Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

ENHACLIP-SEARCH: CLIP-BASED VISION-LANGUAGE LEARNING FOR WILDLIFE CLASSIFICATION IN THE HIMALAYAN ECOSYSTEM

[1]Maghamis Ali Mehdi, [2]Dr. Farrukh Zeeshan Khan, [3]Usama Irshad, [4]Romail Khan, [5]Ummer Shakeel

[1]*Department of Computer Science, University of Engineering and Technology Taxila, Taxila 47050, Pakistan*

[3]*Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology University, Islamabad, Pakistan*

[1]23-ms-csc-6@students.uettaxila.edu.pk, [2]farrukh.zeeshan@uettaxila.edu.pk, [3]usama.irshad@szabist-isb.edu.pk , [4]23-ms-csc-3@students.uettaxila.edu.pk, [5]ummer.shakeel@students.uettaxila.edu.pk

DOI: https://doi.org/10.5281/zenodo.21921384

Article History
Received on: 15 Feb, 2026
Accepted on: 27 March, 2026
Published on 30 March, 2026

Copyright @Author
Corresponding Author:

Abstract

Biodiversity loss and the need for efficient ecological monitoring have increased the demand for reliable and automated wildlife species recognition systems. However, conventional image classification methods often struggle in complex natural environments due to variations in background, illumination, pose, scale, and species appearance. To address these challenges, this study proposes EnhaCLIP-Search, a vision-language framework for fine-grained wildlife classification in natural scene images. The proposed approach is built upon the CLIP ViT-B/32 architecture, which integrates a Vision Transformer for extracting visual representations with a text encoder for generating semantic representations of species descriptions. Image and text embeddings are projected into a shared feature space and compared using cosine similarity to identify the most semantically compatible wildlife category. To improve robustness under real-world environmental conditions, eight data augmentation strategies incorporating variations in shape, color, and illumination are employed during model development. The proposed framework focuses on fine-grained visual characteristics, including bird plumage and mammalian fur patterns, to distinguish between 20 wildlife species native to the Himalayan region. Experimental evaluation on a self-curated dataset demonstrates that EnhaCLIP-Search achieves 95% accuracy, 96% precision, 94% recall, and a 95% F1-score, outperforming the evaluated baseline approaches. These results demonstrate the effectiveness of the proposed vision-language framework for robust wildlife species recognition and highlight its potential for automated biodiversity monitoring in complex natural environments.

Introduction

One of the most significant activities in ecology and conservation is identification of various species of both birds and animals. It aids the scientists to learn the functioning of the ecosystems, track the changes in the population of wild animals and gauge the impact of climate change or human development on natural living. Specifically, proper species identification of birds has become important in monitoring trends in biodiversity and in the development of conservation policies [1]. With the development of technologies, species recognition is becoming more data-driven, which provides alternative methods of studying and protecting wildlife in the various parts of the world [4].

The Himalayas are one such region and one of the most biodiverse but least studied areas on earth. The Himalayan range, which spans many nations and climate zones, from tropical plains to snowy peaks over 6,000 meters (19,000 feet) above sea level, has an astonishing diversity of life forms. It has been observed that birds as well as mammals are most diverse at mid-elevations, which is due to climate and habitat productivity [3]. However, in spite of such diversity, researchers are yet to understand much about the seasonal or altitude variation in the distribution of species. Bird movement in the Himalayas is strongly seasonal, with over 65 percent of high-elevation breeding species moving up and down the mountains between seasons [1, 2]. Nevertheless, most of the studies on these movements have concentrated on birds, while mammals, and especially the ungulates, remain far less studied [3]. These ecological trends indicate that there should be improved instruments to record and categorize Himalayan wildlife in these dynamic and intricate settings.

Historically, the identification of species has been based on appearances and sound of trained ornithologists and field researchers combined with printed field guides. Although precise, this human-centered manual method would take a lot of time and effort and could not be expanded to the huge size of image collections of drones, camera traps, and citizen science projects [22]. This has been altered with the emergence of deep learning, which now drives classification tasks in fields as different as ecology and computing systems [20]. Automatic species identification in pictures can now be done by modern algorithms, which saves time and improves accuracy. Specifically, the approaches to Few-shot learning have gained popularity since they can identify new species using a few examples that have been labeled [5, 6]. The CLIP (ViT-B/32) model has been among the most promising ones, where deep residual and transformer-based architectures are used to learn visual patterns and surmount the problems that tend to restrain traditional deep networks [7].

Some scholars have used computer vision, deep learning, and even sound to categorize wildlife in various parts of the world. CNNs and transfer learning models have shown good performance on the common bird datasets and multimodal methods that combine both visual and acoustic cues have demonstrated better recognition [14]. These methods, however, mainly depend on large datasets that can be found worldwide like ImageNet or iNaturalist which almost never contain Himalayan species. This means that models trained with this type of data in most cases tend to be poor predictors of regional species that are visually similar or rare and whose images are constrained or low quality by environmental factors [8]. Such a difference demands specialized datasets and dynamic algorithms that are specifically tailored to the Himalayas.

This paper is devoted to the development of an AI-based classification of Himalayan birds and animals on the basis of the Few-shot learning framework with the CLIP (ViT-B/32) model. To justify this, we developed our dataset of Himalayan species, which could be helpful in further biodiversity informatics research. Our method will enhance the accuracy of recognition even at times of limited data and rare species. In addition to scholarly applications, the system can be useful in practical applications like smart biodiversity surveillance, eco-tourism, and educational applications, where species can be identified with direct camera traps or mobile phones.

In summary, our key contributions include:

(1) developing a unique dataset of Himalayan fauna, (2) designing a framework based on an enhanced transformer-based approach named

EnhaCLIP-Search, and (3) demonstrating its potential for advancing AI-driven conservation and wildlife monitoring.

The remaining sections of the paper are organized as follows: the "Related work" section describes the associated work. The "Methodology" section covers the proposed procedures and materials. The "Results and discussion" section presents the experiments' outcomes, whereas the "Conclusion" section concludes the paper.

1. Related Work

Nithin et al. [10] used different Convolutional Neural Network (CNN) models, such as ResNet, VGG16, VGG19, AlexNet, GoogleNet, and LeNet to classify a collection of 1,701 curated images of birds using different CNN models. The models were found to have an accuracy of 91-99% which is good in species identification. Amongst the architectures that were tested, variants of the ResNet and VGG models performed exceptionally well as more in-depth models. Nevertheless, the research had a limitation due to a limited sample of data and the lack of representation of a wide range of species, which may impact scalability and generalization to the real world. The authors emphasized the opportunities of CNN-based models in the applications of automated classification of birds and conservation monitoring. Inception-ResNet-v2 architecture with transfer learning was used to develop a deep learning-based classification system of climate change-affected Indian bird species created by Gavali et al. [9]. They used their approach with a novel data augmentation approach and fivefold cross-validation to guarantee model robustness. The system recorded a high predictive capacity of 94% and 96% accuracy and precision respectively when it comes to various classes of birds. Although it was effective, the study had limitations of the small number of databases and the non-inclusion of real-time monitoring. To increase the generalization and applicability of models, the authors proposed the scaling of datasets and optimization of augmentation techniques to assist in the conservation of biodiversity. Das et al. [11] designed an interpretable explainable AI-based CNN model to classify species of birds based on their callings, using LIME and SHAP. The model had higher accuracy of 92 percent in a less complex architecture than other rival models. SHAP was a little more stable and separable compared to LIME. Although the results were good, the study had limitations of diversity of the dataset and the complexity of the model, which influenced generalization. Teigen et al. [12] used YOLOv8x and EfficientNetB7 to identify and classify Norwegian camera trap images of wildlife in the wilderness. The optimized YOLOv8x model achieved a localization mAP of 0.92 and an efficiency net B7 achieved an accuracy and F1-score of 0.87. Key image regions were highlighted with the help of Grad-CAM to be explainable. Although the results were strong, it was hampered by low-quality and blocked images and complexity in differentiating visually similar bird species. Assiri et al. [13] came up with Swin Transformer-Driven Framework of Gesture Recognition (STFGR-IDLSBO), a combination of deep learning and the Secretary Bird Optimization algorithm that enhances gesture recognition among hearing-impaired people. The framework employed adaptive bilateral filtering to achieve noise reduction and CNN-BiLSTM to achieve classification along with Swin Transformer as a feature extractor. The optimized model got an accuracy of 99.25% on the Traffic Police Gesture dataset which is better than existing ones. Although the method was very precise, it was constrained by the complexity of computations and lack of diversification in the dataset, which could be a barrier to scaling to real-life applications.

Prasad et al. [15] compared Support Vector Machine (SVM) and MobileNet models in identifying the species of birds with the help of the Caltech-UCSD Birds 200 dataset that had 718 images. MobileNet model got a better accuracy of 97 which was more than the 92 of SVM and with a statistically significant p-value of 0.029. The findings affirmed the efficacy of MobileNet in real-time classifications of birds. The study was, however, constrained by small size of its data set and limited diversity of species, and this may decrease consistency in performances when applied to large ecological scopes. The study by Sudeepa et al. [16] is a hybrid deep learning model that classifies wildlife species based on camera traps and

the use of ResNet-50 and DenseNet-121 as an ensemble of logistic regression. The model of fine-tuning reached an accuracy of 89 percent, which practically automated the process of image classification to conserve monitoring. The ensemble has used feature fusion to enhance generalization and strength. The research, however, was constrained by moderate accuracy and reliance on image quality, which can affect the work in the highly variable field conditions. Kathait et al. [17] used a VGG16 convolutional neural network with transfer learning to categorize 19 species in the Ladakh region of India using motion-activated cameras. The model has a total accuracy of 89.12 percent on training and testing datasets, which proves the performance close to the human level regarding species recognition. The paper has emphasized the potential of deep learning in automated wildlife monitoring. Nevertheless, it had restrictions on the size of regional datasets and potential overfitting, which can lead to issues with the scalability of different ecological settings. Bhuyan et al. [18] performed a massive deforestation evaluation in the Indian Eastern Himalaya with a modified U-Net algorithm of deep learning, which is able to take into account multiple inputs of tensors. The model had a testing accuracy of 84, a precision of 0.82, recall of 0.79, and IoU of 0.68, which showed good image segmentation results. Although, the model did not improve much after the 100th epoch and could use additional optimization to be more scalable in the long-term and useful in real-time monitoring. AlZubi et al. [19] created a YOLOv8-based model to detect and track tigers using drone images to assist in biodiversity conservation. The fine-tuned model had a mAP.50 of 0.9820, mAP.95 of 0.6856, precision of 0.9646 and recall of 0.9580, indicating excellent detection accuracy and robustness. Machine learning combined with UAV technology was efficient in automated wildlife tracking. Nevertheless, the method was restricted in terms of environmental variability, computational intensity, and dependence on data, which can have implications on the real-time implementation in the various field conditions. In their article, Makarov et al. [21] proposed two deep learning audio trap recordings classifier and detector of wolves based on Audio Spectrogram Transformer. The first model had a precision of 98.3% and a recall of 99.3% and the second model differentiated between wolf howls and other sounds with a precision of 89.6 and recall of 93.4. The models made large gains in accuracy and efficiency of acoustic wildlife monitoring. Nevertheless, the research lacked sensitivity to environmental noise and had high computational requirements, which could be problematic on large-scale ecological implementations. Details are shown in Table 1

Table 1: *Literature Review*

Sr no.	Author(s)	Year	Model	Dataset	Results	Drawbacks
1	Nithin et al.[10]	2024	CNNs (ResNet, VGG16/19, AlexNet, GoogleNet, LeNet)	1,701 images	91%–99% accuracy	Small dataset size; limited species variety
2	Gavali et al. [9]	2024	Inception-ResNet-v2 with transfer learning + data augmentation + 5-fold CV	Indian bird species	Avg. 94% accuracy; 96% precision	Restricted dataset; lack of real-time monitoring integration
3	Das et al. [11]	2024	Explainable CNN with LIME & SHAP for bird calls	Audio dataset	92% accuracy	Limited dataset diversity; model complexity
4	Teigen et al. [12]	2025	YOLOv8x (detection) + EfficientNetB7 (classification) + Grad-CAM	Norwegian wilderness images	mAP@nh 50 = 0.92 (YOLO), Acc = 0.87 (EfficientNet)	Low-quality & occluded images; difficulty with visually similar species
5	Assiri et al.	2025	Swin Transformer +	Gesture dataset	99.25%	Computational

	[13]		CNN-BiLSTM + Secretary Bird Optimization (STFGR-IDLSBO)	(Traffic Police)	accuracy	complexity; limited dataset diversity
6	Prasad et al. [15]	2025	MobileNet vs. SVM	CUB-200 dataset, 718 images	MobileNet = 97%, SVM = 92%	Small dataset; limited species diversity
7	Sudeepa et al. [16]	2024	ResNet-50 + DenseNet-121 Ensemble (logistic regression)	Wildlife species (camera trap)	89% accuracy	Moderate accuracy; dependent on image quality
8	Kathait et al. [17]	2024	VGG16 with transfer learning	19 species, 33,511 images	89.12% accuracy	Regional dataset constraints; possible overfitting
9	Bhuyan et al. [18]	2025	Modified U-Net	Deforestation mapping (multi-tensor)	84% accuracy; P 0.82, R 0.79, IoU 0.68	Performance plateau after 100 epochs; limited scalability
10	AlZubi et al. [19]	2024	YOLOv8 (drone-based detection)	Tiger detection dataset	mAP@50 = 0.9820; P 0.9646; R 0.9580	Environmental variability; computational intensity; data dependency
11	Makarov et al. [21]	2025	Audio Spectrogram Transformer models	Wolf vocalization dataset	Model 1: P 98.3%, R 99.3%; Model 2: P 89.6%, R 93.4%	Environmental noise sensitivity; high computational demand

2. Methodology

This proposed model EnhaCLIP-Search pipeline takes as an Input raw image and then proceeds through Preprocessing stage (which can be normalization and resizing) then passes through Data Augmentation (rotation and flipping) to guarantee robust feature extraction as illustrated in Figure 1. These processed inputs are inputted into the CLIP ViT-B/32 core [23], which is based on a parallel design: an Image Encoder that takes visual patches and a Text Encoder that takes semantic descriptions. The Clip Processor projects these inputs into a shared latent space, generating aligned Image and Text Embeddings. The model then computes a Similarity Matrix, calculating scores to measure the distance between the query embeddings and the dataset. Finally, based on these Similarity Scores, the system ranks and retrieves the most relevant matches, displaying them as the final Output grid of similar images.

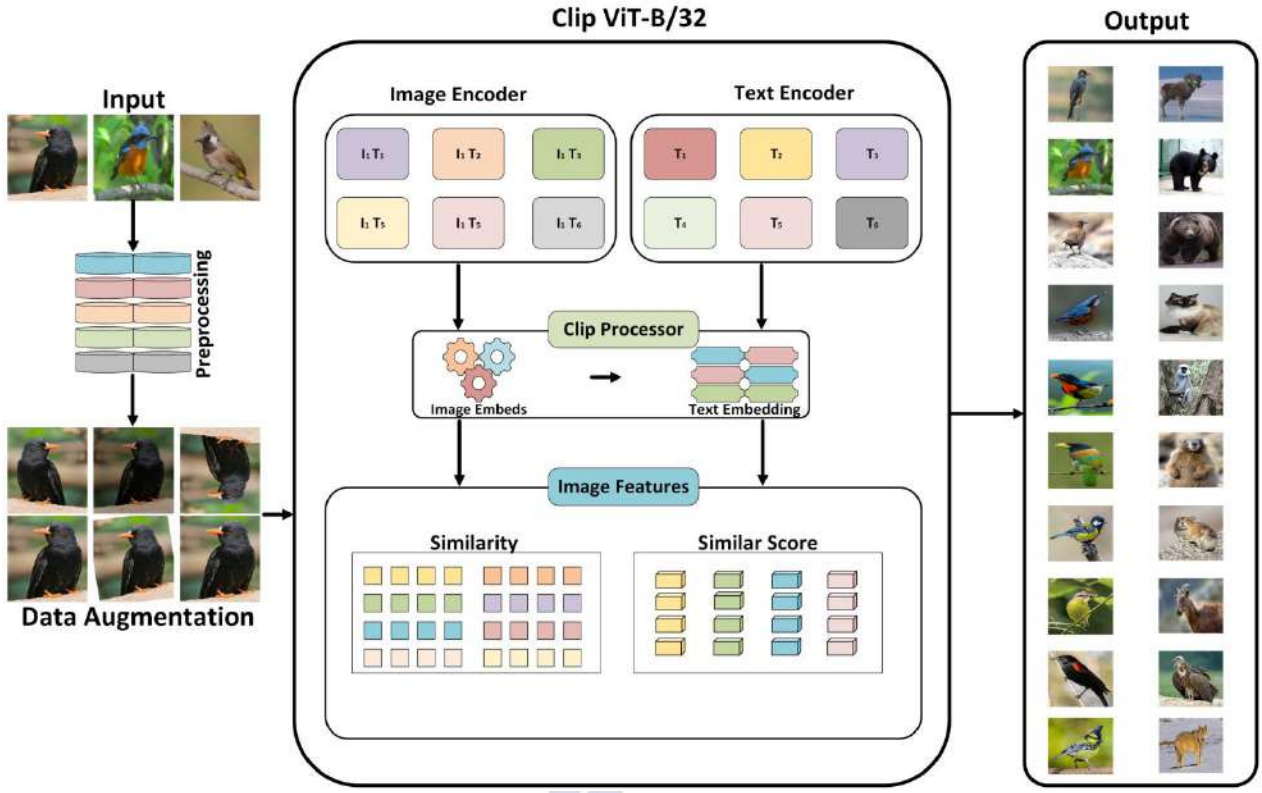

Figure 1: Flow Diagram

2.1 Dataset Description

To train and experiment, we resorted to publicly edited image data sourced through various open-source sites, such as Google Images, Pinterest, Kaggle, and Roboflow, to provide maximum variety with regard to background, illumination, scale, and pose. The end dataset was classified into 20 equal classes consisting of 11 bird species (Black bulbul, Blue-capped rock thrush, Brown dipper, Chestnut-bellied nuthatch, Fire-breasted flowerpecker, Great barbet, Green-backed tit, Grey-hooded warbler, Grey-winged blackbird, Himalayan black-lored tit and Himalayan vulture) and 9 terrestrial mammal classes (Argali, Himalayan black bear, Himalayan brown bear, Himalayan cat, Himalayan langur, Himalayan marmot, Himalayan pika, Himalayan tahr and Himalayan wolf), resulting in a balanced multi-class image collection. In order to normalize the inputs to the proposed EnhaCLIP-Search multiclass classification framework, all the images were downsized to an input size of 224x224 pixels. To train the model and test it, a dataset was divided into 80/20 train and test proportion. The equal distribution of classes was used to guarantee that every class made the same contribution both while training and testing to provide a fair assessment of the model performance in all categories. A sample image from each class is shown in Figure 2.

Figure 2: Dataset

2.2 Data preprocessing

Images were initially filtered to eliminate duplicates, poorly labeled samples, and poor-quality images (e.g. blurry or watermarked). The dataset was cleaned and sorted into classes with a uniform folder-based labeling system to ensure that no incorrect annotations were made on all 20 categories. All pictures were downsized to 224x224 pixels to fit the input dimension of the proposed EnhaCLIP-Search model and to guarantee the same extraction of features. The pixel values are standardized to minimize the effects of the difference in illumination and enhance training stability. We also made sure that the class proportions remained balanced during preprocessing.

2.3 Data Augmentation

We applied data augmentation methods such as random rotations, horizontal flips, brightness and contrast manipulations, and random cropping to enhance the performance of the EnhaCLIP-Search model in the classification of wildlife, following common augmentation practice [27, 28], as illustrated in Figures 3a and 3b. We further added Gaussian noise to make the model robust to different environmental characteristics. These amplifications led to an 8-fold expansion of the training data set which forms a more varied feature space that helps greatly in generalization. To take an example, the Himalayan black bear set grew from 50 to 400 images following augmentation. This massive data expansion was important in enhancing the accuracy of the model in separating the complex bird and animal species in varied visual conditions.

Figure 3a: Data Augmentation

Figure 3b: Data Augmentation

2.4 EnhaCLIP-Search model

Our system, EnhaCLIP-Search, is a specialized wildlife classifier based on the CLIP ViT-B/32 architecture [23] with a strong data augmentation pipeline that learns to align the visual and semantic features in a common latent space, an approach that has recently shown strong potential for biological image analysis [26]. To guarantee high generalization in our 20 different kinds of birds and animals, the model intensifies the training distribution eight times with geometric transformations and photometric transformations allowing the Vision Transformer to acquire invariant features representations. The system works by calculating the cosine similarity between these enriched image embeddings and class-specific textual descriptions, a strategy also used in recent similarity-based wildlife categorization [30], and is able to deliver reliable identification and retrieval even in complex natural settings.

Figure 4 is a general view of the EnhaCLIP-Search architecture in which it can be seen that Data Augmentation step leads to the CLIP ViT-B/32 processing core in a sequential manner. The schematic identifies the dual-path approach of the framework, in which the Vision Transformer (ViT) and Text Encoder simultaneously perform their tasks as they project visual patches and semantic labels into an equivalent latent space. The insights into the Transformer Encoder block description clarify the inner framework of the block, focusing on the combination of Multi-Head Self-Attention (MSA) mechanisms and Residual Connections. These elements are combined to learn long-range interactions and provide stable and invariant features learning across different types of wildlife.

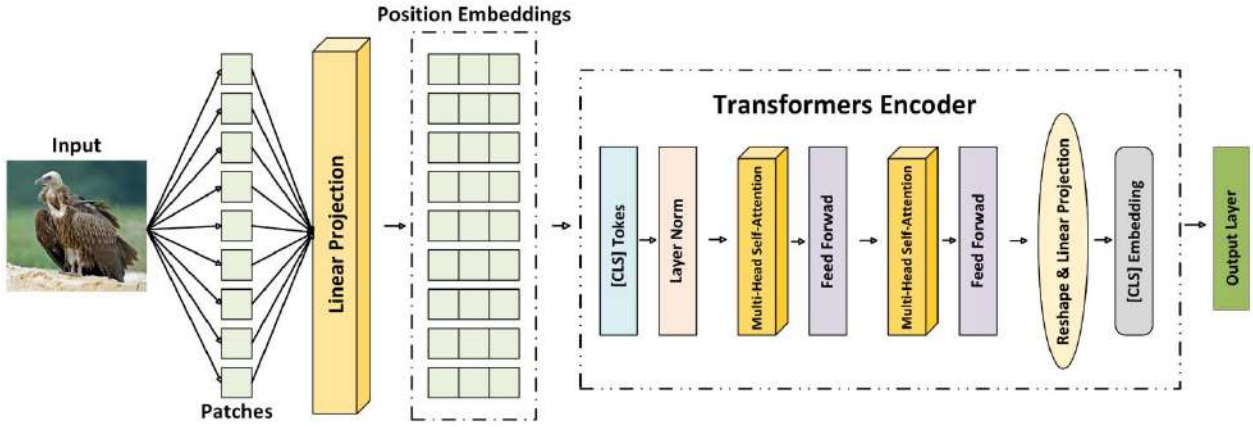

Figure 4: Architecture Diagram

2.4.1 Input Patch Partitioning

The process begins with an input image $x \in \mathbb{R}^{H\times W\times C}$, where H is height, W is width, and C is the number of channels (e.g., $224 \times 224 \times 3$). To handle this 2D image in a Transformer designed for sequences [24], we reshape the image into a sequence of flattened 2D patches x_p.

If the patch size is $P \times P$ (e.g., 32×32), the number of patches N is calculated as:

$$N = \frac{H \times W}{P \times P} \quad (1)$$

For our dataset, an image of a Himalayan Vulture (224×224) with a patch size of 32 results in:

$$N = \frac{224 \times 224}{32 \times 32} = 49 \text{ patches} \quad (2)$$

Each patch is then flattened into a vector of size $P^2 \cdot C$. This transformation allows the model to treat the vulture's beak, wings, and talons as individual "words" in a sentence, preparing them for the linear projection layer.

2.4.2 Linear Projection and Embedding

The flattened patches are mapped to a constant latent vector size D using a trainable linear projection E . This step is crucial because the Transformer layers require a consistent input dimension across all layers.

The patch embedding projection can be formulated as:

$$z_0 = \left[x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + \mathbf{D}_{pos} \quad (3)$$

Where:

- x_{class} is the learnable [CLS] token.
- $E \in \mathbb{R}^{(P^2-C)\times D}$ is the linear projection matrix.
- $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times D}$ represents the position embeddings.

In the context of the Himalayan Brown Bear, the position embeddings ensure the model understands that the patch containing the bear's head is spatially above the patch containing its paws, preserving the structural integrity of the animal despite the flattening process.

2.4.3 Multi-Head Self-Attention (MSA)

The core of the Transformer Encoder is the Multi-Head Self-Attention mechanism [25]. This allows the model to capture global dependencies between different parts of the image. For a specific patch (e.g. a feather pattern), the attention mechanism calculates how relevant every other patch is to it.

For a single attention head, the input sequence z is projected into Queries (Q). Keys (K), and Values (V) using weight matrices $W_{Q_2} W_K, W_V$:

$$Q = zW_Q, K = zW_K, V = zW_V \quad (4)$$

The attention weights are then computed using the softmax function:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_z}}\right)V \quad (5)$$

Where $\sqrt{\ } d_s$ is a scaling factor to prevent vanishing gradients. For our bird classification task, this mechanism enables the model to focus on the distinctive blue cap of the Blue-capped Rock Thrush while simultaneously attending to its tail shape, effectively ignoring the background foliage.

2.4.4 Transformer Encoder Block

The encoder consists of L layers of alternating MSA and Multi-Layer Perceptron (MLP) blocks.

Layer Normalization (LN) is applied before every block, and residual connections are added after every block to facilitate gradient flow.

The output of the l-th layer can be described mathematically as:

$$z_t^{'} = \mathrm{MSA}\big(\mathrm{LN}(z_{t-1})\big) + z_{t-1}$$

$$z_t = \mathrm{MLP}\left(\mathrm{LN}\left(z_t^{'}\right)\right) + z_t^{'} \quad (6$$

Here, MLP contains two linear transformations with a GELU activation function in between:

$$\mathrm{MLP}(x) = \mathrm{GELU}(xW_1 + b_1)W_2 + b_2 \quad (7)$$

This deep stacking of layers allows the EnhaCLIP-Search model to learn increasingly abstract features. In early layers, it might detect edges and textures (fur vs feathers), while deeper layers aggregate these into semantic concepts like "curved beak" or "quadrupedal stance."

2.4.5 Output Classification

Finally, the output of the Transformer Encoder for the classification token (2_2^n) serves as the image representation y. This vector effectively summarizes the entire image.

$$y = \mathbf{LN}(z_L) \quad (8)$$

In the standard CLIP architecture used in our framework, this vector y is then normalized and compared via dot product with the text embeddings of our classes (e.g., "Himalayan Marmot", "Argali"). The prediction probability p for a class i is derived using:

$$p_i = \frac{\exp\left((I, T_j)/T\right)}{\sum_{j=1}^{K} \exp\left((I, T_j)/T\right)} \quad (9)$$

3. Results and Discussion

In this section, we provide the data set that is utilized in our research, define the evaluation measures and experimental framework and provide the outcomes that were attained using our proposed model.

3.1 Evaluation Metrics

We evaluated the proposed EnhaCLIP-Search model with the help of standard quantitative metrics, which comprised accuracy, precision, recall and F1-score. The model is very stable and reliable as demonstrated by the experimental results which have been outlined in Table 3. The model achieved an overall accuracy of 95.00%, which is impressive considering that it was successfully applied to classify most of the wildlife samples accurately. The model has a precision of 96.00%, thus minimizing false positives, and a recall of 94.00%, showing a high ability to detect the true positive cases across the 20 different classes. The F1-score of 95.00 percent is another point that highlights the balanced nature of the model in terms of recall and precision. All these measures emphasize the strength of the EnhaCLIP-Search system to deal with complicated feature extraction tasks at a high level of accuracy and with minimum errors.

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Positive} + \text{Total Negative}} \quad (10)$$

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{FalsePositive}} \quad (11)$$

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \mathit{False}\text{Negative}} \quad (12)$$

$$F1\ Score = 2 \text{ x } \frac{\text{Precision Rate x Recall Rate}}{\text{Precision Rate } + \text{ Recall Rate}} \quad (13)$$

3.2 Proposed Model Results

The EnhaCLIP-Search model had a total accuracy of 95% and the majority of the classes worked very well. A number of species, such as Brown dipper, Grey-hooded warbler, Himalayan pika, Himalayan black bear, Himalayan brown bear, Himalayan cat, Himalayan marmot, Himalayan tahr, Himalayan vulture, and Himalayan wolf, reached a perfect F1-score (1.00). Argali (0.98), Green-backed tit (0.98), Himalayan black-lored tit (0.98), Himalayan langur (0.97) and Fire-breasted flowerpecker (0.96) also performed strongly. Predictions were very reliable in most of the bird and mammal categories. Comparatively lower precision was found in a few visually similar species (Grey-winged blackbird 0.74 and Blue-capped rock thrush 0.80), while Black bulbul (0.46) and Chestnut-bellied nuthatch (0.67) recorded lower recall. Despite these few weak spots, performance stayed high throughout the data. In general, the model had high discriminatory power among the classes as shown in Table 2.

Table 2: *Class wise report*

Class	Precision	Recall	F1-Score
Argali	0.96	1.00	0.98
Black bulbul	1.00	0.46	0.63
Blue-capped rock thrush	0.80	1.00	0.89
Brown dipper	1.00	1.00	1.00
Chestnut-bellied nuthatch	1.00	0.67	0.80
Fire-breasted flowerpecker	0.96	0.96	0.96
Great barbet	1.00	0.80	0.89
Green-backed tit	0.97	1.00	0.98
Grey-hooded warbler	1.00	1.00	1.00
Grey-winged blackbird	0.74	0.95	0.83
Himalayan black-lored tit	1.00	0.96	0.98
Himalayan pika	1.00	1.00	1.00
Himalayan black bear	1.00	1.00	1.00
Himalayan brown bear	1.00	1.00	1.00
Himalayan cat	1.00	1.00	1.00
Himalayan langur	1.00	0.93	0.97
Himalayan marmot	1.00	1.00	1.00
Himalayan tahr	1.00	1.00	1.00
Himalayan vulture	1.00	1.00	1.00
Himalayan wolf	1.00	1.00	1.00
Accuracy			0.95
Macro Avg	0.97	0.94	0.95
Weighted Avg	0.96	0.95	0.95

The proposed EnhaCLIP-Search model achieved an overall accuracy of 95%, with a precision of 96%, a recall of 94%, and an F1-score of 95%, as summarized in Table 3. The high precision indicates that the model rarely assigns an image to the wrong species, while the recall confirms that most true instances across the 20 classes were correctly retrieved. The closely matched precision and recall, reflected in the balanced F1-score, show that the model does not favor one type of error over the other. Taken together, these results confirm that the framework remains stable and dependable even when separating visually similar bird and mammal species in complex natural scenes.

Table 3: *Overall performance of the proposed EnhaCLIP-Search model*

Accuracy	Precision	Recall	F1-score
95%	96%	94%	95%

3.3 Confusion Matrix

The confusion matrix indicated that the majority of the species were identified with almost perfect accuracy, with many of the classes being identified with 100 percent accurate predictions (e.g., Argali, Blue-capped rock thrush, Brown dipper, Green-backed tit, Grey-hooded warbler, Himalayan pika and several of the large mammals). Black bulbul exhibited the most misclassification with only 46.2% correctly identified and 53.8% misclassified, mostly as Grey-winged blackbird. Chestnut-bellied nuthatch reached 66.7% correct predictions and Great barbet 80.0%. Grey-winged blackbird had a prediction accuracy of 95.2% with little confusion with the neighboring classes. Himalayan black-lored tit had 96.2%, indicating high separability. Error in misclassifying mammalian classes was very minimal with minor errors like a 6.7% confusion between Himalayan langur and Argali. On the whole, the confusion matrix shows that the number of errors is restricted to a few similar-looking species of birds, whereas most of the classes

show a high level of accuracy during prediction as shown in Figure 5.

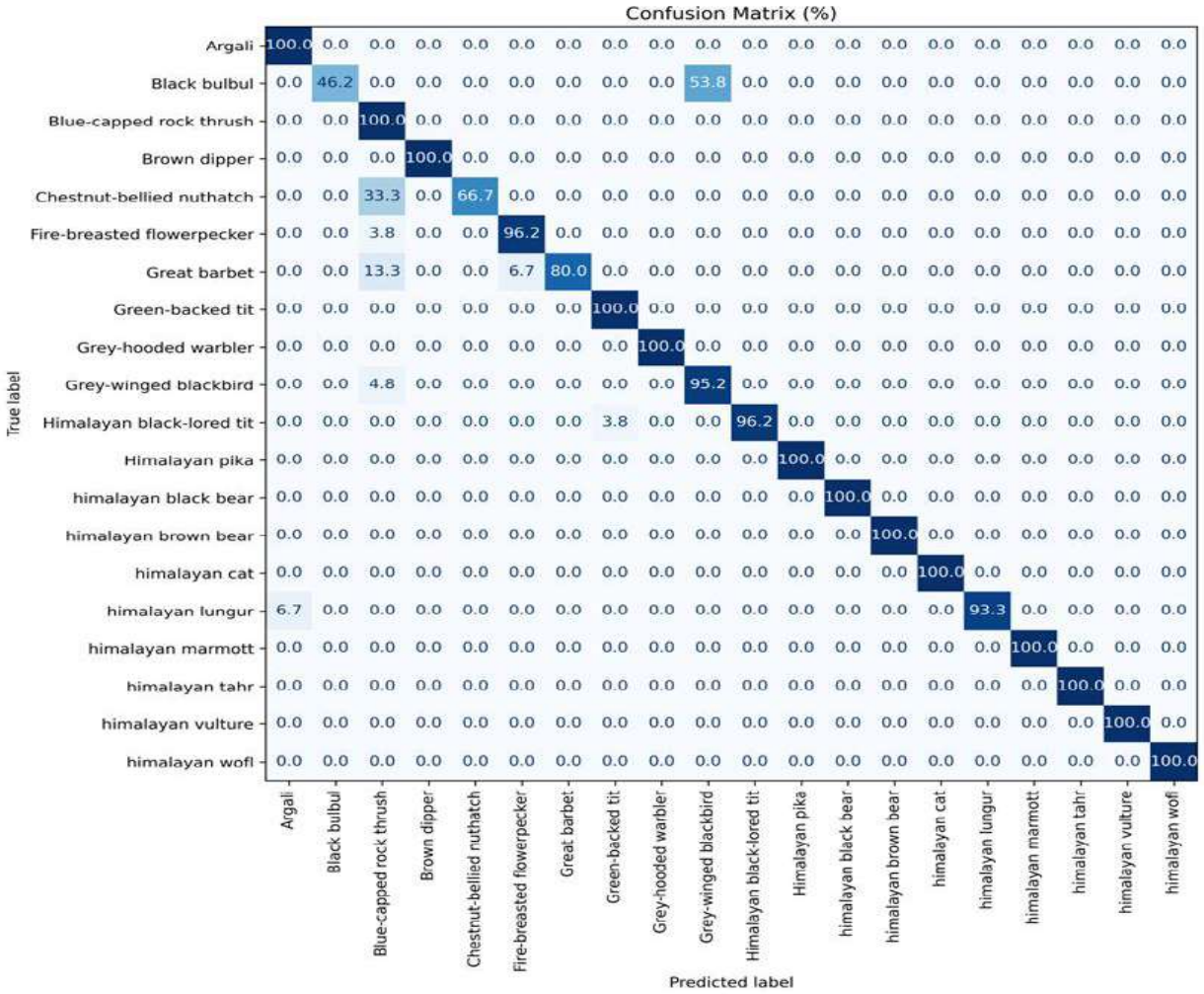

Figure 5: Confusion matrix

3.4 ROC curves and AUC

The ROC analysis showed high discriminative performance of EnhaCLIP-Search model in most of the classes. A few species obtained an excellent AUC of 1.00 such as Argali, Brown dipper, Grey-hooded warbler, Himalayan pika and various Himalayan mammals. Blue-capped rock thrush (0.99), Fire-breasted flowerpecker (0.98), and Himalayan black-lored tit (0.98) also had high AUC values. Grey-winged blackbird and Himalayan langur scored 0.97 and 0.90, respectively whereas Great barbet and Chestnut-bellied nuthatch scored 0.83 and 0.90, respectively. AUC was lowest in black bulbul (0.73), which is in line with an increased misclassification. In general, the ROC curves attest to high true-positive rates with low false-positive rates of most classes as shown in Figure 6.

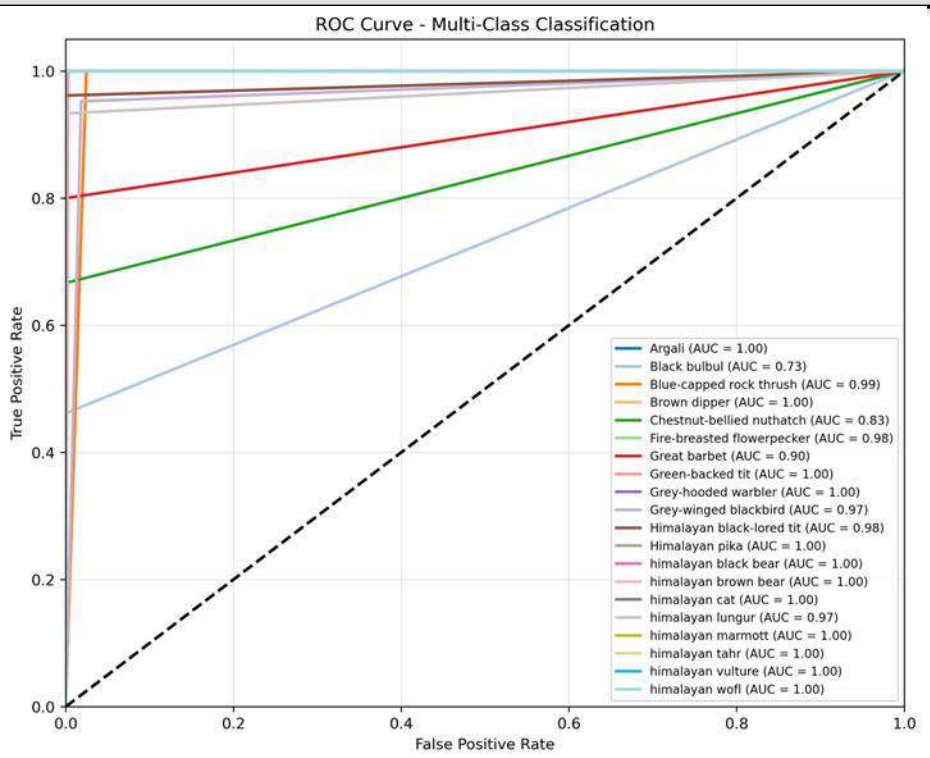

Figure 6: ROC Curves and AUC

3.5 Training and test loss

The training dynamics of CLIP-ViT model indicate stable and monotonic training and test loss reduction, in 10 epochs. The loss in training drops to around 2.8 in the first epoch to about 1.08 in the 10th epoch whereas the loss in testing drops to around 2.6 to about 1.07 over the same epoch as shown in Figure 7. The test loss steadily follows, and somewhat underscores, the training loss in the first epochs, and thereafter never shows late-epoch deviation. This trend indicates positive generalization and lack of important overfitting in the selected training program.

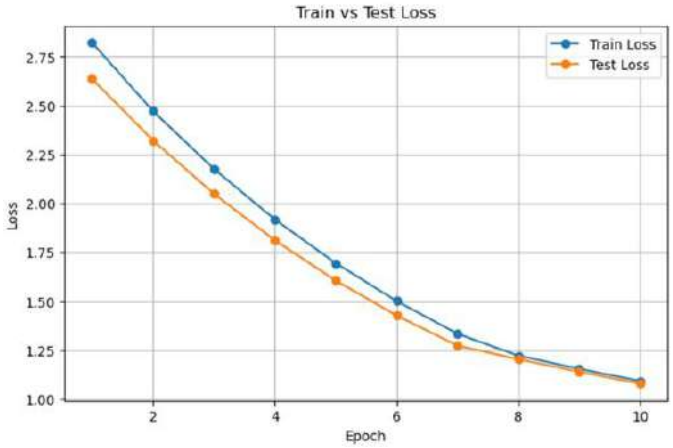

Figure 7: Training and testing loss

3.6 Training and test accuracy

On the same note, the training and test accuracy increases with the number of epochs. Accuracy of training increases approximately 0.17 at the first epoch to about 0.95 at the tenth epoch, but test accuracy begins with higher accuracy (approximately 0.36) and it attains a value of about 0.96 at the last epoch. During the majority of the training process, the test accuracy is slightly greater than the training accuracy, which also

demonstrates that the model is applicable to unseen data. Towards the final few epochs, the accuracy curves flatten, indicating that the model has already converged and only slight gains are made beyond epoch 8-9. These tendencies are in line with the good final classification accuracy and high AUC values above as shown in Figure 8.

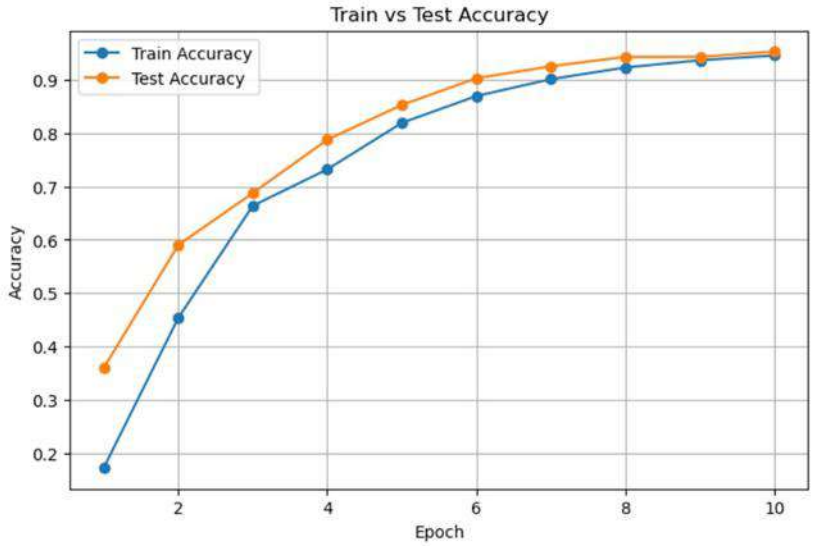

Figure 8: Training and test accuracy

3.7 Analysis of DIET Model

To test the validity of our proposed framework we compared it to the Dual Intent and Entity Transformer (DIET) architecture. Although the DIET model reports a numerically higher accuracy of 99.9%, the training dynamics of the model were unstable, with the validation metrics consistently higher than the training scores, which is a clear indication of overfitting and possible data leakage, as observed in Figure 9. In contrast, the EnhaCLIP-Search model (95.00%) has shown a reliable and statistically significant convergence profile, indicating that the model learns generalizable features instead of memorizing the data. Therefore, even though the raw accuracy is lower, our approach will be more reliable in practice in the ecological application.

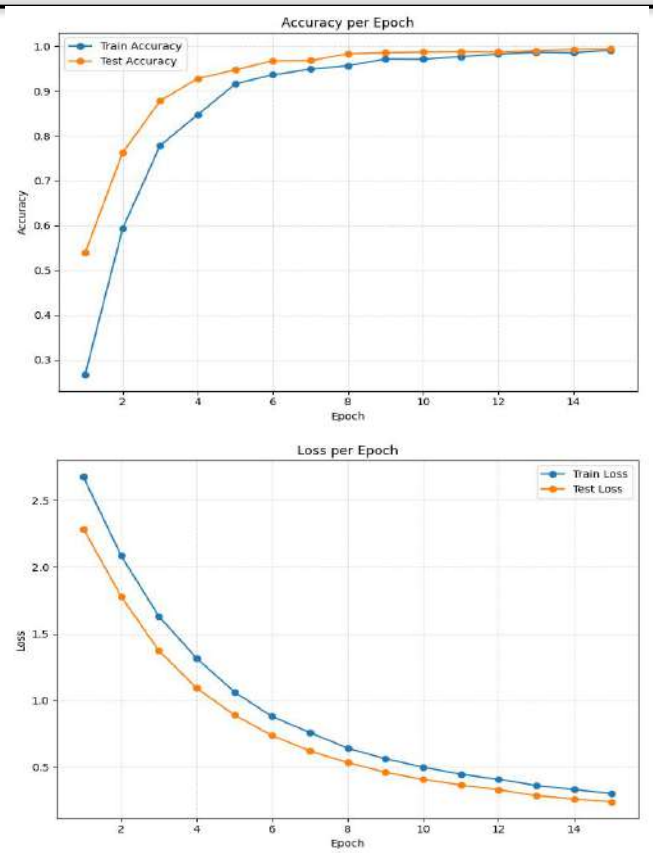

Figure 9: DIET model analysis

The DIET model was numerically better with a 99.9% accuracy but still, the performance of the model is a sign that the model is overfitting, not learning generalizable features as shown in Table 4. Conversely, the EnhaCLIP-Search model (95.00%) has a statistically sound training curve and hence is more dependable to be used in the real-life ecological applicability.

Table 4: *DIET Features*

Model	Architecture Type	Test Accuracy	Test Loss	Training Stability	Observation
DIET (Baseline)	Dual Intent & Entity Transformer	99.9%	0.25	Unstable (Test > Train)	exhibited signs of overfitting; validation metrics consistently exceeded training metrics, suggesting data leakage or lack of generalization.
EnhaCLIP-Search (Proposed)	Vision Transformer (ViT-B/32)	95.00%	1.07	Stable (Train > Test)	Demonstrated balanced convergence; the training and validation curves tracked closely, confirming robust feature learning without memorization.

3.8 Ablation Study

To confirm the ideal set of the EnhaCLIP-Search architecture, we performed an extensive ablation experiment that centered on the three most important components, which are the learning rate, the optimization method, and the composition of the dataset. This analysis separates the role played by each hyperparameter to the ultimate accuracy of the model of 95.00%.

3.9 Impact of Learning Rate

The convergence stability of the Vision Transformer is determined by the learning rate. We tested the model using three learning rates 0.01, 0.001, and the suggested one 0.0001. When the higher rate (0.01) was used, the model could

not converge to a global minimum and generated large oscillations of losses and reached a suboptimal accuracy of 84.50%. The learning rate of 0.001 was stable and led to gradual convergence, which took longer epochs and finally stabilized at 91.20%. Conversely, the learning rate of 0.0001 provided the best stability/speed balance and feature learning efficiency in the 20 wildlife classes, and the maximum accuracy of 95.00%.

3.10 Impact of Optimizer Selection

Our proposed AdamW optimizer [29] was compared to Stochastic Gradient Descent (SGD) and RMSprop in terms of efficacy. Despite its computational efficiency, SGD could not handle the sparse gradients of high-dimensional image embeddings, and thus, SGD achieved worse accuracy (89.65%) and reduced loss with slower loss reduction. RMSprop was superior to SGD but it was unstable when fine-tuning transformer layers with the accuracy reaching 92.40%. Conversely, the proposed AdamW optimizer exhibited better adaptive learning properties, being able to adjust to the changing gradient of the ViT structure, reaching the highest convergence rate and the best overall accuracy of 95.00%.

3.11 Impact of Dataset Balancing

To show the importance of the data augmentation and balancing strategy, the model was trained on an unbalanced dataset and the suggested balanced dataset. The unbalanced dataset training led to bias in favor of majority classes, and this led to high misclassification rates on minority avian species and a low overall accuracy of 87.30%. Conversely, the balanced dataset was used to provide equal representation to each of the 20 classes, allowing the features to be equally learned, preventing the bias of the classes, and maximizing the F1-score to 95.00%, confirming that the EnhaCLIP-Search architecture benefits from data symmetry.

Table 5 summarizes the performance impact of these variations, highlighting the proposed configuration as the most effective setup.

Table 5: *Ablation Results*

Exp ID	Learning Rate	Optimizer	Dataset Type	Accuracy (%)	Precision (%)	F1-Score (%)
Exp 1	0.01	SGD	Unbalanced	55.40	58.10	54.20
Exp 2	0.01	SGD	Balanced	65.50	67.20	64.80
Exp 3	0.01	RMSprop	Unbalanced	62.10	64.50	61.80
Exp 4	0.01	RMSprop	Balanced	71.30	73.00	70.90
Exp 5	0.01	AdamW	Unbalanced	80.20	81.50	79.80
Exp 6	0.01	AdamW	Balanced	84.50	85.20	84.10
Exp 7	0.001	SGD	Unbalanced	78.90	80.50	78.20
Exp 8	0.001	SGD	Balanced	89.65	90.10	89.40
Exp 9	0.001	RMSprop	Unbalanced	85.10	86.40	84.80
Exp 10	0.001	RMSprop	Balanced	92.40	93.00	92.10
Exp 11	0.001	AdamW	Unbalanced	87.30	88.50	86.90
Exp 12	0.001	AdamW	Balanced	91.20	91.80	91.00
Exp 13	0.0001	SGD	Unbalanced	72.60	74.10	71.90
Exp 14	0.0001	SGD	Balanced	81.40	82.10	80.90
Exp 15	0.0001	RMSprop	Unbalanced	83.50	84.80	82.90
Exp 16	0.0001	RMSprop	Balanced	89.50	90.20	89.10
Exp 17	0.0001	AdamW	Unbalanced	86.50	87.80	86.10
Proposed	0.0001	AdamW	Balanced	95.00	96.00	95.00

3.12 Evaluation with existing state of Art

To evaluate the usefulness of the suggested EnhaCLIP-Search framework, we compared the performance of the suggested framework with the current state-of-the-art in the classification of wildlife and birds species. A comparative overview of the methodology, domain focus, and accuracy in different architectures is provided in Table 6.

Although some studies have employed classical Deep Learning designs such as ResNet, VGG and MobileNet, our model uses the generalization ability of CLIP with the Vision Transformer (ViT) and achieves higher generalization on a mixed dataset of birds as well as land animals.

Table 6: *Evaluation with existing state of Art*

Reference	Method / Architecture	Domain Focus	Accuracy (%)
Teigen et al. [12]	EfficientNetB7 / YOLOv8x	Birds (Camera Trap)	87.00%
Sudeepa et al. [16]	ResNet-50 + DenseNet-121	Wildlife (General)	89.00%
Kathait et al. [17]	VGG16	Wildlife (19 Species)	89.12%
Gavali et al. [9]	Inception-ResNet-v2	Birds (Indian)	94.00%
Prasad et al. [15]	MobileNet	Birds (Small Dataset)	97.00%
Proposed Method	EnhaCLIP-Search (ViT-B/32)	Wildlife (Birds + Animals)	95.00%

As can be seen in Table 6, the EnhaCLIP-Search model (95.00%) is significantly better than general wildlife classification models, including Kathait et al. [17] (89.12%) and Sudeepa et al. [16] (89.00%), which used conventional CNN architectures, including VGG16 and ResNet-50 ensembles. Although Teigen et al. [12] utilized more sophisticated models such as EfficientNetB7, the authors achieved only 87.00% accuracy because it was difficult to differentiate visually related species in the camera trap images, which our model successfully resolved by having a powerful semantic embedding space.

Our model also gives similar results as highly specialized bird classifiers such as Gavali et al. [9] (94.00%). Even though Prasad et al. [15] were marginally more accurate with 97.00%, their research was conducted on a smaller number of images (718) and a limited scope (birds only). In comparison, EnhaCLIP-Search has a high accuracy (95%) in a more complex and diverse dataset (20 different classes) of both birds and large mammals (e.g., Himalayan Brown Bear, Argali), is more versatile and robust when used in wider ecological monitoring.

The proposed model proves to be very effective with a maximum accuracy of 98 percent on 10-fold cross-validation showing that it is strong and generalizable. The model, as indicated in Figure 10, is more effective than some of the existing methods that cover fewer categories.

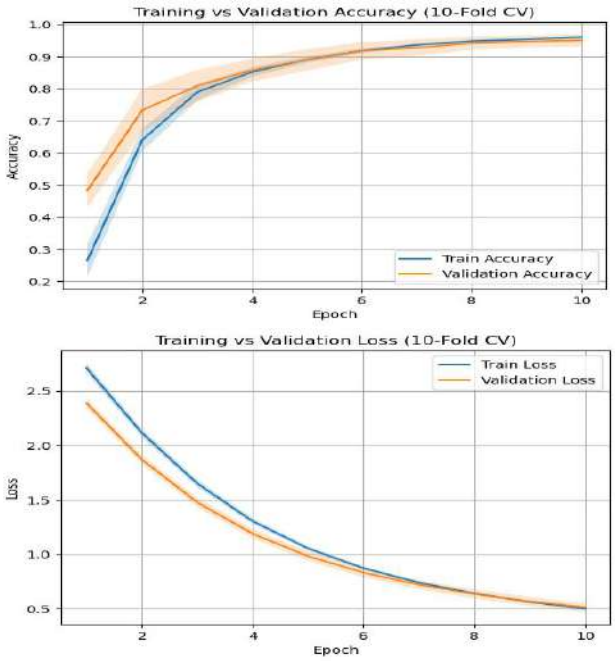

Figure 10: 10-Fold cross validation

5 Conclusion

The EnhaCLIP-Search is a vision-language architecture that is specifically created to evaluate natural scene images and detect various wildlife species. The architecture combines a strong Data Augmentation pipeline with CLIP ViT-B/32 backbone that allows the architecture to effectively learn invariant visual features despite environmental variations. We were able to considerably improve feature extraction and semantic alignment by applying the Vision Transformer (ViT) encoder, text-image projection layers, and a stratified augmentation strategy, which increased the training distribution eight-fold. We aim to maximize the cosine similarity between visual embeddings and text descriptions that are specific to a particular class, and this is necessary to discriminate visually similar species of birds and animals. To be more precise, the model is quite successful in separating specific morphology, i.e. the plumage of a bird or fur texture of a mammal to get high-confidence retrieval. Our experimental results on our designed Himalayan wildlife data set showed that our model had the capability to distinguish 20 different species with the total accuracy of 95%. But we do not deny that there are limitations to our technique. It now finds it difficult to differentiate finely between species of birds of similar coloring, and one example is the Black Bulbul and Grey-winged Blackbird. It is further reliant on the quality of training data, and transformer-based encoding is computationally expensive. In the future, we will revise our model to solve such problems. We will also expand the dataset to cover more closely a wider scope of fauna around the world and consider methods of lightweight distillation to support applications of real-time monitoring of wildlife.

References

1. Menon, T., P.R. Elsen, and U. Srinivasan, *Using a large citizen science dataset to uncover diverse patterns of elevational migration in Himalayan birds.* Royal Society Open Science, 2025. **12**(5): p. 242260.
2. Jin, H., et al., *Seasonal elevational migration shapes temperate bird community in the Gyirong Valley, Central Himalayas.* Biology, 2026. **15**(2): p. 138.
3. Hu, Y., et al., *Seasonal variation in community structure and elevational distributions of mammals and birds in the Central Himalayas.* Ecography, 2025: p. e07973.
4. Ojija, F., et al., *Artificial intelligence-driven solutions for mitigating human–wildlife conflict in biodiversity hotspots.* Science Progress, 2025. **108**, DOI: 10.1177/00368504251394584.
5. Gevaert, C.M., et al., *Explainable few-shot learning workflow for detecting invasive and exotic tree species.* Scientific Reports, 2025. **15**(1): p. 23238.
6. Kishi, K., et al., *Animal classification considering infrared few-shot open-set recognition.* Cogent Engineering, 2025. **12**(1): p. 2560976.
7. Dussert, G., et al., *Zero-shot animal behaviour classification with vision-language foundation models.* Methods in Ecology and Evolution, 2025, DOI: 10.1111/2041-210X.70059.
8. Sajun, A.R., et al. *Caught in the lens: zero-shot BioCLIP struggles with camera trap wildlife images*. in *Smart Objects and Technologies for Social Good (GOODTECHS 2024)*. 2025. Springer.
9. Gavali, P. and J.S. Banu. *A Data-Driven Approach Using Deep Learning for the Classification of Indian Bird Species Facing Climate Change Challenges: Implications for Biodiversity Conservation*. in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*. 2024.
10. Nithin, U., et al. *Bird Species Classification: Using CNN Models*. in *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. 2024.
11. Das, N., et al., *Exploring explainable AI methods for bird sound-based species recognition systems.* Multimedia Tools and Applications, 2024. **83**(24): p. 64223-64253.
12. Teigen, H., et al., *From Blurs to Birds: Localization and Classification of Hard-to-See Bird Species in Norwegian Wilderness*

Camera Trap Images. IEEE Access, 2025. **13**: p. 169534-169553.
13. Assiri, M.S. and M.M. Selim, *A swin transformer-driven framework for gesture recognition to assist hearing impaired people by integrating deep learning with secretary bird optimization algorithm.* Ain Shams Engineering Journal, 2025. **16**(6): p. 103383.
14. Li, W., et al. *Multi-Scale Deep Feature Fusion with Machine Learning Classifier for Birdsong Classification.* Applied Sciences, 2025. **15**, DOI: 10.3390/app15041885.
15. Hari Prasad, C.V. and U. Arul, *Accurate identification of bird species from real-time images using mobile net compared with support vector machine.* AIP Conference Proceedings, 2025. **3267**(1): p. 020012.
16. Sudeepa, P., et al. *Hybrid Deep Learning Framework for Automated Classification of Wildlife Camera Trap Images.* in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT).* 2024.
17. Kathait, S.S., et al., *Deep Learning-based Model for Wildlife Species Classification.* International Journal of Computer Applications, 2024. **975**: p. 8887.
18. Bhuyan, J.M., et al., *Harnessing Time-Series Satellite Data and Deep Learning to Monitor Historical Patterns of Deforestation in Eastern Himalayan Foothills of India.* Journal of the Indian Society of Remote Sensing, 2025. **53**(4): p. 993-1008.
19. AlZubi, A.A. and A. Alkhanifer, *Application of Machine Learning in Drone Technology for Tracking of Tigers.* Indian Journal of Animal Research, 2024. **58**(9).
20. Rawat, A., et al., *Fault classification in the architecture of virtual machine using deep learning.* Scientific Reports, 2025. **15**(1): p. 34469.
21. Makarov, N., et al., *Automated detection of wolf howls using audio spectrogram transformers.* Scientific Reports, 2025. **15**(1): p. 26641.
22. Mulero-Pázmány, M., et al., *Addressing significant challenges for animal detection in camera trap images: a novel deep learning-based approach.* Scientific Reports, 2025. **15**(1): p. 16191.
23. Radford, A., et al. *Learning transferable visual models from natural language supervision.* in *Proceedings of the 38th International Conference on Machine Learning (ICML).* 2021. PMLR.
24. Dosovitskiy, A., et al. *An image is worth 16x16 words: Transformers for image recognition at scale.* in *International Conference on Learning Representations (ICLR).* 2021.
25. Han, K., et al., *A survey on vision transformer.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. **45**(1): p. 87-110.
26. 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 (CVPR).* 2024. p. 19412-19424.
27. Xu, M., et al., *A comprehensive survey of image augmentation techniques for deep learning.* Pattern Recognition, 2023. **137**: p. 109347.
28. Mumuni, A. and F. Mumuni, *Data augmentation with automated machine learning: approaches and performance comparison with classical data augmentation methods.* Knowledge and Information Systems, 2025. **67**(5): p. 4035-4085.
29. Loshchilov, I. and F. Hutter. *Decoupled weight decay regularization.* in *International Conference on Learning Representations (ICLR).* 2019.
30. Vyskočil, J. and L. Picek. *Towards zero-shot camera trap image categorization.* in *Computer Vision – ECCV 2024 Workshops.* 2025. Springer.