Triple-stream Deep Metric Learning of Great Ape Behavioural Actions

Otto Brookes[1], Majid Mirmehdi[1], Hjalmar Kühl[2] and Tilo Burghardt[1]
[1]*Department of Computer Science, University of Bristol, U.K.*
[2]*Evolutionary and Anthropocene Ecology, iDiv, Leipzig, Germany*

Keywords: Animal Biometrics, Multi-Stream Deep Metric Learning, Animal Behaviour, Great Apes, PanAf-500 Dataset.

Abstract: We propose the first metric learning system for the recognition of great ape behavioural actions. Our proposed triple stream embedding architecture works on camera trap videos taken directly in the wild and demonstrates that the utilisation of an explicit DensePose-C chimpanzee body part segmentation stream effectively complements traditional RGB appearance and optical flow streams. We evaluate system variants with different feature fusion techniques and long-tail recognition approaches. Results and ablations show performance improvements of $\sim 12\%$ in top-1 accuracy over previous results achieved on the PanAf-500 dataset containing 180,000 manually annotated frames across nine behavioural actions. Furthermore, we provide a qualitative analysis of our findings and augment the metric learning system with long-tail recognition techniques showing that average per class accuracy – critical in the domain – can be improved by $\sim 23\%$ compared to the literature on that dataset. Finally, since our embedding spaces are constructed as metric, we provide first data-driven visualisations of the great ape behavioural action spaces revealing emerging geometry and topology. We hope that the work sparks further interest in this vital application area of computer vision for the benefit of endangered great apes. We provide all key source code and network weights alongside this publication.

1 INTRODUCTION

As the climate crisis gathers pace, the threat to many endangered species grows ever more perilous (Almond et al., 2022). All species of great apes are, for instance, listed as endangered or critically endangered according to the IUCN Red List (IUCN, 2022).

Consequently, there is urgent need for methods that can help to monitor population status and assess the effectiveness of conservation interventions (Kühl and Burghardt, 2013; Congdon et al., 2022; Tuia et al., 2022). This includes the recognition of behaviors and variation therein, as an integral part of biological diversity (Dominoni et al., 2020; Carvalho et al., 2022).

Previous works have employed deep neural networks which leverage multiple modalities, such as RGB, optical flow, and audio (Sakib and Burghardt, 2020; Bain et al., 2021), for the classification of great ape behaviours and actions. However, higher level abstractions such as *pose* or *body part* information have remained unexplored for addressing this task. In response, we propose utilising the latter *together* with RGB and optical flow in a triple-stream metric learning system (see Fig. 1) for improved classification results and domain visualisations relevant to biologists.

Great Ape Activities. This paper will focus on *great ape activity recognition*, where the coarse activity classes used are illustrated in Fig. 2 for the utilised PanAf-500 dataset (see Sec. 3). Note that computer vision would traditionally categorise these classes as actions whilst in the biological realm they represent behaviour (or aspects thereof) often captured in ethograms (Nishida et al., 1999; Zamma and Matsusaka, 2015). For clarity, in this paper we will refer to these classes as *behavioural actions* recognising historical traditions in both disciplines.

We will approach the classification task via a deep *metric* learning system (Karaderi et al., 2022) that embeds inputs into a latent space and uses geometric distances to form distributions that align with the semantic similarity captured by the classes (Hermans et al., 2017; Musgrave et al., 2020). A major advantage over standard supervised systems is that sample distances in visualisations of the latent space always relate to learned similarity and, thus, are more naturally interpretable by experts. We will also analyse

Brookes, O., Mirmehdi, M., Kühl, H. and Burghardt, T.
Triple-stream Deep Metric Learning of Great Ape Behavioural Actions.
DOI: 10.5220/0011798400003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 294-302
ISBN: 978-989-758-634-7; ISSN: 2184-4321

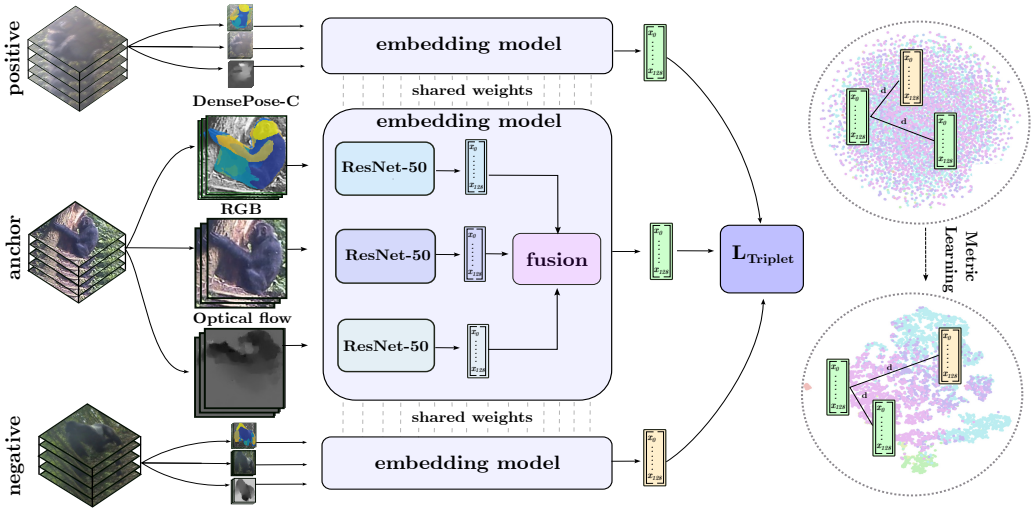

Figure 1: **System Overview**. Our proposed triple-stream metric learning approach utilises all RGB appearance, optical flow, and DensePose-C segmentations of chimps in videos. Exploiting hybrid reciprocal triplet and cross entropy losses, the model is then trained to map embeddings representing great ape behavioural actions onto a metric space, where semantically similar representations are geometrically close forming natural clusters. This pipeline improves on state-of-the-art classification performance and allows for visualisations of the underpinning space of behavioural actions (best viewed zoomed).

the role that additional DensePose-Chimp information (Sanakoyeu et al., 2020) can play in improving recognition performance compared to systems that utilise RGB and optical flow only. Lastly, as shown by Sakib and Burghardt (Sakib and Burghardt, 2020), there are significant challenges in correctly classifying behavioural actions which occur infrequently and form the distribution tail (see Fig. 2). To address this, we will employ three long-tailed recognition (LTR) techniques to improve performance on tail classes; (i) logit adjustment (Menon et al., 2020); (ii) class balanced focal loss (Cui et al., 2019); and (iii) weight balancing (Alshammari et al., 2022).

In summary, our contributions are as follows: (i) we implement the first deep *metric* learning system for recognising great ape behavioural actions; (ii) we show that utilising explicit pose information has a significant positive effect on recognition performance in this domain; and (iii) we establish that existing LTR techniques can be applied in a metric learning setting to improve performance on tail classes for the problem. The proposed approaches improve the state-of-the-art performance benchmarks with respect to top-1 ($\sim$ 85%) and average per class ($\sim$ 65%) accuracy on the PanAf-500 dataset.

2 RELATED WORK

Action recognition aims to classify actions observed in video (Kalfaoglu et al., 2020; Shaikh and Chai, 2021). Learning spatio-temporal features characteristic for actions (Simonyan and Zisserman, 2014) via various deep learning paradigms forms the approach of choice in the domain of human action recognition (HAR). We will briefly review concepts from this field, before discussing specifc relevant great ape behavioural action recognition and LTR methods.

Human Action Recognition. Although there are numerous deep learning approaches to action recognition (Zhou et al., 2018; Lin et al., 2019; Tran et al., 2019; Kalfaoglu et al., 2020; Pan et al., 2019; Majd and Safabakhsh, 2020; Sharir et al., 2021; Zhang et al., 2021a) this work focuses on multi-stream architectures, which address key aspects of the action recognition problem (e.g., spatial and temporal) independently and explicitly. Feichtenhofer et al. (Feichtenhofer et al., 2019) introduced the SlowFast architecture which employs two streams, each operating at different frame rates; a slow, low frame-rate pathway captures spatial information while the fast, high frame-rate pathway captures fine temporal detail. Other types of multi-stream networks process different visual modalities. Simonyan (Simonyan and Zisserman, 2014) introduced a two-stream network that processes RGB and optical flow to exploit spatial and temporal semantics, respectively. Since then, several networks that utilise additional modalities, such as motion saliency (Zong et al., 2021) and audio (Wang et al., 2021), have been introduced. Recently, the introduction of pose, which is critical for the perception of actions (Le et al., 2022), has shown promising results in multi-stream architectures (Hong et al., 2019;

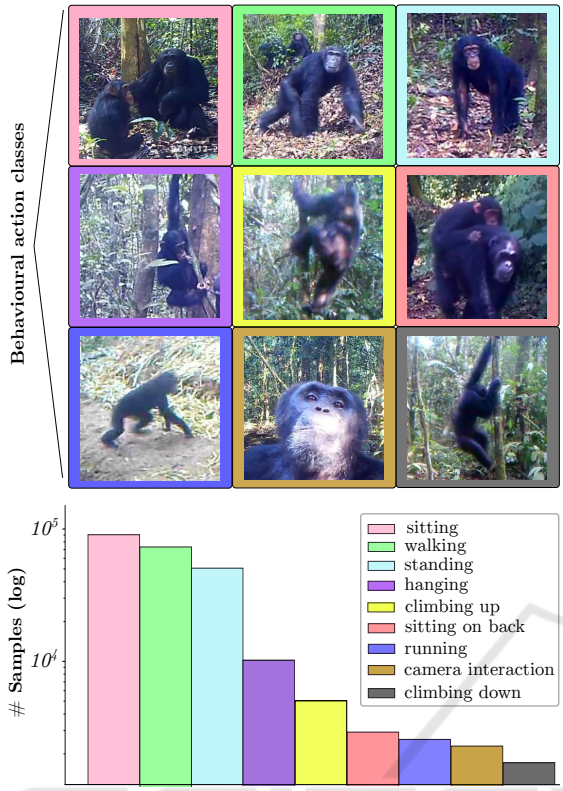

Figure 2: **Behavioural Actions in the PanAf-500 Data**. Examples of each one of the nine behavioural action classes (*top*) and their distribution across the approx. 180k frames in the dataset (*bottom*). Note the imbalance of two orders of magnitude in the distribution (best viewed zoomed).

Hayakawa and Dariush, 2020; Duan et al., 2021; Li et al., 2022). In particular, the DensePose format provides an opportunity to exploit fine-grained, segmentation map-based pose representations for action recognition. Hayakawa et al. (Hayakawa and Dariush, 2020) combine RGB and DensePose estimations in a two-stream network and demonstrate strong performance on egocentric footage of humans. Whilst such significant progress has been made in the domain of HAR, research into great ape behavioural action recognition is still in its infancy and few systems have been tested on natural datasets.

Great Ape Domain. To date, two systems have attempted automated great ape behavioural action recognition, both are multi-stream architectures. The first (Sakib and Burghardt, 2020) is based on the two-stream convolutional architecture by Simonyan et al. (Simonyan and Zisserman, 2014) and used 3D ResNet-18s for feature extraction and LSTM-based fusion of RGB and optical flow features. They report top-1 accuracy of 73.52% across the nine behavioural actions in the PanAf-500 dataset (see Sec. 3) and a relatively low average per class accuracy (42.33%), highlighting the issue of tail class performance. The second, proposed by Bain et al. (Bain et al., 2021), is a deep learning system that requires both audio and video inputs and detects two specific behaviours; buttress drumming and nut cracking. Their system utilised a 3D ResNet-18 and a 2D ResNet-18 for extraction of visual and assisting audio features, respectively, in different streams. They achieved an average precision of 87% for buttress drumming and 85% for nut cracking on their unpublished dataset. However, the multi-modal method is not applicable to all camera trap settings since many older models do not provide audio. It cannot be utilised on the PanAf-500 dataset since many clips there do not contain audio.

Long-Tailed Recognition. Most natural recorded data exhibits long-tailed class distributions (Liu et al., 2019). This is true of great ape camera-trap footage which is dominated by commonly occurring behaviours - even with only the nine classes of the PanAf-500 data the distribution shows a clear tail (see Fig. 2). Without addressing this issue, models trained on such data often exhibit poor performance on rare classes. Various counter-measures have been proposed (Verma et al., 2018; Kang et al., 2019; Zhang et al., 2021b). Class balanced losses assign additional weights, typically determined by inverse class frequencies, to samples from rare classes and have yielded strong results when coupled with techniques to reduce per-class redundancy (Cui et al., 2019). Similarly, logit adjustment uses class frequencies to directly offset output logits in favour of minority classes during training (Menon et al., 2020). An orthogonal approach, based on the observation that weight norms for rare classes are smaller in naively trained classifiers, is to perform weight balancing (Alshammari et al., 2022). These techniques have achieved strong results on several LTR benchmarks.

Before detailing how we use triple-stream metric learning with explicit DensePose-Chimp processing and LTR extensions for behavioural action recognition, we will briefly outline the utilised dataset.

3 DATASET

The *Pan-African* dataset, gathered by the Pan African Programme: 'The Cultured Chimpanzee', comprises $\sim$ 20,000 videos from footage gathered at 39 study sites spanning 15 African countries. Here we utilise a 500 video subset, PanAf-500, specifically ground-truth labelled for use in computer vision under re-

Figure 3: **Frame-by-frame Ground Truth Annotations**. Four still frames from PanAf-500 videos with annotations of location (green boxes) and behavioural actions (visualised as text) of the apes in-frame (best viewed zoomed).

producible and comparable benchmarks. It includes frame-by-frame annotations for full-body locations of great apes and nine behavioural actions (Sakib and Burghardt, 2020) across approximately 180k frames (see. Fig. 3). Fig. 2 displays the behavioural actions classes in focus together with their distribution. We utilised the PanAf-500 dataset for all experiments and employ the same training and test partitions described in (Sakib and Burghardt, 2020).

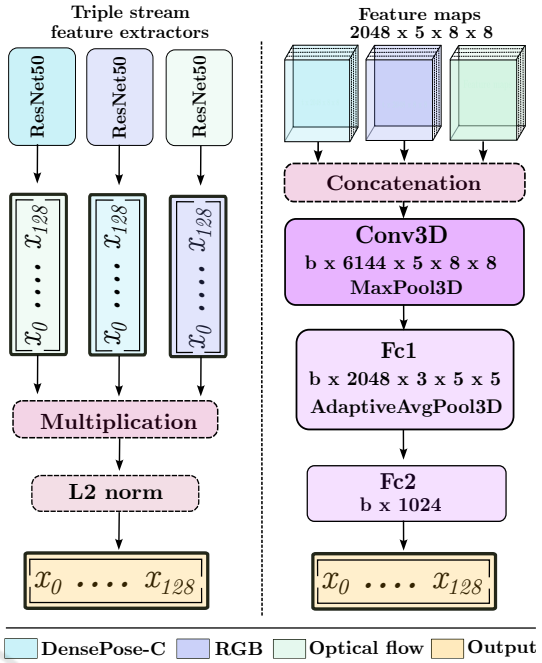

Figure 4: **Fusion Head Schematics**. A component breakdown of fusion by element-wise multiplication (*left*) and convolutional fusion (*right*) as applied for our work to explore their impact on performance.

4 METHOD

The proposed system utilises three visual modalities as input; RGB, optical flow, and DensePose-C estimations (Sanakoyeu et al., 2020), as illustrated in Fig. 1). All optical flow images are pre-computed using OpenCV's implementation of the Dual TV L1 algorithm (Zach et al., 2007). We employ the model developed by Sanakoyeu et al. (Sanakoyeu et al., 2020) to generate DensePose-C segmentations describing chimpanzee pose. The model predicts dense correspondences between image pixels and a 3-D object mesh where each mesh represents a chimpanzee body part specified by a selector I and local surface coordinates within each mesh indexed by U and V. Frame-by-frame application to each of the PanAf-500 videos yields DensePose-C estimates expressed in IUV coordinates.

Each of the three input modalities is fed into a 3D ResNet-50 (Du Tran et al., 2017) backbone, which together act as a feature extractor (see Fig. 1). The input tensors into the backbones are 3D since inputs are processed in snippets, that is each stream accepts a sequence of n consecutive RGB frames, optical flow images, or IUV coordinates, respectively. The final fully-connected layer outputs an n-dimensional encoding for each stream. These are fused into a single embedding using three popular approaches; (i) simple averaging across streams; (ii) convolutional fusion whereby stream features are concatenated and passed to a 3D convolutional layer as a volume; and (iii) element-wise multiplication of all three embedding vectors followed by $L2$ normalisation. The latter two approaches are illustrated in detail in Fig. 4. A linear layer at the end of the fusion head finally outputs the unified embedding as logits. Whilst this system was trained via metric learning - visually sketched in Fig. 1 (right) - a k-NN classifier is used to perform inference in the embedding space during evaluation.

Let the parameters of this network $f_\theta(\cdot)$ be denoted by θ. Furthermore, let $f_\theta(x) = x$ be the shorthand for referring to embeddings. Our metric learning objective is, thus, to minimise the distance between anchor-positive embedding pairs $d(x_a, x_p)$ and maximise distance between anchor-negative embedding pairs $d(x_a, x_n)$, where d represents a Euclidean. Instead of using standard triplet loss (Hermans et al., 2017) L_{TL}, we use an improved version (Andrew et al., 2021), where the model is optimised via a hybrid reciprocal triplet and softmax cross-entropy loss:

$$L_{RC} = L_{CE} + \lambda\, L_{RT}. \quad (1)$$

It is assembled from two components balanced by $\lambda = 0.1$ as given in (Andrew et al., 2021). The two

components themselves are evaluated as:

$$L_{RT} = d(x_a, x_p) + \frac{1}{d(x_a, x_n)} \tag{2}$$

$$L_{CE} = -\log\left(\frac{e^{x_y}}{\sum_{i=1}^{C} e^{x_i}}\right), \tag{3}$$

where C denotes the total number of classes and y are the class labels.

In order to extend this system into the LTR domain we substitute the softmax cross-entropy term for losses calculated using; (i) cross-entropy softmax with logit adjustment (Menon et al., 2020) L_{LA}; (ii) class-balanced focal loss (Cui et al., 2019) L_{CB}; and (iii) class-balanced focal loss with weight balancing (Alshammari et al., 2022). The first two losses are evaluated as follows:

$$L_{LA} = -\log\left(\frac{e^{x_y} + \tau \cdot \log \pi_y}{\sum_{i=1}^{C} e^{x_i + \tau \cdot \log \pi_i}}\right), \tag{4}$$

$$L_{CB} = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} (1-p_i)^{\gamma} \log(p_i), \tag{5}$$

where π represents the class priors (i.e., class frequencies in the training set) and temperature factor $\tau = 1$, $\beta = 0.99$ is the re-weighting hyper-parameter, n is the total number of samples, y are the classes, $\gamma = 1$ is the focal loss hyper-parameter and $p_i = \sigma(x_i)$. Balancing the network weights θ is performed via a MaxNorm constraint $\|\theta_{l,i}\|_2^2 \leq \delta^2, \forall i$ given in (Alshammari et al., 2022) imposed on each class filter i in the last layer l of the network where δ is the L2-norm ball radius. We will reference a L_{CB}-based optimisation where weight balancing is performed via L_{WB}.

Methodologically, this described architecture approaches the learning of behavioural great ape actions via five key capabilities: 1) utilisation of multiple relevant input modalities across an entire video snippet; 2) effective streamed content encoding; 3) fusion into a single embedding space; 4) metric space optimisation so that distances naturally reflect semantic similarity; and 5) taking into account class imbalances common to the domain content.

5 EXPERIMENTS

5.1 General Training Setup

We train our architecture via SGD optimisation using batch size 32 and learning rate 10^{-4}. Feature extractor backbones are initialised with Kinetics-400 (Kay et al., 2017) pre-trained weights and training runs are distributed over 8 Tesla V100 GPUs for 100 epochs.

Table 1: **Behavioural Action Recognition Benchmarks.** Top-1 and average per-class (C-Avg) accuracy performance on the PanAf-500 dataset for the current state-of-the-art (row 1), single and dual-stream baselines (rows 2–5), and our triple-stream networks (rows 6–8) for different fusion methodologies and losses tested.

	Models/Streams	Fusion	Loss	Top-1	C-Avg
Sakib et al. 2020					
1	*RGB+OF*	LSTM	L_{FL}	**73.52%**	**42.33%**
Up to Dual-Stream					
2	*RGB only*	None	L_{TL}	55.50%	32.67%
3	*RGB only*	None	L_{RC}	74.24%	55.76%
4	*RGB+OF*	Avg	L_{TL}	62.90%	39.10%
5	*RGB+OF*	Avg	L_{RC}	**75.02%**	**61.97%**
Triple-Stream (Ours)					
6	*RGB+OF+DP*	Avg	L_{RC}	81.71%	46.61%
7	*RGB+OF+DP*	Conv	L_{RC}	82.04%	**56.31%**
8	*RGB+OF+DP*	Elem	L_{RC}	**85.86%**	50.50%

5.2 Baselines and Stream Ablations

As shown in Tab. 1, we first establish performance benchmarks for one and two stream baseline architectures of our system (rows 2–5) against the current state-of-the-art (row 1), which uses a ResNet-18 backbone with focal loss L_{FL}, SGD, and LSTM-based frame fusion (Sakib and Burghardt, 2020). As expected, we confirmed that - using identical setups and losses - adding an optical flow stream is beneficial in the great ape domain mirroring HAR results (see rows 2 vs 4, and 3 vs 5). Additionally, models trained using L_{RC} consistently outperformed standard triplet loss L_{RC} scenarios (see rows 2 vs 3, and 4 vs 5). Finally, a dual-stream version of our proposed architecture trained with L_{RC} outperforms the state-of-the-art by a small margin (see rows 1 vs 5).

5.3 Triple-Stream Recognition

As given in Tab. 1 rows 6–8, our proposed triple-stream architecture significantly outperforms all baselines with regards to top-1 accuracy, achieving up to 85.86%. Thus, explicit DensePose-C information appears a useful information source for boosting behavioural action recognition in great apes. However, without LTR techniques all our triple-stream models are significantly outperformed by a dual-stream setting (row 5) with regards to average per-class accuracy. This reduction is caused by significantly poorer performance on minority classes (see Sec. 5.4).

Since the learned behavioural action embeddings are constructed as metric from the outset, they can be visualised meaningfully – we note that such data-driven visualisations are novel in the primatology do-

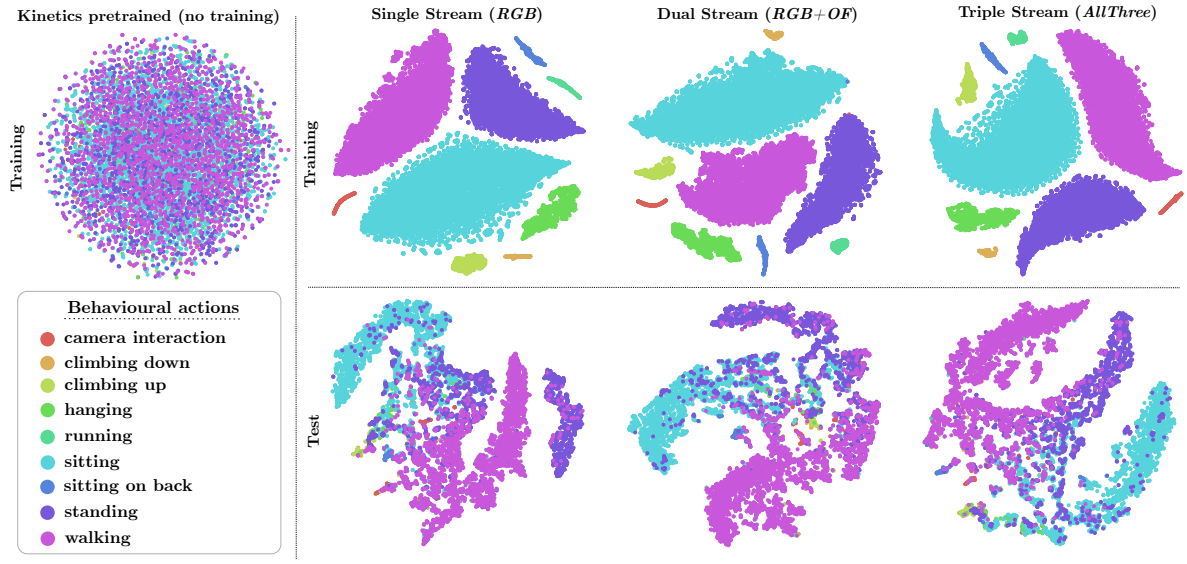

Figure 5: **Visualisations of Great Ape Behavioural Action Spaces**. A 2D t-SNE (Wattenberg et al., 2016) visualisation of the 128-dimensional training (top-right) and test (bottom-right) embeddings produced by the single, dual and three-stream network with convolutional fusion. We can see that training set embeddings from all classes are clustered cleanly. In contrast, test set embeddings show significant overlap and only embeddings from majority classes form distinct clusters. This is consistent with the high top-1 accuracy and relatively low average per-class accuracy reported in Tab. 1.

main. Fig. 5 depicts such learned spaces for our data and architecture where, independent of stream cardinality, embeddings cluster the training data cleanly. This is of course expected given above 99% top-1 *training* accuracy in all settings. Yet, behavioural actions of great apes are highly intricate as well as variable and, even with approx. 144,000 training frames used, the model clearly shows signs of overfitting. As a result, test set embeddings exhibit significant cluster overlap. Sample groups representing sitting, standing, and walking, for instance, blend into one another. In addition to overfitting, this also highlights the transitional nature of these often temporarily adjacent and smoothly changing actions. Thus, future temporally transitional ground truth labelling may be needed to represent behavioural great ape action in the PanAf-500 dataset more authentically.

5.4 Fusing Streams

When looking at the impact of information fusion methods on performance in more detail, we find that benchmarks vary significantly (see Tab. 1 rows 6–8) when we test averaging, element-wise multiplication, and convolutional fusion, as described in Sec. 4. Results show that convolution and element-wise multiplication improve performance slightly across both metrics when compared with averaging: top-1 accuracy improves by 0.33% and 4.1%, respectively (see rows 6–8). However, the most significant gains are observed with respect to average per class accuracy which increases by 3.44% for element-wise multiplication and 9.7% for convolutional fusion. Learnable parameters in the convolution method clearly help blending information even when only fewer samples are available for training. Building on this improvement, we will next investigate the impact of LTR methods in order to benefit tail class performance.

5.5 Long-Tail Recognition

When grouping behavioural actions into *head* (covering sitting, standing, and walking) and remaining *tail* classes based on frequency in the data (see Fig. 2), a significant performance gap becomes apparent even when using the so far best C-Avg performing model (see Tab. 2 row 1). Employing LTR techniques can, however, reduce this gap and improve average per-class accuracy further as quantified across rows 2–4 in Tab. 2). Fig. 6 shows t-SNE visualisations of the three LTR triple-stream approaches when trained with convolutional feature fusion. Particularly for the class-balanced approaches and weight-balancing setups (two rightmost), *tail* class clusters appear more clearly separated and class overlap is generally reduced. Thus, for the great ape domain underrepresented classes are indeed an effective source of information for improving action separability in general.

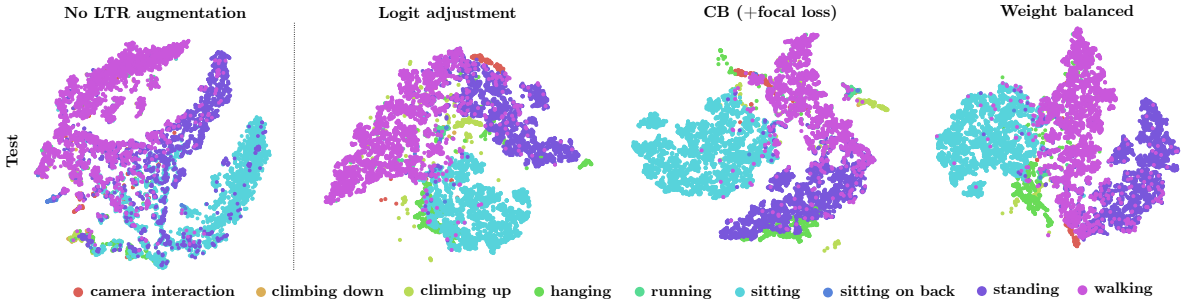

Figure 6: **Long-tail Test Embeddings**. A 2D t-SNE visualisation of the 128-dimensional test embeddings produced by the three-stream network with convolutional fusion alone (leftmost) and augmented with each LTR technique; (i) logit adjustment (ii) CB (+focal loss) and (iii) weight balancing. All LTR-augmented methods improve clustering of embeddings belonging to tail classes. They appear more clearly separated and exhibit less overlap when compared with the non-LTR method.

6 CONCLUSION

In this work we introduced the first deep metric learning system for great ape behavioural action recognition. We demonstrated that the proposed triple-stream architecture can provide leading state-of-the-art performance when tested on the PanAf-500 camera trap dataset covering 180,000 annotated frames across 500 videos taken in the wild. We demonstrated that the addition of a DensePose-C chimpanzee pose estimation stream into the embedding architecture is highly effective and leads to system performance of 85.86% top-1 accuracy on the data. We also showed that adding LTR techniques that address poor tail class performance to the system can improve the average per-class accuracy to 65.66% on the dataset. Despite these improvements we note that both larger annotated datasets to counteract overfitting as well as more temporally blended forms of annotation (e.g. action transition annotations) would benefit the authenticity of data-driven great ape behavioural representations. We hope that the research presented here sparks further interest in this vital application area for the benefit of endangered species such as great apes.

ACKNOWLEDGEMENTS

We thank the Pan African Programme: 'The Cultured Chimpanzee' team and its collaborators for allowing the use of their data for this paper. We thank Amelie Pettrich, Antonio Buzharevski, Eva Martinez Garcia, Ivana Kirchmair, Sebastian Schütte, Linda Gerlach and Fabina Haas. We also thank management and support staff across all sites; specifically Yasmin Moebius, Geoffrey Muhanguzi, Martha Robbins, Henk Eshuis, Sergio Marrocoli and John Hart. Thanks to the team at https://www.chimpandsee.org particularly Briana Harder, Anja Landsmann, Laura K. Lynn, Zuzana Macháčková, Heidi Pfund, Kristeena Sigler and Jane Widness. The work that allowed for the collection of the dataset was funded by the Max Planck Society, Max Planck Society Innovation Fund, and Heinz L. Krekeler. In this respect we would like to thank: Ministre des Eaux et Forêts, Ministère de l'Enseignement supérieur et de la Recherche scientifique in Côte d'Ivoire; Institut Congolais pour la Conservation de la Nature, Ministère de la Recherche Scientifique in Democratic Republic of Congo; Forestry Development Authority in Liberia; Direction Des Eaux Et Forêts, Chasses Et Conservation Des Sols in Senegal; Makerere University Biological Field Station, Uganda National Council for Science and Technology, Uganda Wildlife Authority, National Forestry Authority in Uganda; National Institute for Forestry Development and Protected Area Management, Ministry of Agriculture and Forests, Ministry of Fisheries and Environment in Equatorial Guinea. This work was supported by the UKRI CDT in Interactive AI under grant EP/S022937/1.

Table 2: **LTR-enabled Behavioural Action Recognition Benchmarks.** Average per-class accuracy for our triple-stream network with convolutional fusion for best performing non-LTR method (row1), and three LTR approaches (rows 2–4) targetting poor tail class performance.

	Method/Loss	C-Avg	Head	Tail
Non-LTR Triple-Stream				
1	L_{RC}	56.31	80.57	44.78
LTR Triple-Stream				
2	L_{LA}	61.76	**83.22**	50.7
3	L_{CB}	63.56	77.60	55.95
4	L_{WB}	**65.66**	82.55	**56.26**

REFERENCES

Almond, R., Grooten, M., Juffe Bignoli, D., and Petersen, T. (2022). Wwf (2022) living planet report 2022 - building a nature-positive society. 1

Alshammari, S., Wang, Y.-X., Ramanan, D., and Kong, S. (2022). Long-tailed recognition via weight balancing. In *CVPR*, pages 6897–6907. 2, 3, 5

Andrew, W., Gao, J., Mullan, S., Campbell, N., Dowsey, A. W., and Burghardt, T. (2021). Visual identification of individual holstein-friesian cattle via deep metric learning. *Computers and Electronics in Agriculture*, 185:106133. 4

Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., Hockings, K. J., Matsuzawa, T., Hayashi, M., Biro, D., et al. (2021). Automated audiovisual behavior recognition in wild primates. *Science advances*, 7(46):eabi4883. 1, 3

Carvalho, S., Wessling, E. G., Abwe, E. E., Almeida-Warren, K., Arandjelovic, M., Boesch, C., Danquah, E., Diallo, M. S., Hobaiter, C., Hockings, K., et al. (2022). Using nonhuman culture in conservation requires careful and concerted action. *Conservation Letters*, 15(2):e12860. 1

Congdon, J., Hosseini, M., Gading, E., Masousi, M., Franke, M., and MacDonald, S. (2022). The future of artificial intelligence in monitoring animal identification, health, and behaviour. 1

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277. 2, 3, 5

Dominoni, D. M., Halfwerk, W., Baird, E., Buxton, R. T., Fernández-Juricic, E., Fristrup, K. M., McKenna, M. F., Mennitt, D. J., Perkin, E. K., Seymoure, B. M., et al. (2020). Why conservation biology can benefit from sensory ecology. *Nature Ecology & Evolution*, 4(4):502–511. 1

Du Tran, H. W., Torresani, L., Ray, J., Lecun, Y., and Paluri, M. (2017). A closer look at spatiotemporal convolutions for action recognition.(2017). *OK*. 4

Duan, M., Qiu, H., Zhang, Z., and Wu, Y. (2021). Ntu-densepose: A new benchmark for dense pose action recognition. In *Big Data*, pages 3170–3175. IEEE. 3

Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In *ICCV*, pages 6202–6211. 2

Hayakawa, J. and Dariush, B. (2020). Recognition and 3d localization of pedestrian actions from monocular video. In *ITSC*, pages 1–7. IEEE. 3

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*. 1, 4

Hong, J., Cho, B., Hong, Y. W., and Byun, H. (2019). Contextual action cues from camera sensor for multi-stream action recognition. *Sensors*, 19(6):1382. 3

IUCN (2022). Iucn red list of threatened species version 2022.1. 1

Kalfaoglu, M. E., Kalkan, S., and Alatan, A. A. (2020). Late temporal modeling in 3d cnn architectures with bert for action recognition. In *ECCV*, pages 731–747. Springer. 2

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. (2019). Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*. 3

Karaderi, T., Burghardt, T., Hsiang, A. Y., Ramaer, J., and Schmidt, D. N. (2022). Visual microfossil identification via deep metric learning. In *ICPRAI*, pages 34–46. Springer. 1

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. 5

Kühl, H. S. and Burghardt, T. (2013). Animal biometrics: quantifying and detecting phenotypic appearance. *TREE*, 28(7):432–441. 1

Le, V.-T., Tran-Trung, K., and Hoang, V. T. (2022). A comprehensive review of recent deep learning techniques for human activity recognition. *Computational Intelligence and Neuroscience*, 2022. 2

Li, Y., Lu, Z., Xiong, X., and Huang, J. (2022). Perf-net: Pose empowered rgb-flow net. In *WACV*, pages 513–522. 3

Lin, J., Gan, C., and Han, S. (2019). Tsm: Temporal shift module for efficient video understanding. In *ICCV*, pages 7083–7093. 2

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. (2019). Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546. 3

Majd, M. and Safabakhsh, R. (2020). Correlational convolutional lstm for human action recognition. *Neurocomputing*, 396:224–229. 2

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. (2020). Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*. 2, 3, 5

Musgrave, K., Belongie, S., and Lim, S.-N. (2020). Pytorch metric learning. 1

Nishida, T., Kano, T., Goodall, J., McGrew, W. C., and Nakamura, M. (1999). Ethogram and ethnography of mahale chimpanzees. *Anthropological Science*, 107(2):141–188. 1

Pan, Y., Xu, J., Wang, M., Ye, J., Wang, F., Bai, K., and Xu, Z. (2019). Compressing recurrent neural networks with tensor ring for action recognition. In *AAAI*, volume 33, pages 4683–4690. 2

Sakib, F. and Burghardt, T. (2020). Visual recognition of great ape behaviours in the wild. *VAIB*. 1, 2, 3, 4, 5

Sanakoyeu, A., Khalidov, V., McCarthy, M. S., Vedaldi, A., and Neverova, N. (2020). Transferring dense pose to proximal animal classes. In *CVPR*, pages 5233–5242. 2, 4

Shaikh, M. B. and Chai, D. (2021). Rgb-d data-based action recognition: A review. *Sensors*, 21(12):4246. 2

Sharir, G., Noy, A., and Zelnik-Manor, L. (2021). An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*. 2

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 27. 2, 3

Tran, D., Wang, H., Torresani, L., and Feiszli, M. (2019). Video classification with channel-separated convolutional networks. In *ICCV*, pages 5552–5561. 2

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):1–15. 1

Verma, V., Lamb, A., Beckham, C., Najafi, A., Courville, A., Mitliagkas, I., and Bengio, Y. (2018). Manifold mixup: learning better representations by interpolating hidden states. 3

Wang, L., Yuan, X., Zong, M., Ma, Y., Ji, W., Liu, M., and Wang, R. (2021). Multi-cue based four-stream 3d resnets for video-based action recognition. *Information Sciences*, 575:654–665. 2

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill*. 6

Zach, C., Pock, T., and Bischof, H. (2007). A duality based approach for realtime tv-l 1 optical flow. In *DAGM*, pages 214–223. Springer. 4

Zamma, K. and Matsusaka, T. (2015). *Ethograms and the diversity of behaviors*, page 510–518. Cambridge University Press. 1

Zhang, Y., Li, X., Liu, C., Shuai, B., Zhu, Y., Brattoli, B., Chen, H., Marsic, I., and Tighe, J. (2021a). Vidtr: Video transformer without convolutions. In *ICCV*, pages 13577–13587. 2

Zhang, Y., Wei, X.-S., Zhou, B., and Wu, J. (2021b). Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, volume 35, pages 3447–3455. 3

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). Temporal relational reasoning in videos. In *ECCV*, pages 803–818. 2

Zong, M., Wang, R., Chen, X., Chen, Z., and Gong, Y. (2021). Motion saliency based multi-stream multiplier resnets for action recognition. *Image and Vision Computing*, 107:104108. 2