quad (16)$$

where $v \in \{0,1\}$ denotes the venom-status label. The consistency term is defined as $\mathcal{L}_{\text{cons}} = \|p_{\text{dir}} - p_{\text{marg}}\|^2$. The complete objective is therefore

$$\mathcal{L} = \mathcal{L}_{\text{species}} + \lambda_v \mathcal{L}_{\text{venom}} + \lambda_c \mathcal{L}_{\text{cons}}, \quad (17)$$

where λ_v and λ_c control the contribution of the venom and consistency terms, respectively. This formulation allows the two tasks to be optimized jointly while explicitly encouraging agreement between the independent venom estimate and the venom probability implied by species classification.

3.7. Training, Ensembling, and Test-Time Augmentation

The cross-view fusion module and prediction heads are trained using AdamW [29] with episodic sampling. Each training episode follows a 7-way, 3-shot configuration, with the query set defined according to the experimental protocol. Model selection is based on validation species accuracy with early stopping. Because the small validation set can produce noticeable run-to-run variability, five models are trained using different random seeds, $s = 1, \dots, 5$. Their class probabilities are averaged at inference:

$$p_{\text{ens}}(y|\mathbf{z}) = \frac{1}{5} \sum_{s=1}^5 p_s(y|\mathbf{z}) \quad (18)$$

This probability-level ensemble is used to improve predictive stability and reduce sensitivity to the particular random initialization and training trajectory. Performance variability across the five seeds is quantified separately in Section 4.5. At test time, the framework additionally applies horizontal-flip test-time augmentation (TTA). Predictions from the original and horizontally flipped versions of the input image are averaged:

$$p_{\text{TTA}}(y) = \frac{1}{2} [p(y|I) + p(y|\text{flip}(I))] \quad (19)$$

The same inference procedure is applied to the species and venom predictions. The resulting ensemble and TTA predictions are used for the final held-out test evaluation reported in Section 4. This training and inference strategy separates model development from final evaluation: the validation set is used for early stopping and model selection, whereas the held-out test set is used only for final performance assessment. The five-seed ensemble and TTA are therefore treated as inference procedures rather than additional training stages.

4. Results and Discussion

4.1. Evaluation Protocol and Experimental Setup

The framework is implemented in PyTorch using the timm library [30], with the publicly released DINOv2 ViT-B/14 weights. All experiments are conducted on a single NVIDIA T4 GPU. The task-specific fusion and prediction modules are optimized using AdamW with a learning rate of 10^{-3} and weight decay of 0.05; the remaining model configuration is summarized in Section 3 and Table 3.

Two complementary evaluation protocols are used. The first employs the fixed train/validation/test split for the main held-out evaluation. Results are reported both for individual model runs across five random seeds and for the corresponding five-seed probability ensemble. The second protocol uses five-fold stratified cross-validation over the pooled dataset to assess the sensitivity of performance to the data partition. For the fixed-split protocol, validation data are used for model selection and early stopping, whereas the cross-validation models follow a fixed training schedule without a separate validation set. In both protocols, the samples used for final evaluation are not used for model or hyperparameter selection.

4.2. Evaluation Metrics

Model performance is assessed using complementary metrics chosen according to the characteristics of the two tasks. For the seven-class species classification problem, macro-F1 and macro-recall are treated as the primary metrics, because they give equal weight to all species and therefore better reflect performance across classes with different numbers of images. Overall accuracy and balanced accuracy are reported as supporting measures. For the binary venom-status task, recall on the venomous class is the primary safety-oriented metric, since a false-negative prediction may have more serious consequences than a false positive. Venom accuracy, macro-F1, and ROC-AUC are reported as complementary measures. For the species task, accuracy is defined as

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = y_i] \quad (19)$$

where N is the number of evaluated samples, y_i is the ground-truth label, and $\hat{y}_i$ is the predicted label. Balanced accuracy is the arithmetic mean of the per-class recalls:

$$\text{BAcc} = \frac{1}{K} \sum_{k=1}^K \text{Recall}_k \quad (20)$$

where $K = 7$ is the number of species and

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \quad (21)$$

Precision and F1 are computed for each class as

$$\text{Prec}_k = \frac{TP_k}{TP_k + FP_k} \quad (22)$$

and

$$\text{F1}_k = \frac{2\text{Prec}_k\text{Recall}_k}{\text{Prec}_k + \text{Recall}_k} \quad (23)$$

Macro-precision, macro-recall, and macro-F1 are obtained by averaging the corresponding per-class values with equal weight. The same metrics are used for the binary venom-status task, together with ROC-AUC. In addition, recall for the venomous class is reported explicitly to quantify the rate at which venomous snakes are correctly identified. Prediction reliability is assessed using the expected calibration error (ECE) [31]. Predictions are divided into B equal-width confidence bins, and ECE is computed as

$$\text{ECE} = \sum_{b=1}^B \frac{|B_b|}{N} |\text{acc}(B_b) - \text{conf}(B_b)|, \quad (24)$$

where B_b denotes the samples assigned to bin b , $\text{acc}(B_b)$, and $\text{conf}(B_b)$ is the mean predicted confidence in that bin. Lower ECE indicates closer agreement between predicted confidence and empirical accuracy.

Finally, uncertainty in the reported performance is characterized in two ways. Means and standard deviations are reported across random seeds for the fixed-split experiments and across folds for five-fold cross-validation. For the final held-out species accuracy, a 95% bootstrap confidence interval is additionally estimated using 2,000 resamples of the test set. These

measures provide complementary information about performance variability across model initializations and data partitions.

4.3. Species and Venom Classification Results

Training behaviour is examined first to provide context for the subsequent held-out evaluation. Figure 3 presents the validation species and venom accuracy together with the training loss for a representative seed. Both accuracy curves increase rapidly during the early training stages and subsequently plateau, while the training loss decreases without an evident divergence from validation performance. This behaviour suggests stable optimization of the task-specific modules under the frozen-encoder setting, although a single training trajectory should not be interpreted as evidence of generalization by itself.

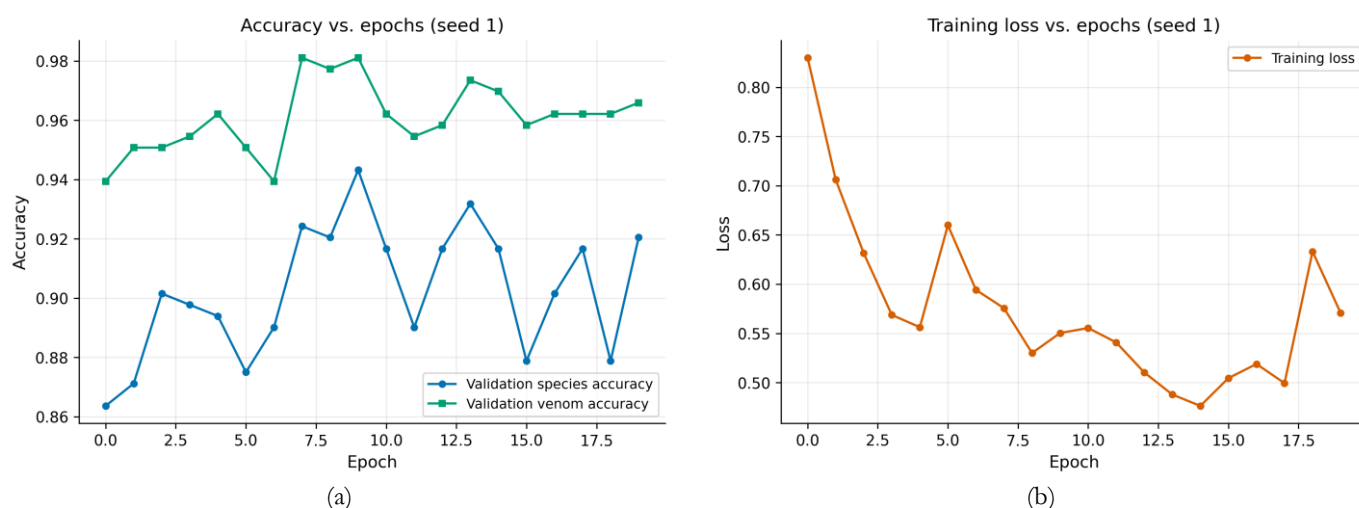


Figure 3. Training dynamics for a representative seed: (a) Validation species and venom accuracy vs. epoch; (b) Training loss vs. epoch.

The final held-out results are summarized in Table 4. For species classification, the five-seed ensemble achieves a macro-F1 of 0.9748 and macro-recall of 0.9757, with an overall accuracy of 0.9742. The close agreement between macro-F1, macro-recall, and accuracy indicates that performance is relatively balanced across the seven species rather than being driven primarily by the three larger non-venomous classes. For venom-status classification, the ensemble achieves 0.9852 accuracy, 0.9852 macro-F1, and 0.9926 recall for the venomous class, indicating that very few venomous samples are missed.

Table 4. Overall performance on the held-out test set.

Metric	Value
Species accuracy (5-seed ensemble)	0.9742
Species accuracy (single model, mean $\pm$ std)	0.9454 $\pm$ 0.0092
Species macro-precision	0.9743
Species macro-recall	0.9757
Species macro-F1	0.9748
Species accuracy, 95% bootstrap CI	[0.9557, 0.9926]
Species ECE	0.0786
Venom accuracy (5-seed ensemble)	0.9852
Venom accuracy (single model, mean $\pm$ std)	0.9734 $\pm$ 0.0072
Venom macro-precision	0.9854
Venom macro-F1	0.9852
Venom ROC-AUC	0.9985
Recall on venomous class	0.9926

The ensemble species accuracy has a 95% bootstrap confidence interval of 0.9557–0.9926, while the single-model results across five seeds reach 0.9454 ± 0.0092 for species accuracy and 0.9734 ± 0.0072 for venom accuracy. The difference between the single-model and ensemble results indicates that averaging predictions across seeds provides a measurable gain on this relatively small dataset. Importantly, the ensemble result should be interpreted as performance under the specified held-out split rather than as an estimate of performance across unseen geographic or species distributions; the latter is examined more conservatively through five-fold cross-validation in Section 4.5.

The per-species results in Table 5 provide a more detailed view of the classification behaviour. All seven species achieve an F1 score above 0.94, with common krait classified without error on the test set ($F1 = 1.000$). King cobra has the lowest F1 score (0.943), primarily because of occasional confusion with Indian cobra and oriental rat snake. The confusion matrix in Figure 4 shows that the seven errors are concentrated in a small number of visually similar species rather than being distributed broadly across the class set. In particular, the remaining errors involve the cobra species and the non-venomous colubrid species. This pattern suggests that the residual difficulty is primarily fine-grained species discrimination rather than a systematic tendency to favour either venom-status category.

Table 5. Per-class precision, recall, F1, and support on the held-out test set.

Species	Venom status	Precision	Recall	F1	Support
Chequered keelback	Non-venomous	0.978	0.978	0.978	45
Common krait	Venomous	1.000	1.000	1.000	34
Indian cobra	Venomous	1.000	0.971	0.985	34
Indian rock python	Non-venomous	0.978	0.978	0.978	45
King cobra	Venomous	0.917	0.971	0.943	34
Oriental rat snake	Non-venomous	0.977	0.933	0.955	45
Russell's viper	Venomous	0.971	1.000	0.986	34
Macro average		0.974	0.976	0.975	271

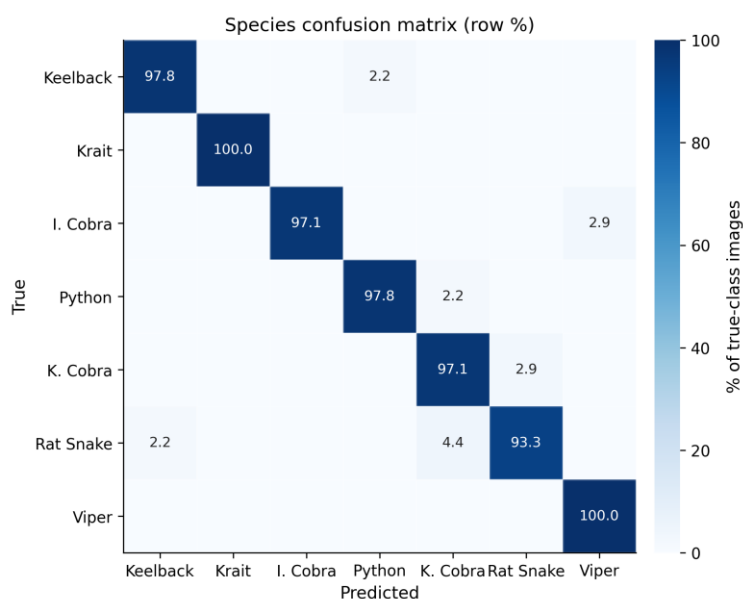


Figure 4. Species confusion matrix on the held-out test set, shown as row-normalized percentages.

The calibration and discrimination diagnostics are shown in Figure 5. The species predictions achieve an ECE of 0.0786, indicating a residual gap between predicted confidence and empirical accuracy across confidence bins. Thus, the model is reasonably calibrated but still exhibits some overconfidence. This result supports the use of confidence estimates as an additional diagnostic, while further calibration would be advisable before any safety-critical

deployment. For venom classification, the ROC-AUC of 0.9985 indicates excellent discrimination across decision thresholds.

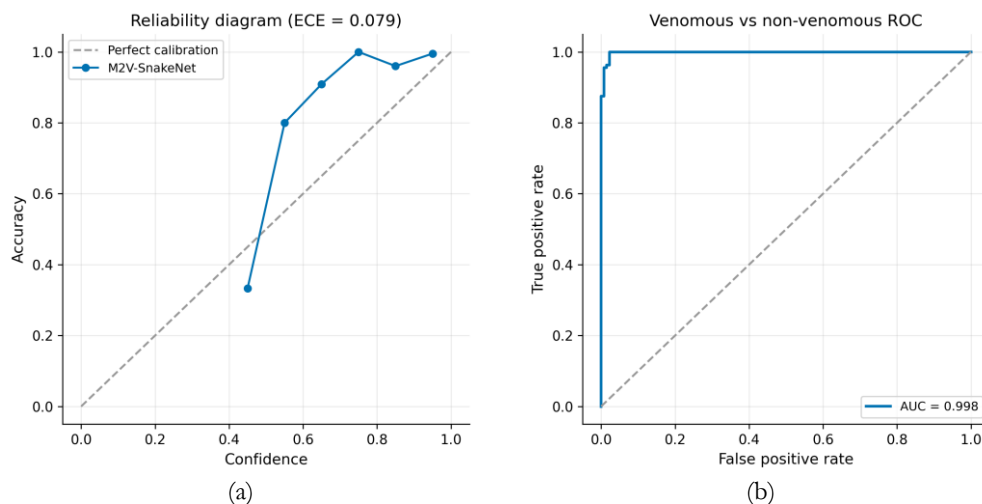


Figure 5. Calibration and discrimination diagnostics for the proposed framework: (a) Reliability diagram for species classification; (b) ROC curve for venom-status classification

Figure 6 provides a two-dimensional t-SNE projection of the fused test embeddings [32]. The visualization shows relatively distinct clusters for the seven species and substantial separation between venomous and non-venomous samples. This qualitative structure is consistent with the quantitative classification results and suggests that the fused representation captures discriminative information relevant to both tasks. However, the t-SNE projection is descriptive rather than a direct test of representation robustness or the absence of background-related cues.

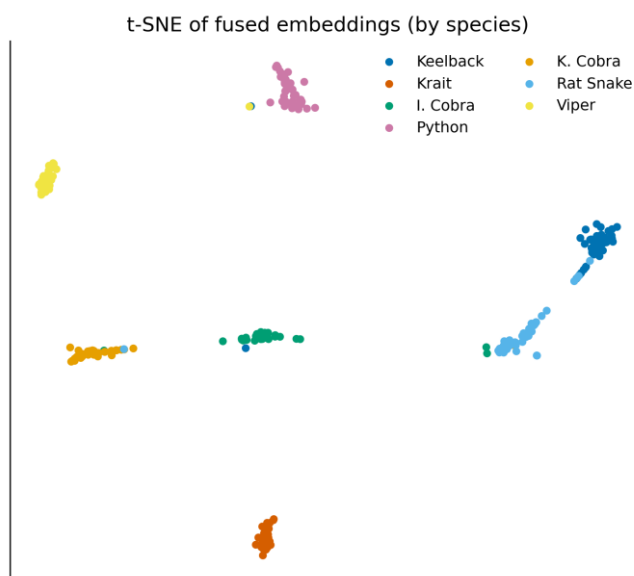


Figure 6. t-SNE projection of the fused test embeddings, coloured by species.

Overall, the main evaluation indicates that the proposed framework performs consistently well across the seven species while maintaining particularly high sensitivity to venomous snakes. The remaining errors are concentrated in visually similar species, whereas the venom-status decision is less affected by these within-status species confusions. The contribution of individual architectural components is examined next through controlled ablation experiments.

4.4. Ablation Study

Ablation experiments are conducted to separate the contributions of the frozen backbone, multi-view representation, fusion strategy, and classification head. Two protocols are used. The backbone comparison keeps the downstream architecture and training procedure fixed and evaluates five random seeds, whereas the component ablations remove or replace one component at a time within the DINOv2-based framework and use three seeds. All variants use the same fixed data split and evaluation procedure within their respective protocols.

4.4.1. Backbone Ablation

Tables 6 and 7 compare the frozen ImageNet-supervised ViT-B [33], [34] with frozen DINOv2 under the same five-seed protocol. The backbone choice produces the largest performance difference among the evaluated ablations. For species classification, replacing the ImageNet-supervised ViT-B with DINOv2 increases single-model accuracy from 0.7712 ± 0.0816 to 0.9454 ± 0.0092 , corresponding to a 17.4-percentage-point improvement. Venom accuracy increases from 0.9262 ± 0.0146 to 0.9734 ± 0.0072 .

The DINOv2 representation is also substantially more stable across seeds: the standard deviation of species accuracy decreases from 0.0816 to 0.0092. This result suggests that the stronger frozen representation reduces the sensitivity of the lightweight task-specific modules to random initialization. A plausible explanation is that self-supervised DINOv2 features provide a more transferable representation for fine-grained visual similarity than features learned primarily through the ImageNet classification objective. This interpretation is consistent with the prototypical head, which relies on distances between embeddings rather than a conventional class-specific softmax layer. With five-seed ensembling, the DINOv2 configuration reaches 0.9742 species accuracy and 0.9852 venom accuracy, matching the corresponding results in Table 4. This agreement confirms that the backbone ablation and the main evaluation use the same fixed-split ensemble protocol.

Table 6. Backbone comparison for species classification (five seeds, mean $\pm$ standard deviation).

Backbone	Accuracy	Precision	Recall	F1
ImageNet ViT-B	0.7712 ± 0.0816	0.7761 ± 0.0888	0.7773 ± 0.0855	0.7741 ± 0.0891
DINOv2 (this work)	0.9454 ± 0.0092	0.9487 ± 0.0098	0.9483 ± 0.0083	0.9474 ± 0.0095

Table 7. Backbone comparison for venom classification (five seeds, mean $\pm$ standard deviation).

Backbone	Accuracy	Precision	Recall	F1
ImageNet ViT-B	0.9262 ± 0.0146	0.9274 ± 0.0135	0.9261 ± 0.0146	0.9261 ± 0.0147
DINOv2 (this work)	0.9734 ± 0.0072	0.9739 ± 0.0070	0.9734 ± 0.0072	0.9734 ± 0.0072

4.4.2. Component Ablation

Tables 8 and 9 isolate the effects of the multi-view representation, fusion mechanism, classification head, and TTA. The results reveal that the benefits of these components are not uniform across tasks. The strongest architectural effect is associated with learned fusion. Replacing the fusion module with parameter-free mean pooling reduces species accuracy from 0.9434 to 0.8339, a decrease of 10.95 percentage points. This result indicates that simply averaging the view embeddings is insufficient for this multi-view setting, where individual region views can differ substantially in informativeness. Removing the multi-view decomposition also reduces species accuracy, from 0.9434 to 0.9065, corresponding to a 3.69-point decrease. The result supports the value of incorporating complementary region information beyond the global foreground view.

The comparison between attention and concatenation is more nuanced. Their species accuracies are nearly identical (0.9434 ± 0.0063 for attention and 0.9422 ± 0.0087 for concatenation), with the difference of only 0.12 percentage points being small relative to the observed seed variability. Thus, the results do not support treating attention as a dominant contributor to species accuracy. The distinction is more apparent for venom classification: concatenation achieves 0.9582 ± 0.0122 venom accuracy compared with 0.9705 ± 0.0030 for attention, together with substantially greater seed variability. One possible explanation is that attention can adaptively modulate the contribution of individual views, whereas

concatenation preserves each view in a fixed positional representation. This interpretation remains a hypothesis rather than a directly verified mechanism.

Replacing the prototypical head with a linear classifier produces a more moderate reduction in species accuracy, from 0.9434 to 0.9274, or approximately 1.6 percentage points. This result indicates that metric-based classification contributes to the overall performance, but it is not the principal source of the observed gains. Interestingly, venom classification is relatively stable across the component variants: accuracy remains between 0.9582 and 0.9717, while ROC-AUC remains above 0.987. This robustness is consistent with the observation that several species-level errors occur between species sharing the same venom status, so an incorrect species prediction does not necessarily change the venom decision.

Table 8. Component ablation for species classification (three seeds, mean $\pm$ standard deviation).

Ablation	Accuracy	Precision	Recall	F1
Full model	0.9434 $\pm$ 0.0063	0.9471 $\pm$ 0.0038	0.9448 $\pm$ 0.0066	0.9451 $\pm$ 0.0054
w/o multi-view	0.9065 $\pm$ 0.0196	0.9168 $\pm$ 0.0149	0.9086 $\pm$ 0.0194	0.9089 $\pm$ 0.0196
Mean-pool fusion	0.8339 $\pm$ 0.0000	0.8363 $\pm$ 0.0000	0.8274 $\pm$ 0.0000	0.8166 $\pm$ 0.0000
Concat fusion	0.9422 $\pm$ 0.0087	0.9452 $\pm$ 0.0098	0.9455 $\pm$ 0.0077	0.9451 $\pm$ 0.0090
Linear head	0.9274 $\pm$ 0.0122	0.9322 $\pm$ 0.0124	0.9307 $\pm$ 0.0089	0.9295 $\pm$ 0.0112
w/o TTA	0.9434 $\pm$ 0.0063	0.9472 $\pm$ 0.0034	0.9452 $\pm$ 0.0062	0.9456 $\pm$ 0.0050

Table 9. Component ablation for venom classification (three seeds, mean $\pm$ standard deviation).

Ablation	Accuracy	Precision	Recall	F1
Full model	0.9705 $\pm$ 0.0030	0.9707 $\pm$ 0.0028	0.9705 $\pm$ 0.0030	0.9705 $\pm$ 0.0030
w/o multi-view	0.9619 $\pm$ 0.0097	0.9619 $\pm$ 0.0096	0.9619 $\pm$ 0.0097	0.9619 $\pm$ 0.0097
Mean-pool fusion	0.9692 $\pm$ 0.0035	0.9697 $\pm$ 0.0036	0.9693 $\pm$ 0.0035	0.9692 $\pm$ 0.0035
Concat fusion	0.9582 $\pm$ 0.0122	0.9589 $\pm$ 0.0121	0.9582 $\pm$ 0.0122	0.9582 $\pm$ 0.0122
Linear head	0.9705 $\pm$ 0.0030	0.9710 $\pm$ 0.0030	0.9705 $\pm$ 0.0030	0.9705 $\pm$ 0.0030
w/o TTA	0.9717 $\pm$ 0.0035	0.9719 $\pm$ 0.0032	0.9717 $\pm$ 0.0034	0.9717 $\pm$ 0.0035

Taken together, the ablations suggest a clear hierarchy of contributions. The pretrained representation accounts for the largest performance difference, learned multi-view fusion provides a substantial additional benefit, the multi-view decomposition contributes a smaller but consistent gain, and the prototypical head provides a moderate improvement. The near-equivalence of attention and concatenation for species classification also indicates that the principal benefit of the fusion stage lies in learning how to combine views rather than in attention itself being universally superior.

4.5. Cross-Validation Analysis

To assess the sensitivity of the results to the choice of data partition, five-fold stratified cross-validation was performed over the pooled dataset. One model was trained for each fold, with the folds preserving the species distribution. Unlike the fixed-split evaluation, the cross-validation protocol uses a fixed training schedule without a separate validation set. The resulting performance is summarized in Table 10.

Across the five folds, species accuracy is 93.4% $\pm$ 1.5%, with macro-precision of 94.0% $\pm$ 1.7%, macro-recall of 93.5% $\pm$ 1.3%, and macro-F1 of 93.6% $\pm$ 1.5%. For venom classification, the corresponding accuracy and F1 are both 96.9% $\pm$ 1.4%, while ROC-AUC reaches 0.995 $\pm$ 0.005. The relatively small standard deviations indicate that performance remains reasonably stable across the five partitions, although the lower-performing third fold shows that some partitions are more challenging than others.

The cross-validation accuracy is lower than the 97.4% obtained by the five-seed ensemble on the fixed held-out split. This difference should not be interpreted as evidence that cross-validation models are disadvantaged by having less training data. Each cross-validation model uses approximately 1,421 images (four fifths of the 1,776-image dataset), compared with 1,241 images in the fixed-split training set. Instead, the two estimates represent different evaluation procedures. The fixed-split result uses a five-seed ensemble and validation-based

early stopping, whereas each cross-validation fold uses a single model under a fixed training schedule without a separate validation set. In addition, the fixed-split result is measured on one particular test partition of 271 images, while cross-validation averages performance over five rotating test folds covering the entire dataset.

The distinction between these estimates is therefore important when interpreting the reported performance. The 93.4% cross-validation accuracy provides the more conservative estimate of performance across alternative data partitions, whereas the 97.4% ensemble result represents the performance achieved under the specified held-out split and ensemble procedure. The latter should not be interpreted as an expected performance level for unseen geographic regions or species.

Table 10. Five-fold stratified cross-validation results.

Fold	Species Acc.	Species Prec.	Species Rec.	Species F1	Venom Acc.	Venom F1	Venom AUC
1	0.9333	0.9440	0.9317	0.9352	0.9750	0.9750	0.9988
2	0.9412	0.9469	0.9439	0.9452	0.9860	0.9860	0.9985
3	0.9065	0.9059	0.9102	0.9065	0.9462	0.9461	0.9853
4	0.9405	0.9464	0.9392	0.9414	0.9603	0.9603	0.9979
5	0.9490	0.9547	0.9482	0.9504	0.9773	0.9773	0.9948
Mean $\pm$ SD	0.934 $\pm$ 0.015	0.940 $\pm$ 0.017	0.935 $\pm$ 0.013	0.936 $\pm$ 0.015	0.969 $\pm$ 0.014	0.969 $\pm$ 0.014	0.995 $\pm$ 0.005

The cross-validation findings therefore complement rather than replace the fixed-split results. Together, they indicate that the model retains strong performance across different partitions, while also showing that the higher ensemble accuracy on the held-out split should be interpreted in the context of its specific evaluation protocol.

4.6. Frozen Encoder Design and External Meta-Pretraining

The backbone ablation in Section 4.4 indicates that the choice of visual representation has the largest effect among the evaluated design factors. The frozen DINOv2 encoder therefore appears particularly effective for the low-data setting considered in this study, allowing the task-specific modules to adapt a strong pretrained representation without updating the large backbone.

Given the meta-learning formulation of M2V-SnakeNet, an additional experiment examined whether meta-pretraining on an external snake corpus could provide further benefits. The external data were drawn from SnakeCLEF [22]. Two variants were evaluated: the first meta-trained the complete model, whereas the second meta-trained only the backbone before adapting it to the target task. Neither variant improved upon the frozen DINOv2 configuration. The full-model meta-pretraining variant reached a validation species accuracy of 0.72, while the backbone-only variant reached 0.80, compared with 0.89 for the frozen-encoder configuration at the corresponding validation stage as shown in Figure 7.

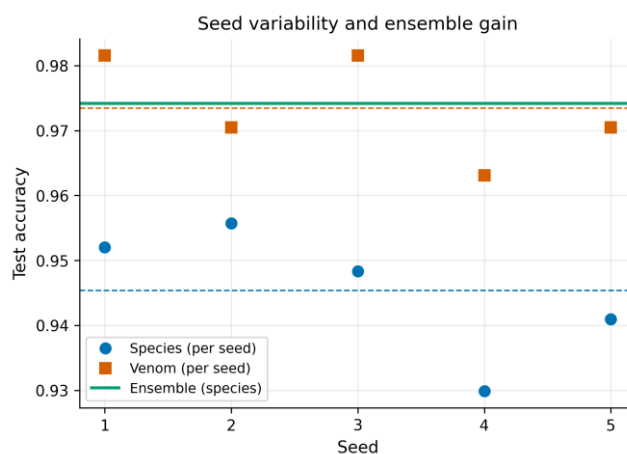


Figure 7. Validation species accuracy of the frozen DINOv2 baseline and the two external SnakeCLEF meta-pretraining variants.

These results suggest that additional task-specific meta-pretraining is not necessarily beneficial when a strong self-supervised foundation representation is already available. One possible explanation is the mismatch between the species distribution in the external SnakeCLEF corpus and the seven target species. Updating the pretrained representation may also reduce some of the transferable features provided by DINOv2. However, these mechanisms were not isolated experimentally and should therefore be regarded as plausible explanations rather than directly established causes.

The negative result is nevertheless relevant to the overall design rationale. Under the dataset scale and target-species setting examined here, retaining DINOv2 in frozen form was more effective than introducing additional external meta-pretraining. Thus, the role of meta-learning in M2V-SnakeNet is better characterized as task-specific metric adaptation over a strong frozen representation, rather than reliance on extensive external meta-pretraining. This finding also reinforces the importance of evaluating additional training stages empirically rather than assuming that more task-specific adaptation will necessarily improve a pretrained foundation model.

5. Comparison with State of the Art

Table 11 provides an indicative comparison between M2V-SnakeNet and representative prior studies. The comparison should be interpreted with caution because the studies differ substantially in dataset size and composition, number of species, task definition, data partitioning, and evaluation protocol. In particular, some studies address binary venom classification, others perform fine-grained species recognition, and several target global long-tailed benchmarks containing hundreds or thousands of species. Consequently, the reported values are not directly comparable and should not be interpreted as a strict performance ranking.

Table 11. Comparison of M2V-SnakeNet with representative prior image-based snake classification studies.

Ref.	Dataset / task	Method	Species Acc. (%)	Venom Acc. (%)	Other reported metric
[5]	6 species (Iran)	MobileNetV2; LDA+SVM baseline	93.16	–	84 (LDA+SVM)
[6]	1,766 images; binary venom classification	CNN + transfer learning + ensemble	–	91.3 (5-fold CV)	–
[8]	45 species	Salient-object detection + VGG16	97.09	–	–
[9]	Moroccan species; SnakeCLEF pretraining	EfficientNetB0, fully fine-tuned	92.23	–	F1 93.67
[10]	SnakeCLEF 2022; 1,572 species, global	Challenge overview; best of 31 teams	94.01	–	Macro-F1 86.47
[11]	SnakeCLEF 2021; 772 species	EfficientNet-B4 + ViT ensemble + location prior	90.42	–	Macro-F1 78.75; country-F1 82.88
[12]	SnakeCLEF 2021; 772 species	ResNeSt/ResNeXt/ResNet ensemble + country filter	91.6	–	Macro-F1 83.0
[18]	Small-scale curated snake set	Swin Transformer + improved SimCLR	97.48	–	Macro-F1 97.0
[19]	SnakeCLEF 2024; 182,261 images, 1,784 species	Frozen DINOv2 + linear probe	–	–	Track-1 score 39.69; reported F1/accuracy affected by label-indexing error
[21]	772 species, worldwide	Vision Transformer	96.0	–	Macro-F1 92.2
[22]	SnakeCLEF 2023; 1,784 species	Challenge overview; best of five evaluated teams	–	–	Track-1 metric 91.43; best macro-F1 61.39
Ours	1,776 images; 7 species	Frozen DINOv2 + multi-view + prototypical head	97.4	98.5	Macro-F1 97.5; CV species acc. 93.4 $\pm$ 1.5%; CV venom acc. 96.9 $\pm$ 1.4%

For completeness, the table reports the principal classification results available from each selected study. Where accuracy is not reported or is not meaningful for the corresponding task, the metric originally reported by the study is retained in the final column. Challenge-specific composite scores and other task-specific metrics are therefore not interpreted as equivalent to classification accuracy. The SnakeCLEF 2024 result [19] also requires particular

caution because the reported zero accuracy and F1 resulted from a label-indexing error acknowledged by the authors; its reported challenge score is therefore the more informative value.

The comparison highlights two complementary aspects of the proposed framework. First, its 97.4% species accuracy is within the range of the strongest results reported on relatively small curated datasets, including 97.09% [8] and 97.48% [18]. However, these datasets and protocols are not identical, so this should not be interpreted as evidence of superiority. Second, M2V-SnakeNet additionally evaluates venom status on the same dataset, achieving 98.5% accuracy and 99.3% recall for the venomous class, addressing a task that is not evaluated jointly with species identification in the listed species-classification studies.

More importantly, the results should be interpreted together with the cross-validation analysis rather than through the highest held-out accuracy alone. The five-fold cross-validation result of $93.4\% \pm 1.5\%$ species accuracy is substantially more conservative than the five-seed ensemble result of 97.4%, reflecting the sensitivity of performance to the evaluation protocol and data partition. Thus, the principal distinction of M2V-SnakeNet is not simply its absolute accuracy, but the combination of frozen foundation features, multi-view representation, metric-based classification, joint species–venom prediction, and evaluation across multiple seeds and data partitions.

This comparison also reinforces the scope of the contribution. The proposed model is designed for a small, fixed regional species set, rather than the global long-tailed recognition problem addressed by SnakeCLEF [10]–[12], [19], [22]. Its results therefore provide evidence for the effectiveness of the proposed low-data formulation within the studied setting, but do not establish superiority or generalization across the much broader species and geographic distributions represented by global benchmarks.

6. Conclusions

This study presented M2V-SnakeNet, a low-data framework that combines a frozen self-supervised DINOv2 encoder with presence-aware multi-view generation, cross-view attention fusion, a prototypical metric-based head, and a hierarchical species–venom prediction head. The experiments show that the quality of the pretrained representation is the dominant factor in this setting: replacing the frozen ImageNet-supervised ViT-B with DINOv2 improves single-model species accuracy by 17.4 percentage points while substantially reducing seed-to-seed variability. Multi-view representation and learned fusion provide additional gains, with removal of multi-view reducing species accuracy by 3.69 points and mean pooling causing a 10.95-point reduction. The prototypical head provides a more moderate improvement, while attention fusion is particularly beneficial for the stability of venom prediction.

Under five-fold cross-validation, M2V-SnakeNet achieves $93.4\% \pm 1.5\%$ species accuracy, providing a more conservative estimate than the 97.4% obtained with the five-seed ensemble on the fixed held-out split. For venom classification, the framework achieves 99.3% recall on the venomous class, demonstrating strong sensitivity to the safety-critical class within the evaluated dataset. The additional meta-pretraining experiment produced a negative result, indicating that external task-specific meta-pretraining did not improve the frozen DINOv2 configuration. The results also show that higher ensemble accuracy does not necessarily imply better calibration, highlighting the need for post-hoc calibration before any operational deployment.

The main limitation is the scope of the evaluation: the study covers seven species and 1,776 curated images from a single geographic region, without external validation. Therefore, the results should not be interpreted as evidence of broad geographic, species-level, or in-the-wild generalization. Future work should evaluate larger and geographically diverse datasets, improve probability calibration and part discovery, explore higher-resolution foundation models, and assess the framework under realistic field conditions.

Author Contributions: Conceptualization: H.M.K. and W.A.; Methodology: H.M.K. and W.A.; Software: H.M.K.; Validation: H.M.K. and S.M.A.; Formal analysis: H.M.K.; Investigation: H.M.K.; Resources: W.A. and S.M.A.; Data curation: H.M.K.; Writing—original draft preparation: H.M.K.; Writing—review and editing: W.A. and S.M.A.; Visualization: H.M.K.;

Supervision: W.A. and S.M.A.; Project administration: W.A.; Funding acquisition: Not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement: The curated snake image dataset, including the per-species composition, the venom-status labels, and the exact training, validation and test splits reported in Section 3.2, is publicly available at <https://www.kaggle.com/datasets/hamzamaqsoodkhan/snake-dataset>. The training and evaluation code (including the embedding-cache, seed-ensemble, ablation, and cross-validation procedures) is available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used AI-assisted tools for formatting the manuscript into the journal template, for verifying bibliographic records against their published versions of record, and for language and consistency checking. No AI tool was used in the study design, the implementation of the framework, the execution of experiments, the analysis of results, or the interpretation of findings. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this manuscript.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

- [1] J. M. Gutiérrez, J. J. Calvete, A. G. Habib, R. A. Harrison, D. J. Williams, and D. A. Warrell, "Snakebite envenoming," *Nat. Rev. Dis. Prim.*, vol. 3, no. 1, p. 17063, Sep. 2017, doi: 10.1038/nrdp.2017.63.
- [2] J. Aquino, I. L. E. Lima, F. R. Costa, and I. Nunes, "Trends in diagnostic characters of Snakes species (Squamata) in the first two decades of the 21st century," *Zootaxa*, vol. 5689, no. 3, pp. 439–450, Sep. 2025, doi: 10.11646/zootaxa.5689.3.2.
- [3] A. Patel, L. Cheung, N. Khatod, I. Matijosaitiene, A. Arteaga, and J. W. Gilkey, "Revealing the Unknown: Real-Time Recognition of Galápagos Snake Species Using Deep Learning," *Animals*, vol. 10, no. 5, p. 806, May 2020, doi: 10.3390/ani10050806.
- [4] M. Vasmatkar, I. Zare, P. Kumbha, S. Pimpalkar, and A. Sharma, "Snake Species Identification and Recognition," in *2020 IEEE Bombay Section Signature Conference (IBSSC)*, Dec. 2020, pp. 1–5. doi: 10.1109/IBSSC51096.2020.9332218.
- [5] M. Rajabizadeh and M. Rezghi, "A comparative study on image-based snake identification using machine learning," *Sci. Rep.*, vol. 11, no. 1, p. 19142, Sep. 2021, doi: 10.1038/s41598-021-96031-1.
- [6] N. I. Progg, N. Rezoana, M. S. Hossain, R. U. Islam, and K. Andersson, "A CNN Based Model for Venomous and Non-venomous Snake Classification," in *Communications in Computer and Information Science*, 2021, pp. 216–231. doi: 10.1007/978-3-030-82269-9_17.
- [7] P. Joshi, S. Sati, T. Choudhury, T. Bajaj, K. Kotecha, and A. Sar, "Classification of Venomous and Non-venomous Snakes Using Transfer Learning with MobileNetV2," in *Lecture Notes in Networks and Systems*, 2024, pp. 427–438. doi: 10.1007/978-981-97-3810-6_35.
- [8] K. Ahmed, M. A. Gad, and A. E. Aboutabl, "Snake species classification using deep learning techniques," *Multimed. Tools Appl.*, vol. 83, no. 12, pp. 35117–35158, Sep. 2023, doi: 10.1007/s11042-023-16773-0.
- [9] M. Iguernane, M. Ouzziki, Y. Es-Saady, M. El Hajji, A. Lansari, and A. Bouazza, "Deep Learning-Based Snake Species Identification for Enhanced Snakebite Management," *AI*, vol. 6, no. 2, p. 21, Jan. 2025, doi: 10.3390/ai6020021.
- [10] L. Picek, M. Hruz, A. M. Durso, and I. Bolon, "Overview of SnakeCLEF 2022: Automated Snake Species Identification on a Global Scale," in *CLEF 2022: Conference and Labs of the Evaluation Forum*, 2022. [Online]. Available: <https://ceur-ws.org/Vol-3180/paper-156.pdf>
- [11] L. Bloch and C. M. Friedrich, "EfficientNets and Vision Transformers for Snake Species Identification Using Image and Location Information," in *CLEF 2021– Conference and Labs of the Evaluation Forum*, 2021. [Online]. Available: <https://ceur-ws.org/Vol-2936/paper-126.pdf>
- [12] R. Chamidullin, M. Šulc, J. Matas, and L. Picek, "A Deep Learning Method for Visual Recognition of Snake Species," in *CLEF 2021– Conference and Labs of the Evaluation Forum*, 2021. [Online]. Available: <https://ceur-ws.org/Vol-2936/paper-128.pdf>
- [13] M. Palaniappan, K. Desingu, H. Bharathi, E. A. Chodisetty, and A. Bhaskar, "Deep Learning and Gradient Boosting Ensembles for Classification of Snake Species," in *CLEF 2022: Conference and Labs of the Evaluation Forum*, 2022. [Online]. Available: <https://ceur-ws.org/Vol-3180/paper-173.pdf>
- [14] F. Hu *et al.*, "Watch out Venomous Snake Species: A Solution to SnakeCLEF2023," in *Conference and Labs of the Evaluation Forum*, Jul. 2023. [Online]. Available: <http://arxiv.org/abs/2307.09748>
- [15] K. Desingu, M. Palaniappan, and J. Kumar, "Snake Species Classification using Transfer Learning Technique," in *CLEF 2021 – Conference and Labs of the Evaluation Forum*, 2021. [Online]. Available: <https://ceur-ws.org/Vol-2936/paper-133.pdf>
- [16] L. Kalinathan, P. Balasundaram, P. Ganesh, S. S. Bathala, and R. K. Mukesh, "Automatic Snake Classification using Deep Learning Algorithm," in *CLEF 2021 – Conference and Labs of the Evaluation Forum*, 2021. [Online]. Available: <https://ceur-ws.org/Vol-2936/paper-135.pdf>
- [17] H. Wang *et al.*, "Snake-DETR: a lightweight and efficient model for fine-grained snake detection in complex natural environments," *Sci. Rep.*, vol. 15, no. 1, p. 1282, Jan. 2025, doi: 10.1038/s41598-024-84328-w.
- [18] L. Li, R. Kang, W. Huang, and W. Feng, "A Study on Small-Scale Snake Image Classification Based on Improved SimCLR," *Appl. Sci.*, vol. 15, no. 11, p. 6290, Jun. 2025, doi: 10.3390/app15116290.

- [19] A. Miyaguchi, M. Gustineli, A. Fischer, and R. Lundqvist, "Transfer Learning with Self-Supervised Vision Transformers for Snake Identification," in *CLEF 2024: Conference and Labs of the Evaluation Forum*, 2024. [Online]. Available: <https://ceur-ws.org/Vol-3740/paper-203.pdf>
- [20] A. M. Durso, G. K. Moorthy, S. P. Mohanty, I. Bolon, M. Salathé, and R. Ruiz de Castañeda, "Supervised Learning Computer Vision Benchmark for Snake Species Identification From Photographs: Implications for Herpetology and Global Health," *Front. Artif. Intell.*, vol. 4, Apr. 2021, doi: 10.3389/frai.2021.582110.
- [21] I. Bolon, L. Pícek, A. M. Durso, G. Alcoba, F. Chappuis, and R. Ruiz de Castañeda, "An artificial intelligence model to identify snakes from across the world: Opportunities and challenges for global health and herpetology," *PLoS Negl. Trop. Dis.*, vol. 16, no. 8, p. e0010647, Aug. 2022, doi: 10.1371/journal.pntd.0010647.
- [22] L. Pícek, R. Chamidullin, M. Hruz, and A. M. Durso, "Overview of SnakeCLEF 2023: Snake Identification in Medically Important Scenarios," in *CLEF 2023: Conference and Labs of the Evaluation Forum*, 2023. [Online]. Available: <https://ceur-ws.org/Vol-3497/paper-168.pdf>
- [23] F. M. Firnando, D. R. I. M. Setiadi, A. R. Muslikh, and S. W. Iriananda, "Analyzing InceptionV3 and InceptionResNetV2 with Data Augmentation for Rice Leaf Disease Classification," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 1–11, May 2024, doi: 10.62411/faith.2024-4.
- [24] iNaturalist, "Snakes (suborder Serpentes)," *iNaturalist*. <https://www.inaturalist.org/taxa/85553-Serpentes> (accessed Jan. 02, 2026).
- [25] K. Muhamada *et al.*, "Exploring Machine Learning and Deep Learning Techniques for Occluded Face Recognition: A Comprehensive Survey and Comparative Analysis," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 160–173, Sep. 2024, doi: 10.62411/faith.2024-30.
- [26] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *Trans. Mach. Learn. Res.*, Feb. 2024, doi: 10.48550/arxiv.2304.07193.
- [27] A. Vaswani *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [28] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical Networks for Few-shot Learning," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Jun. 2017, pp. 4080–4090. [Online]. Available: <http://arxiv.org/abs/1703.05175>
- [29] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *ArXiv*. Jan. 04, 2019. [Online]. Available: <http://arxiv.org/abs/1711.05101>
- [30] R. Wightman *et al.*, "PyTorch Image Models (timm)," *Zenodo*. <https://github.com/huggingface/pytorch-image-models>
- [31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, Aug. 2017, pp. 1321–1330. [Online]. Available: <http://arxiv.org/abs/1706.04599>
- [32] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008, [Online]. Available: <https://www.jmlr.org/papers/v9/vandermaaten08a.html>
- [33] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR 2021*, Jun. 2021. [Online]. Available: <http://arxiv.org/abs/2010.11929>
- [34] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.