Wildlife Re-Identification using MegaDescriptor

Abhinav Ujjawal
2021120
abhinav21120@iiitd.ac.in

Abhishek Sushil
2021441
abhishek21441@iiitd.ac.in

Aditya Jain
2021511
aditya21511@iiitd.ac.in

Abstract

Wildlife re-identification is a challenging task with applications in conservation and wildlife monitoring. In this report, we explore the use of MegaDescriptors with Swin Transformer models for wildlife re-identification. We trained the models on three datasets (LionData, MacaqueFaces, FreisianCattle2015) and tested them on the MPDD dataset. Our results demonstrate the effectiveness of the approach, achieving high accuracy in re-identifying wildlife species.

1. Introduction

Wildlife Re-Identification involves identifying individual animals from images, which is crucial for conservation efforts and population monitoring. Traditional methods rely on manual identification, which is time-consuming and error-prone. Automated re-identification using deep learning models offers a promising solution to this problem.

2. Problem statement

This project aims to explore and evaluate the performance of two MegaDescriptors (Swin image feature models), on wildlife datasets. Specifically, we will replicate the results of recent papers in this domain by conducting inference using pre-trained models on one of the datasets used in the papers. Additionally, we will design and execute an analysis experiment to assess the performance differences and types of errors between the two approaches using an alternative wildlife dataset.

3. Literature Review

3.1. Shifting Window Transformers

Why do we need SWin Transformers? Attempt to adapt Transformers as the backbone for computer vision, as it does for NLP and as CNNs do in vision.

Why has this previously been difficult? Scale : Unlike the word tokens that serve as the basic elements of processing in language Transformers, visual elements can vary substantially in scale. This is a problem in tasks like object detection. Resolution : Much higher resolution of pixels in images compared to words in passages of text. Problem of computation in dense tasks like semantic segmentation.

How to overcome these limitations? SWin Transformer architecture that constructs hierarchical feature maps and has linear computational complexity to image size. It constructs a hierarchical representation by starting from small-sized patches and gradually merging neighbouring patches in deeper Transformer layers. Thus it can conveniently leverage advanced techniques for dense prediction such as feature pyramid networks or U-Net. It computes self-attention locally within non-overlapping windows that partition an image. The number of patches in each window is fixed, and thus the complexity becomes linear to image size.

How is it different from an earlier Sliding Window approach? It shifts the window partition between consecutive self-attention layers. This is useful as it provides connections between layers (as most patches remain common between a shift). It is also efficient in regards to real-world latency: all query patches within a window share the same key set. This is in contrast to the sliding window as it had different key sets for different query pixels.[2]

The research paper WildlifeDatasets[1] introduces the WildlifeDatasets toolkit, an open-source Python library designed primarily for ecologists and researchers. Its main objective is to streamline access to publicly available wildlife datasets and provide a wide range of tools that can be used for pre-processing, performance analysis, and model fine-tuning. The toolkit is showcased through various scenarios and baseline experiments, including a comprehensive comparison of datasets and methods for wildlife re-identification.

The paper identifies the importance of animal re-identification in various aspects of wildlife study, such as population monitoring, migration patterns, behavioural studies, and wildlife management. It highlights the need for automated methods due to the increasing size of collected data and the demand for reducing labour-intensive manual processing. However, it notes a lack of standardization in algorithmic procedures, evaluation metrics, and dataset uti-

lization across the literature, hindering comparability and reproducibility.

In response to these challenges, the WildlifeDatasets toolkit is developed, offering features like easy access to wildlife datasets, advanced dataset splitting, and accessible feature extraction and matching algorithms. The toolkit includes methods based on local descriptors and deep learning approaches, with the introduction of MegaDescriptor, a foundation model for individual re-identification across species. MegaDescriptor achieves state-of-the-art performance and is available through the HuggingFace hub.

The related work section discusses existing methods and datasets for automated animal re-identification, highlighting the need for improved standardization and reproducibility. It categorizes approaches into local descriptors, deep descriptors, and species-specific methods, emphasizing the importance of addressing these issues to advance the field.

A methodology section details the approach to developing MegaDescriptor, focusing on local features and metric learning approaches. Ablation studies are conducted to validate design choices and select optimal configurations. Results from these studies inform the performance evaluation, where MegaDescriptor consistently outperforms other methods on all datasets.

Further analysis explores MegaDescriptor's performance in seen and unseen domains, demonstrating its ability to generalize beyond the datasets it was trained on. The conclusion summarizes the contributions of the WildlifeDatasets toolkit and MegaDescriptor, anticipating their widespread use in wildlife conservation and research.

Triplet loss compares a triplet of images: an anchor image x_a, a positive image x_p, and a negative image x_n. The anchor image x_a shares the same label as the positive image x_p but differs from the negative image x_n. Triplet loss aims to learn an embedding space where the distance between x_a and x_p is smaller than between x_a and x_n by at least a margin m. The model is trained to learn embeddings such that similar images have a low distance, while dissimilar images have a large distance.

ArcFace adds an angular margin to the standard softmax loss to improve the discriminative power of the learned embeddings. The embeddings are normalized and scaled to place them at a hypersphere of radius s which is selected as hyperparameter.

Overall, this paper presents a comprehensive toolkit and methodology for wildlife re-identification, addressing key challenges in the field and offering a significant advancement in automated animal identification techniques.

4. Dataset Description

Wildlife Dataset, published in year 2024, is a collection of 33 publicly available dataset each containing one or more animals[1]. It also provides utilities to mass download and convert them into a unified format. We used 3 datasets from this collection for verifying the performance of the model which are LionData, MacaqueFaces, FreisianCattle2015. LionData contain images of lions collected from Mara Masia project in Kenya. It includes images containing various lion details such as ears or noses. Two examples of LionData are shown below. MacaqueFaces dataset shows the faces of group-housed rhesus macaques at a breeding facility in large indoor enclosures. Their face was semi-automatically extracted from high-definition videos. FriesianCattle2015 dataset captures images of Holstein-Friesian cows from an aerial standpoint which are extracted from videos.

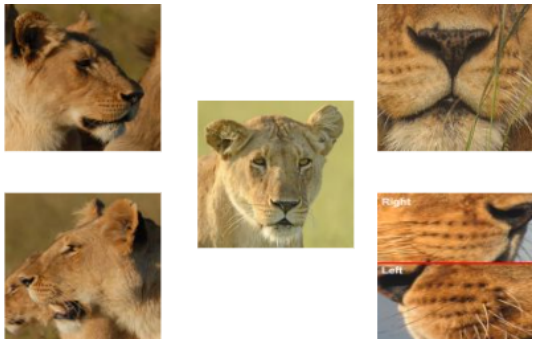


Figure 1. Charm from LionData

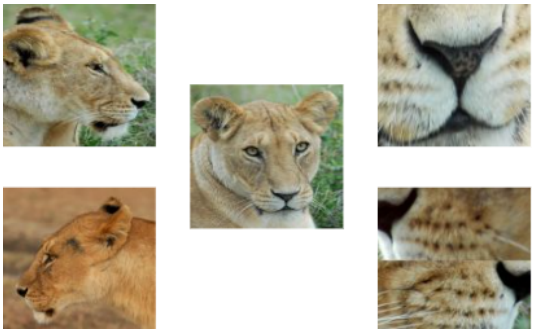
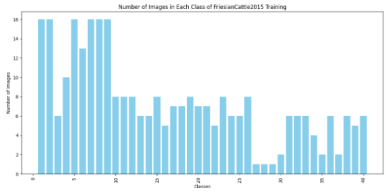


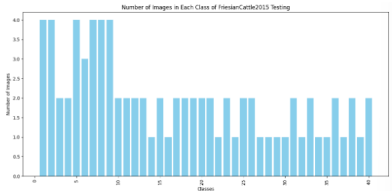
Figure 2. Cleopatra from LionData

5. Approach

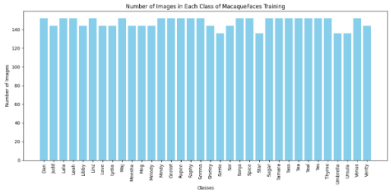
We utilized the L-224 and T-224 models, which are Swin Transformer variants, for our experiments. These models were pre-trained on a collection of large-scale wildlife datasets but we are verifying the results on only three datasets which are LionData, MacaqueFaces, FreisianCattle2015. We used the MPDD(Multi-pose dog dataset) for testing. The code for the T-224 model can be found HERE. The code for the L-224 model can be found HERE.



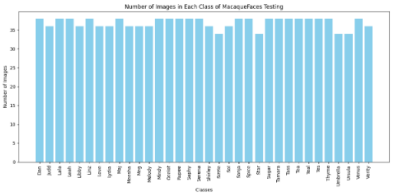
(a) Bar graph for training for cattle dataset



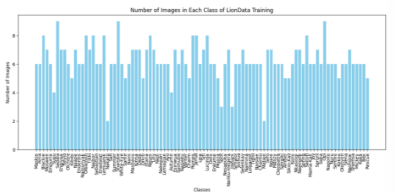
(b) Bar graph for testing for cattle dataset



(c) Bar graph for training for macaque dataset



(d) Bar graph for testing for macaque dataset



(e) Bar graph for training for lion dataset

(f) Bar graph for testing for lion dataset

Figure 3. Bar graphs for training and testing for different datasets

6. Results of the Paper

We were able to replicate the results which were mentioned in the paper. The L-224 model shows better results than the T-224 model in terms of accuracy and F1-score but was computationally more expensive.

Based on the experimental results, it can be observed that the performance of the Swin Transformer models varies across different datasets. The tables below summarizes the accuracy, F1 score, and sensitivity of the models on three datasets: FriesianCattle2015, LionData, and MacaqueFaces.

'Accuracy': 0.8413173652694611, 'F1 Score': 0.8284107110454416, 'Sensitivity (Recall)': 0.8413173652694611

Dataset	Accuracy	F1 Score	Sensitivity
FriesianCattle2015	0.55	0.543	0.55
LionData	0.148	0.140	0.148
MacaqueFaces	0.990	0.990	0.990

Table 1. Results of the experiments on different datasets for T-224.

Dataset	Accuracy	F1 Score	Sensitivity
FriesianCattle2015	0.55	0.542	0.55
LionData	0.206	0.188	0.206
MacaqueFaces	0.990	0.990	0.990

Table 2. Results of the experiments on different datasets for L-224.

The MacaqueFaces dataset demonstrates exceptional performance, with an accuracy, F1 score, and sensitivity of approximately 0.99 across both models which is consistent across multiple papers. This indicates that the model was able to accurately reidentify individual animals from the dataset.

In contrast, the LionData dataset shows lower performance in both models, with an accuracy of only 0.206 in L-224 and an accuracy of 0.148 in T-224. This could be attributed to the challenges associated with identifying individual lions based on their features, which may be less distinctive compared to other species which is also taken as hypothesis testing mentioned later.

The FriesianCattle2015 dataset falls in between, with an accuracy of 0.55 across both models. While this is higher than the LionData dataset, it is still lower than the performance on the MacaqueFaces dataset.

On the testing dataset MPDD, the accuracy, F1-score and sensitivity for the model T-224 was 0.841 and for L-224 it was 0.919, which shows that L-224 results are better than T-224 on a dataset which is unseen by the model and did not involve any fine-tuning.

Model	Accuracy	F1 Score	Sensitivity
T-224	0.841	0.828	0.841
L-224	0.919	0.914	0.919

Table 3. Results of the experiments on different models for MPDD.

Overall, these results highlight the importance of dataset selection and the challenges associated with wildlife reidentification. Further research and experimentation may be required to improve the performance of the model on challenging datasets such as LionData.

7. Discussion

Our approach highlights the potential of MegaDescriptors and Swin Transformer models in wildlife re-identification. The ability to accurately identify individual animals from images can significantly aid conservation efforts and wildlife management.

8. Experiments

Since the Mega-Descriptor works by leveraging the different S-Win Transformer models to extract features hierarchically and stores a large dataset of these. Test or inference images first have their features extracted and then a KNN-Extractor is used to compare with the existing database to predict the class.

Our above inference, which aligns with the paper results shows that some classes like Macaque are nearly solved or have comparably higher accuracies than other models like Dino-V2 while the lion class for instance still suffers from a lot of inaccuracies.

Having only the final feature embeddings to work with, we propose that accurate classification will depend on the extent of similarity/dissimilarity images of a sub-class have with other entities within the class. (I.e. how different are the cows within the Friesian Cattle class for instance.)

To attempt to find a correlation we at random selected a subset of test and train sample embeddings in an 20:80 ratio and plotted the highest cosine similarities with in each class (intra-class similarity), in an attempt to discover any obvious correlation with the accuracies. The results of both these experiments are discussed below.

Another hypothesis we wanted to test was whether the dataset accuracy would be affected by the size of dataset alone. Essentially, how much of Macaque class's accuracy can be attributed to the fact that it has more entries in its database to match images against. We independently conducted a KNN matching (with k=1) for using the subsets of the train and test sets as the database and test set respectively.

9. Experimental Results

Intra-class similarity was much higher for Lion class than Macaque class. If we are to assume a correlation then our considerably well-performing test set of MPDD should also have intra-class similarities on the lower side. This is aligning with the results of the 3 classes.

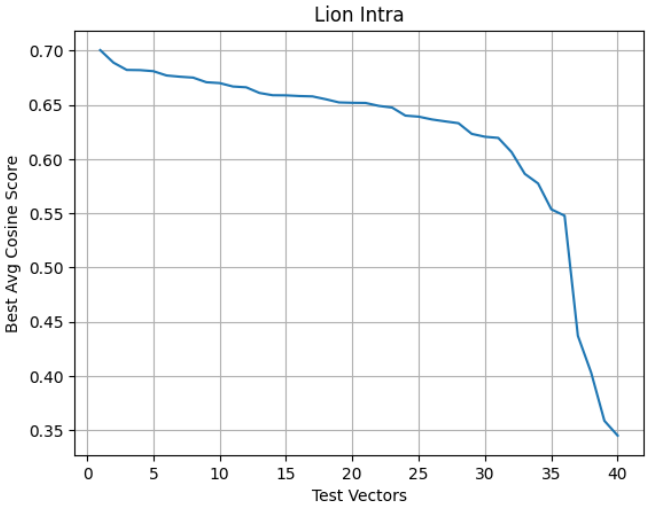


Figure 4. High Cosine Similarity Values

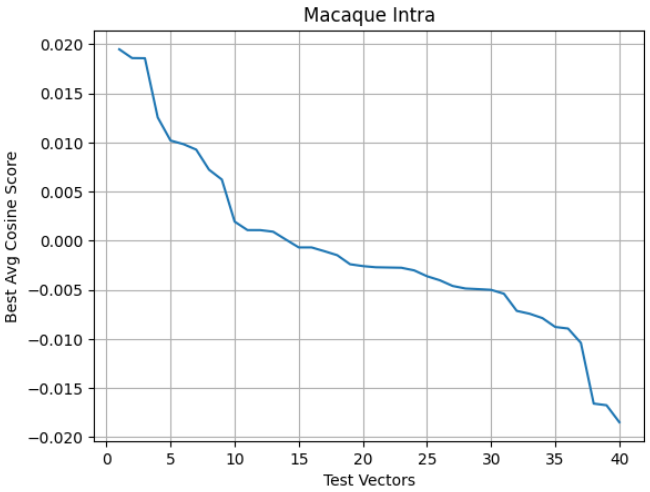


Figure 5. Low Cosine Similarity Values

In the case of testing whether the frequency difference was the cause of the increased accuracy, we cant report that this alone cannot be a factor for the high accuracy reported by the macaque class. As the accuracy on the Macaque Subset was also 1.0, whereas that of the Lion Subset was 0.1875.

10. Conclusion

In conclusion, our study demonstrates the effectiveness of MegaDescriptors with Swin Transformer models in wildlife re-identification. Future work could focus on further improving the models' performance across already existing dataset, making and training the models on re-identification of other animals and exploring their application in other

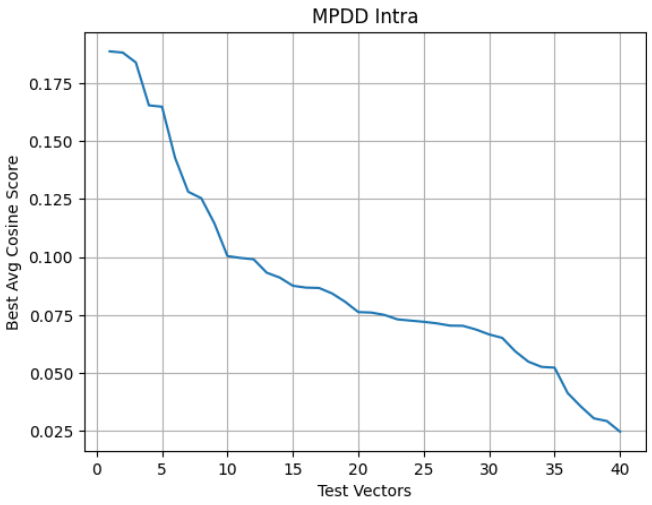


Figure 6. Relatively Low Cosine Similarity Values

wildlife conservation tasks.

11. Individual Commentaries

11.1. Aditya Jain

We have used two models T-224 and L-224 which are Swin Transformers for the purpose of Wildlife Re-identification. We used these models on three datasets(LionData, MacaqueFaces, FreisianCattle2015) from Wildlife Datasets which is a collection of 33 wildlife datasets and replicated the results given in the paper and also tested it on MPDD. We noticed that the results for the LionData were comparatively poor. From the Exploratory Data Analysis, we noted that the similarity in images for the lions is high, so we took the Null Hypothesis as if it was harder to distinguish amongst a class (say lions) it would be harder to achieve high accuracy for our experiment. We tried to find the cosine similarity between 160 samples of the training and 40 samples of the testing for LionData and Macaque, since Macaque was giving a good accuracy. We found that the cosine similarity values for LionData was very high, and Macaque was low. We also took MPDD dataset to validate the hypothesis and found the cosine similarity values to be low as well. We also noted that MacaqueFace had a larger number of images in the training dataset as compared to other datasets. We took the second hypothesis as if the size of the training data is the reason for the increase in accuracy. We had to reject this hypothesis because if we took sizes of them to be the same then also MacaqueFace gave a much better accuracy than LionData.

11.2. Abhishek Sushil

We chose 2 of the 4 Shifting Window Transformers used in the Megadescriptor paper, namely the T-224 and L-224. We used these models to extract feature vectors from 3 of the 33 datasets that are a part of the larger Wildlife Dataset. We also used MPDD as an external test set. The model essentially extracted features from the images and used a KNN Classifier to attempt to predict them base on the cosine similarity of the feature vectors in its training data.

We observed poor performance of Lion class, moderate performance of Fressian Cattle class and almost perfect accuracy of Macaque class. In order to explain these results we used 2 hypotheses, namely:- Whether differentiability between animals within a class would play a role in determining accuracy and whether the size of the train data i.e. the values one could reference against would favor higher accuracy.

We observed that our initial hypothesis seemed to align with our assumption although it would be ideal to perform this over many classes. Our second hypothesis, on the other hand, was not accurate as despite reducing the sizes of the Macaque and Lion class to the same number, the Macaque class still outperformed the Lion Class considerably.

11.3. Abhinav Ujjawal

From the research paper, we decide to use two of the megadescriptor models (T-224 and L-224) and reproduce their results on the widllife datasets. For this task, we chose three datasets that have three different species.

First is FresianCattle2015 (based on cattle), which had an average performance. The Second was LionData (based on lions), which had a very poor performance. The third dataset was MacaqueFaces (based on Macaques), which had a very high performance.

Apart from this, we picked up another dataset, MPDD (based on dogs), on which the model was not trained. We get a high performance on this dataset.

The difference between L-224 and T-224 becomes starkly visible in the case of the LionData dataset. The L-224 model has approximately 6% more accuracy than the T-224 model. After noticing the variation in these results, we decided to conduct some experiments and decide on the validity of two hypotheses.

In the first hypothesis, we see whether the similarity of animals within the dataset itself played a role in determining the accuracy of that dataset. To check this, we used the cosine similarity over a sample of data from the datasets LionData (low accuracy) and MacaqueFaces (high accuracy). We found out this hypothesis is valid as the cosine similarity value for lions was high, but it was very low for macaques.

In the second hypothesis, we see whether the size of our train data in the datasets impacts the model's performance. However, this hypothesis is not valid as even after taking the same size of data for both species, the Macaque dataset's performance was fairly better than the Lion dataset's.

References

[1] Jan Cermak et al. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2024. 1, 2

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1