Lightweight Multi-Branch Network for Animal Individual Re-Identification

Bertan Karacora*

September 15, 2021

Abstract

The specific identification of animal individuals is fundamental for a broad range of biological research applications such as measuring biodiversity or studying animal behaviour. Using camera trap data and image-based deep learning systems, wild animals can be identified upon re-encounter avoiding any invasive methods, which often include capturing and sometimes injuring them. Thus, their behaviour is not effected and investigations can be automated with low logistical requirements. While recently, an increasing amount of Person Re-Identification approaches have been proposed, current advances in this field might also be applied to Animal Re-Identification. The Lightweight Multi-Branch Network, as introduced by Herzog et al. in 2021, focuses on resource-efficient, but precise re-identification and achieves state-of-the-art results on person datasets *Market1501* and *CUHK03*. In this work, the Lightweight Multi-Branch Network is adjusted for training on two animal datasets *BayWald* and *Wildpark*, that are built from wild animal video recordings. On these datasets, the model reaches 99.95% and 93.64% mAP, respectively.

Contents

*Matriculation number: 3226384, bertan.karacora@uni-bonn.de, Röckumstr. 46a, 53121 Bonn

1 Introduction

Animal Individual Re-Identification means recognizing a reoccurring animal identity from a gallery of images given only one example. These images can vary in recording conditions including camera position, perspective, time of recording, lighting, poses, background, movement, camera effects, resolution and of course the depicted animal individual. Unlike in traditional classification or identification tasks, there is no predetermined database of classes or identities and detection is also a premise. Instead, re-identification systems aim to find a similarity measure, so that for a single given query image, all pictures from the gallery that show the exact same identity as the query are ranked highest according to this measure (figure 1) [7].

The ability to re-identify animal individuals is of great value for biological and ecological research. For example, it is needed in observation of social behaviour in order to determine hierarchies, group memberships, or in general, which individuals are interacting with each other [16]. Especially if an animal leaves the camera field of vision and re-enters at a later time or in a different camera's view, Re-Identification is essential for consistent data acquisition. In their study, Schindler and Steinhage [14] give an indicative example how Animal Re-identification could serve as a supplement to animal tracking and action observation. Moreover, the measurement of biodiversity is a crucial prerequisite for nature conservation. Particularly if combined with large-scale-compatible camera traps, Animal Re-Identification enables researchers to estimate species distribution as well as population size and density of wild animal live stocks, which is important for ensuring environmental sustainability and ecological stability [15].

Naturally, many different methods for this task have been established, however, most of them do not involve image-based re-identification. Some currently practised processes are attaching physical tags to the animal body, visible implants, microchip implants, GPS tracking or even genetic engineering. All of these are bound to logistical effort and costly resources, but most of all, these methods conclude to an incursion of the ecosystem and have an unpredictable effect on the animal behaviour [16]. In some cases, small injuries serve as unique markers, for instance in scarring, ear-notching or toe-clipping procedures [12]. In consideration of these practices, the advantages of an image-based, non-invasive re-identification system seem apparent.

Although there have been many deep learning approaches to Person Re-Identification in the past and in the present [18, 20], only little attention is paid to Animal Individual Re-Identification. A noteworthy contribution has been made by Freytag et al. [5], who implemented the AlexNet neural network for face recognition of chimpanzees. Brust et al. [1] reproduced their technique on a gorilla dataset. Therefore, they deployed camera traps and cropped the resulting images to have only the faces remain. Lastly, Schneider et al. [15] give a further perspective on the capabilities of deep learning in this domain and how to set up a wildlife animal re-identification system in detail.

This work is based on the studies of Herzog et al. [7] in their publication *Lightweight Multi-Branch Network for Person Re-Identification* from 2021. Therein, they propose a lightweight network structure consisting of an OSNet-backbone [22] and three separated branches, each

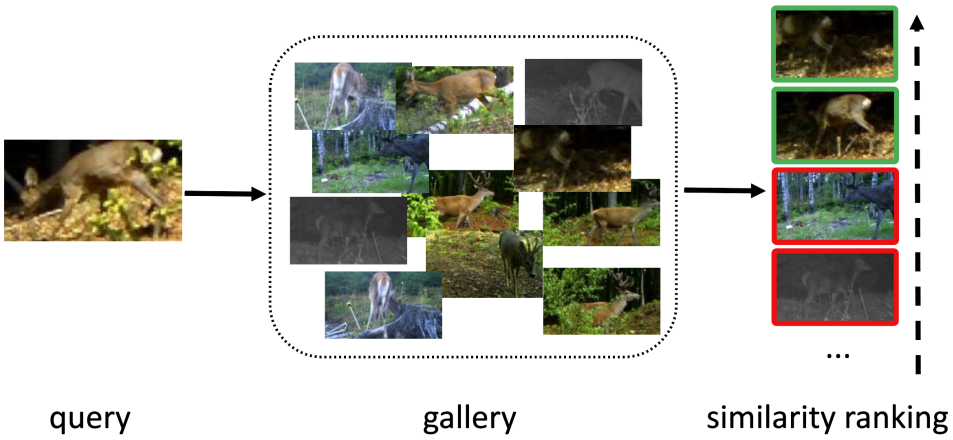

Figure 1: Gallery images are matched to the query and ranked according to a similarity measure. If top results show the same individual as the query, re-identification is successful.

one of which is responsible for global, channel-based or partial features. Making use of suitable training routines, their network achieves state-of-the-art performances on Market1501 and CUHK03 [9, 19]. In recent years, deep learning approaches for Person Re-Identification tend to become increasingly complex and resource-heavy. The Lightweight Multi-Branch Network outperforms other multi-branch architectures on those datasets without increasing the overall model complexity extensively. Herzog et al. [7] also shared their code and some pretrained models publicly. On this basis, an implementation for animal video data is made in this work. Therefore, programming language Python as well as libraries Pytorch [11] and Torchreid [21] are used to accomplish this.

The following elaborates how videos of wild animals can be used for training the Lightweight Multi-Branch Network. In order to do so, two datasets were taken into account: BayWald and Wildpark, which partially have been annotated as part of this work. Both show only a small amount of different individuals and because of the nature of videos data, the images have to run through some preprocessing prior to the training. The original method of Herzog et al. [7] is analyzed, reproduced and eventually, modifications are tested accordingly for use on wild animal video data. While adapting training parameters, the results are evaluated with assistance of visualization means such as activation maps and similarity rankings.

2 Data and annotation

As Schneider et al. [15] stated, a data library for Animal Re-Identification should optimally not be based on video frames because the training and testing data could suffer from background biases. Additionally, it should provide at least 500 individuals and all of them should have multiple sightings from different perspectives. Whereas Market1501 and CUHK03 had different camera positions, only a single camera was used each for BayWald and Wildpark. Both datasets show wild deers (*Cervus elaphus* and *Capreolus capreolus*) and are constructed from videos that contain only a few individuals and even less instances with multiple sightings. Instead of many individuals with a few different images of them, there are only a few identities with many images for everyone of them. So in this case, these datasets are initially not perfectly suited

for a Re-Identification task. However, if there is a sufficient time gap between the video frames and some animal movement, background bias can be reduced regarding only the bounding boxes. Furthermore, using segmentation masks of the animals the background could be cut out completely. BayWald has been provided with mask annotations in CSV format. Since either bounding boxes or segmentation masks are needed, in this work, for the annotation of Wildpark the tools *MiVOS* [2] and *VGG Image Annotator* [3] were used. This task is accomplished in cooperation of Iva Ewert for a parallel investigation [4] that is based on the work of Quispe et al. and the Top-DB-Net [13].

2.1 BayWald and Wildpark

In this study, videos from the datasets BayWald and Wildpark are used. The data material for the BayWald dataset is provided by the *Nationalpark Bayerischer Wald*, Germany. The videos of the Wildpark dataset were recorded in the *Wald- und Wildpark Rolandseck* by Frank Schindler with permission of the *Waldpark Rolandseck GmbH*, Germany.

BayWald consists of a total of 6292 images assembled from video recordings, which have been already annotated by Frank Schindler. Each video is recorded with either 8 or 15 frames per second, they show 20 different individuals and 3 instances with multiple sightings across the videos. The depicted animals are roe deers and red deers. The dataset includes pictures recorded at day or at night, they differ in lighting, camera effects, surroundings, distances to the camera. There are animals moving across the scene and resting ones.

Similarly, Wildpark contains a total of 3600 images, built from only 4 Videos each with 30 frames per second. Per video 5-12 individuals can be perceived, which results in 34 animals total, but only up to 21 different identities can be included. So some of these individuals reoccur in multiple videos, but unfortunately, they can not be further identified. Also, only red deers are included. Because of multiple instances on each video at the same time, some animals are covered by each other or by trees, so sometimes only a small part of the animal body can be seen. In general, most of the individuals are further away from the camera than in BayWald, which result in a lower resolution for these animals. A particular difficulty when handling this dataset is that if a bounding box is drawn around an animal, more than one individual can be seen due to the mentioned close vicinity between them. Given a bounding box picture with multiple individuals as query, the neural network could re-identify both of them in the gallery. Then again, it might re-identify a different animal in the gallery than the one, the bounding box belongs to. This way, Wildpark represents additional challenges for a re-identification system.

2.2 MiVOS

MiVOS (Modular Interactive Video Object Segmentation) [2] is an open-source interactive video annotation tool, which is composed of an interaction-to-mask module, a propagation module and a fusion module. First, the user constructs a segmentation mask of an object on a single video frame. The segmentation tool supports different interaction options, e.g. clicks or free-hand drawing. Then the propagation module generalizes across the video and adds the object mask to the remaining video frames. In an iterative process, manual annotation and propagation alternate, so corrections are made and updated throughout the entire video. The fusion module, that is used for further corrections, takes user-made changes into account and is difference-aware. If the result is satisfying, MiVOS generates images of the segmentation

Figure 2: Comparison of annotation results after transforming segmentation masks to contour polygons. Left: MiVOS. Right: VGG Image Annotator

masks using different colors for the annotated objects with a black background and also overlay images where semi-transparent masks are added on top of the original frames.

MiVOS has already been used effectively for annotation of the BayWald dataset. Hence, it appeared to be a promising framework for Wildpark annotation as well. Due to hardware requirements, longer videos need to be cut into small clips and put together afterwards. In this work, MiVOS was used for the annotation of only one of the four videos of Wildpark. In the other videos, it showed restrictions in number of objects and resolution. The images contain several animals at any time and many of the individuals in Wildpark have low resolution from the beginning. These limitations discouraged the use of MiVOS for video data. Although those limitations were removed during the course of this work, so that any specific resolution and an unlimited number of instances can be annotated, MiVOS images turned out to have significantly less quality than hand-made segmentation masks using the VGG Image Annotator (figure 2). Conclusively, MiVOS is less suitable than manual segmentation for video data segmentation including multiple objects and sub-optimal conditions such as coverage and low resolution.

2.3 VGG Image Annotator

The VGG Image Annotator [3] is a versatile open-source image annotation framework. It works in a web browser without any further installation or setup requirements. In contrast to MiVOS, annotation is done manually by either predetermined shapes like rectangles, circles, ellipses or by hand-drawn closed polygonal chains. For video data, a copy function for these polygons could be used if the animal does not noticeably move between the video frames. This is how 23613 polygons could be produced with equal contribution of Iva Ewert for the 3600 images of Wildpark with 5 to 12 instances on every frame. The VGG Image Annotator makes it possible to import or export a CSV file in the format that was used for BayWald annotation as well. Whereas MiVOS is bound to powerful hardware and resulted in mostly less profitable segmentation masks, the hand-made polygons do not suffer from pixelated curves or missing animal limbs even though annotation by hand is more time-intensive. Figure 2 shows a direct comparison between polygons created with MiVOS after transforming the mask images and with the VGG Image Annotator.

3 Methodology

In the following section, the methods of this work are explained in more detail. Essentially, these can be broken down into three parts: First, the Lightweight Multi-Branch Network and its architecture. Second, the optimization of the training process and lastly, the video data processing allowing a viable input for training and testing the image-based re-identification system.

3.1 Network architecture

Different from most previous multi-branch approaches towards Person Re-Identification, the Lightweight Multi-Branch Network focuses on a well-justified balance between high performance using multiple branches and at the same time low structural complexity. It is assembled of three branches on top of an OSNet [22] backbone. Figure 3 illustrates the architecture of the Lightweight Multi-Branch Network.

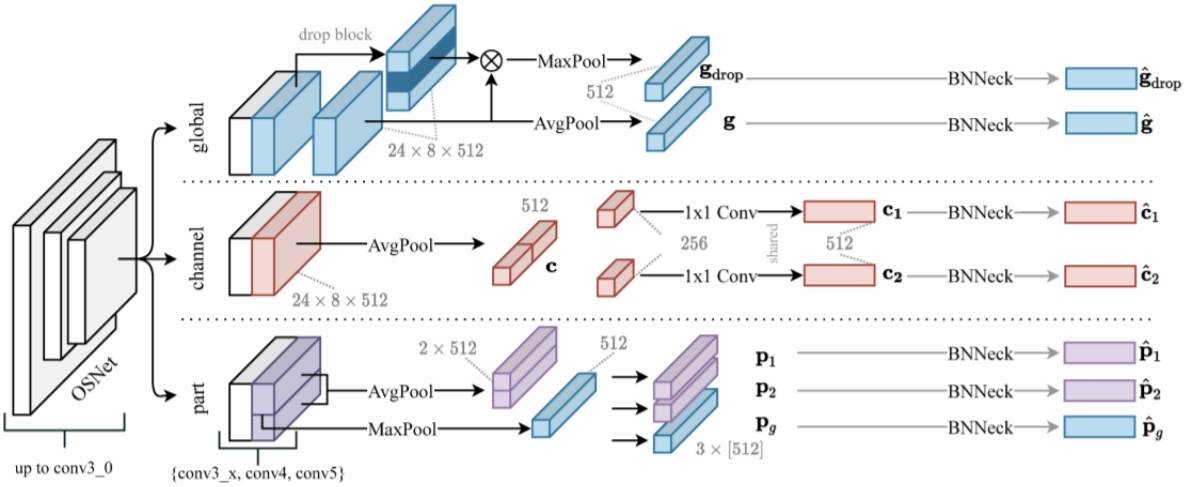

Figure 3: Network architecture. The first three stages of the OSNet backbone [22] extract omni-scaled features. Subsequently, the network is split into three separated branches: the global, channel and part branch. After additional pooling and convolution steps, BNNeck blocks [10] are applied to resulting feature vectors for optimization regarding different loss functions and to retrieve final output. Figure by Herzog et al. [7].

ResNet50 [6] has been used in similar architectures before, but the OSNet was picked in favor of a ResNet backbone due to stronger results while also keeping a smaller model size. The OSNet introduces omni-scaled features. Basically, these are able to combine smaller local discriminative details with the surrounding context of an image across different scales. Regional attributes and large-scale features are extracted equally and dynamically weighted to either focus on one of those or any mixture. Regarding Animal Re-Identification this method allows to recognize individuals based on e.g. overall head proportions in combination with ear shape. Framed into the network, a OSNet backbone pretrained on ImageNet [8] is applied.

The network branches differ on the following aspects: the global branch receives all extracted features. One global sub-branch obtains global features through average pooling. The other

half of the global branch applies the drop block technique [4, 13] prior to maximum pooling in order to increase robustness. Therefore, the features of the most discriminative areas are dropped, so that the system is forced to focus on more solid information.

In the channel branch, behind an average pooling step, the received set of $1 \times 1 \times 512$ input channels is split into halves as it can be seen in figure 3. This figure refers to Market1501 specifically, since tensor sizes $24 \times 8 \times 512$ assume image size 384×128 as input and the preassigned number of 512 feature maps. After that, 1×1 convolution is used to retrieve original scale.

In a similar manner, the part branch cuts the feature maps in its height. Here, the average pooling produces $2 \times 1 \times 512$ values, so that p_1 and p_2 reflect the upper and lower image half, respectively. Furthermore, another global vector is obtained in this branch using maximum pooling this time. At the end, BNNeck [10] blocks turn all resulting vectors into employable normalized representations. Unlike a standard neck structures of neural networks, BNNeck optimizes output for ranking loss as well as identity loss, since these require different metric space embeddings.

3.2 Training optimization

Herzog et al. [7] propose a combination of standard Cross-Entropy and Multi-Similarity Loss [17] for training. Other training optimization techniques used by Herzog et al. include random erasing augmentation, label smoothing and warmup cosine annealing for learning rate. For random erasing simply a random small rectangle of the image is cut out and replaced by the mean value. The label smoothing trick inserts a uniform noise to network predictions during training to avoid overfitting and overconfidence. As learning rate schedule warmup cosine annealing works as follows. First, learning rate increases linearly, then it is reduced with cosine decay. For this work, all these techniques have been adopted, but a reduced learning rate was used with regards to the small number of identities to learn in BayWald and Wildpark.

3.3 Video data processing

As a consequence of its construction, both animal datasets differ heavily from the originally used person datasets. This is why input video data needs to be preprocessed to allow an image-based training with the network. The preprocessing pipeline includes: (1) Load annotation files from CSV format and video frames. (2) Crop images to either bounding boxes or masks using annotation polygons. (3) Split data into training set, query-set and gallery. (4) Relabel identities for training. (5) Forward data on image-level. The implementation is based on Pytorch [11] and Torchreid [21, 22]. It is designed for easy use in a python notebook. Code was provided by Herzog et al. [7] and modified in this work.

The videos of BayWald and Wildpark show a fixed scene. So to obtain input for the network in a comparable manner the images need to be cropped to the animals bounding boxes or masked to remove background entirely. This is a necessary since detection is part of the re-identification task. Figure 4 demonstrates the resulting images. Due to different distances and poses the bounding box of the polygons obtained from annotation have not all the same dimensions, so often times some edge strips need to be preserved when cropping to keep the original aspect ratio.

BayWald was split into a training set, the query set and the gallery as follows: 15 of the 25 videos containing 4 night recordings and 13 of 20 different individuals are assigned for

Figure 4: Cropping procedure. The animal body forms only a small part of the scene.

training. Accordingly, 10 videos containing 2 night recordings and 10 instances belong to the query and gallery subsets. In more detail, initially one frame per second is used as query and the remaining frames as gallery. Wildpark contains 3600 images with 5 to 12 instances each resulting in a much higher number of 23617 images for this dataset after cropping. 2 of 4 videos are included in training and test sets, respectively. Query subset and the gallery are obtained the same way as for BayWald.

4 Experimental results

The Lightweight Multi-Branch Network and aforementioned training and data processing techniques are evaluated in the following section. In all experiments, training is performed for 30 epochs on BayWald and 20 epochs on Wildpark. In both cases, batchsize was set to 30, built from 10 images each for 3 identities. Input pixel resolution was set to 320×180 for BayWald and 240×180 for Wildpark. All evaluation metrics are utilized in the same manner as in the original study of Herzog et al. [7]. The rank-k accuracy metrics describe if the query identity is found in the top k matched gallery images according to the trained similarity measure. The mean average precision (mAP) is the mean value over all query images for the average precision (AP) of the top n predictions. So if the highest ranked gallery image is a correct prediction, precision for $m = 1$ is 1. If $m = 2$, the top 2 predictions are considered. The AP is calculated

Method	BayWald	Wildpark
baseline	0.9873	0.9516
FD	0.9949	0.9467
SQG	0.9833	0.9350
FD + SQG	0.9938	0.9364
FD + SQG + LS + WCA	0.9885	0.9320
FD + SQG + MA	0.9994	0.9225
FD + SQG + LS + WCA + MA	0.9995	0.9225
TRA + FD + SQG	-	0.9241

Table 1: mAP per method. FD: frame dropping. SQG: separation of query and gallery subsets. LS: label smoothing trick. WCA: warmup cosine annealing. MA: use of masks instead of bounding boxes. TRA: training on BayWald, evaluation on Wildpark

as the average of all m precisions up to a given limit n.

Tests on Wildpark induce an additional challenge: because it is not known which individuals are re-occurring among the videos, evaluation can either be done within-video, so that gallery contains always only frames of the same video for a given query or alternatively, one could label all instances as different individuals even though they might have the same identity. Experiments showed that this has no significant effect on evaluation, so this is not further regarded in the following. For training on Wildpark the latter method was used. For evaluation on Market1501 and CUHK03 images of the same identity from the same camera are discarded. This is not applicable for BayWald and Wildpark because only a single camera was used.

Table 1 shows an overview of utilized methods and performance. A combination of all techniques increases performance on BayWald to the maximum whereas masks are no improvement for Wildpark.

4.1 General performance on animal video data

Figures 6 and 5 exemplary illustrate mAP progression and loss during training. Generally, training progress converges after few epochs. It is striking that the model achieves high performance before training on animal datasets has started. This is caused by a combination of the already pretrained OSNet backbone and the nature of the video datasets. These contain only a small number of identities to recognize and also equal recording conditions across the video frames. Still, the precision of the network is raised substantially and optimized using abovementioned methods. The visualization by means of activation maps and similarity rankings (figure 7 and 8 demonstrates its capabilities to re-identify animal individuals and find corresponding images in the gallery. Thereby, the head region and especially ears reach highest activation. The top ranked images of the gallery usually show frames very close to the query, since the overall images are more similar. Table 2 shows model parameters describing the overall complexity if standard architecture is used. It can be observed that the network keeps its light weight, as number of parameters and flops retain the original order of magnitude.

Model complexity	Parameters	Flops
BayWald	$5.07 \cdot 10^6$	$2.80 \cdot 10^9$
Wildpark	$6.70 \cdot 10^6$	$2.10 \cdot 10^9$
Market1501	$9.35 \cdot 10^6$	$2.43 \cdot 10^9$

Table 2: Estimates on model complexity for different datasets. Parameters: weights and parameters to describe model. Flops: floating point operations per second.

4.2 Frame dropping

The idea of dropping frames aims to reduce background bias and redundancy. Consecutive frames show nearly identical animal poses and recording conditions. Therefore, given a query image the re-identification system can find many similar images in the gallery if every frame is used. Naturally, background becomes more significant than the actual identity. Frame dropping can also be utilized for training depending on dataset size, video length, framerate and amount of animal movement. While decreasing redundancy in the gallery, frame dropping has no further influence because similar and less similar video frames are dropped equally.

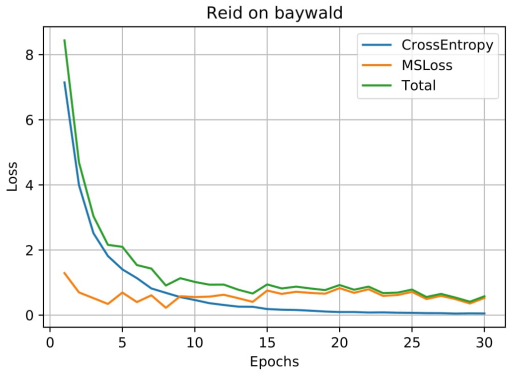

Figure 5: Loss during training on BayWald.

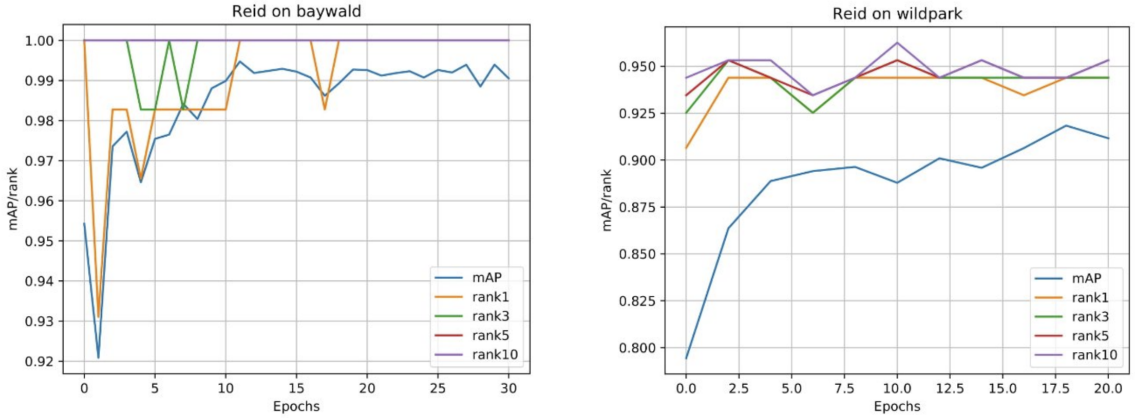

Figure 6: mAP/epoch for BayWald (left) and Wildpark (right). Configuration: FD + SQG.

Figure 9 shows that frame dropping results in a more volatile progression of mAP. Also in this case, the rank-k accuracy is not constant due to less similar gallery images. In this work different grades of frame dropping were tested with the observation that moderate frame dropping does not effect overall performance significantly while decreasing redundancy. That means on the other hand, with the lack of animal movement it fails to reduce background bias.

4.3 Time-based separation of query and gallery

Having the same goal as frame dropping, the purpose of a time-based separation of the query and gallery sets is to reduce background and recording condition similarity between these two sets by leaving out a group of consecutive frames in the middle of the video. Query images are then obtained from the beginning or the end and gallery images the other way around. When comparing figures 6 and 9 it can be seen that the starting mAP of the untrained system is lower if separation is applied. Re-identification becomes more challenging because surrounding frames of the query can not be found in the gallery anymore. At the same time, the model can handle this complication after some training on animal datasets. In Wildpark, some individuals are only for a part of the video duration on scene, so the separation needs to be dynamically for every individual to be represented in query and gallery.

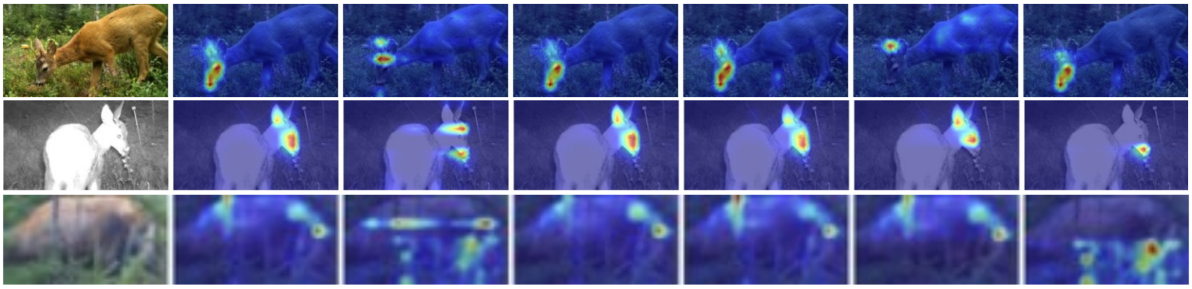

Figure 7: Activation maps. From left to right: original image, global features, global + drop block, first half of channel features, second half, upper part, lower part.

Figure 8: Similarity ranking. Top ranked gallery images given a query (left).

4.4 Influence of Backbone and Branches

The design of the Lightweight Multi-Branch Network architecture relies on both a strong performance using multiple branches as well as high efficiency mainly achieved through the OSNet backbone. For this work, modification of the code and the network structure were made in order to evaluate single branches separately (figure 3). Of the three branches the channel branch performed the best and the global branch the poorest. This might indicate that global features play a less important role for animal video data oppositely to the observations of Herzog et al. regarding person re-identification. Also the system achieves worse results if the OSNet backbone is replaced by a ResNet backbone even though the latter does increase model parameters slightly and flops by the factor of 5.

Architecture	BayWald	Wildpark
OS + G	0.9836	0.8944
OS + C	0.9958	0.9322
OS + P	0.9928	0.9273
OS + G + C + P	0.9995	0.9320
RES + G + C + P	0.9957	0.9117

Table 3: mAP per network architecture. OS: OSNet backbone. RES: ResNet50 backbone. G: global branch. C: channel branch. P: part branch.

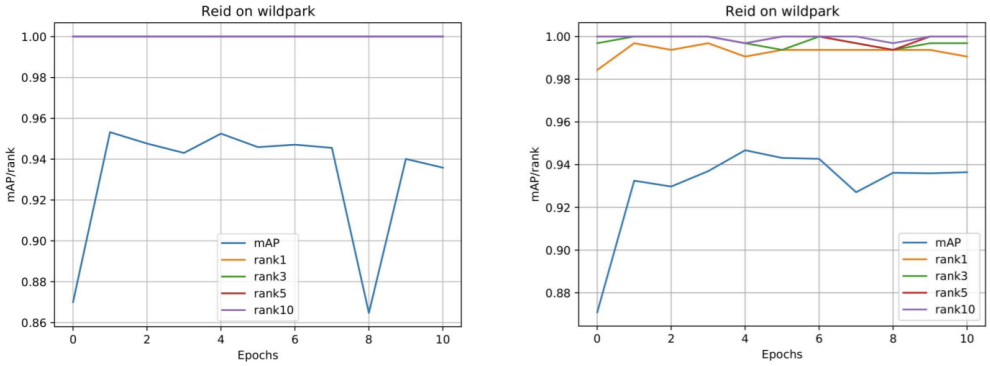

Figure 9: Effect of frame dropping on mAP. Left: baseline. Right: frame dropping.

4.5 Cropping to masks instead of bounding boxes

If images are cropped using masks instead of bounding boxes, the similarity can only be measured by regarding only the depicted animal individual. At the same time, optimally, the network learns that background information is irrelevant for individual re-identification and if the background is dropped out, this might not be the case. Since feature activation extracted by the pretrained OSNet already is mostly only focused on the animal body (figure 7, no enhancement by using masks can be found in visualization in that regard. However, table 1 shows improved performance on BayWald and decreased mAP on Wildpark. Performance is always evaluated using bounding boxes as before. As mentioned, evaluation on Wildpark does not include re-identification across videos. That is why similar background can lead to an imminent recognition and non-masked images facilitate the prediction. So conclusively, results led to the belief that training on masked images improves the networks ability measuring similarity of animal images.

5 Conclusion

In this work the Lightweight Multi-Branch Network as introduced by Herzog et al. [7] is applied for re-identification of animal individuals in two video datasets BayWald and Wildpark. Wildpark is partially annotated using video-based framework MiVOS and image-based annotation tool VGG Image Annotator. The original methods are optimized and complemented by video data processing additions such as image-cropping, frame dropping and query-gallery-separation. These techniques achieve an improvement of both training and evaluation process. Model complexity, performance, feature activation and ranking output are inspected, visualized and evaluated considering different configurations of methods and network structure. The results demonstrate that the Lightweight Multi-Branch architecture is well-suited for Animal Individual Re-Identification. For future work, the light weight of the network and the application possibilities of easy-to-maintain camera traps suggest considering data of a significantly larger scale including multiple camera positions and a sizable number of individuals. Studies of Schneider et al. [15] show that similarity learning networks such as the Lightweight Multi-Branch Network can also re-identify across different species and outperform human recognition. In combination with animal detection the re-identification system can be utilized for animal population estimation and behaviour observation among other applications.

References

[1] Brust, C. ; Burghardt, T. ; Groenenberg, M. ; Käding, Christoph ; Kühl, H. ; Manguette, Marie L. ; Denzler, Joachim: Towards Automated Visual Monitoring of Individual Gorillas in the Wild. In: *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (2017), S. 2820–2830

[2] Cheng, Ho K. ; Tai, Yu-Wing ; Tang, Chi-Keung: Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion. In: *CoRR* abs/2103.07941 (2021). – URL `https://arxiv.org/abs/2103.07941`

[3] Dutta, Abhishek ; Zisserman, Andrew: The VIA Annotation Software for Images, Audio and Video. In: *Proceedings of the 27th ACM International Conference on Multimedia*. New York, NY, USA : Association for Computing Machinery, 2019 (MM '19), S. 2276–2279. – URL `https://doi.org/10.1145/3343031.3350535`. – ISBN 9781450368896

[4] Ewert, Iva: *Video-based wild animal Re-Identification using Top-DB-Net architecture*. 2021. – Unpublished

[5] Freytag, Alexander ; Rodner, Erik ; Simon, Marcel ; Loos, Alexander ; Kühl, Hjalmar S. ; Denzler, Joachim: Chimpanzee Faces in the Wild: Log-Euclidean CNNs for Predicting Identities and Attributes of Primates. (2016), S. 51–63. ISBN 978-3-319-45886-1

[6] He, Kaiming ; Zhang, Xiangyu ; Ren, Shaoqing ; Sun, Jian: Deep Residual Learning for Image Recognition. In: *CoRR* abs/1512.03385 (2015). – URL `http://arxiv.org/abs/1512.03385`

[7] Herzog, Fabian ; Ji, Xunbo ; Teepe, Torben ; Hörmann, Stefan ; Gilg, Johannes ; Rigoll, Gerhard: Lightweight Multi-Branch Network for Person Re-Identification. In: *CoRR* abs/2101.10774 (2021). – URL `https://arxiv.org/abs/2101.10774`

[8] Krizhevsky, Alex ; Sutskever, Ilya ; Hinton, Geoffrey E.: ImageNet Classification with Deep Convolutional Neural Networks. In: *Commun. ACM* 60 (2017), Mai, Nr. 6, S. 84–90. – URL `https://doi.org/10.1145/3065386`. – ISSN 0001-0782

[9] Li, Wei ; Zhao, Rui ; Xiao, Tong ; Wang, Xiaogang: DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, S. 152–159

[10] Luo, Hao ; Gu, Youzhi ; Liao, Xingyu ; Lai, Shenqi ; Jiang, Wei: Bag of Tricks and A Strong Baseline for Deep Person Re-identification. In: *CoRR* abs/1903.07071 (2019). – URL `http://arxiv.org/abs/1903.07071`

[11] Paszke, Adam ; Gross, Sam ; Massa, Francisco ; Lerer, Adam ; Bradbury, James ; Chanan, Gregory ; Killeen, Trevor ; Lin, Zeming ; Gimelshein, Natalia ; Antiga, Luca ; Desmaison, Alban ; Köpf, Andreas ; Yang, Edward ; DeVito, Zach ; Raison, Martin ; Tejani, Alykhan ; Chilamkurthy, Sasank ; Steiner, Benoit ; Fang, Lu ; Bai, Junjie ; Chintala, Soumith: PyTorch: An Imperative Style,

High-Performance Deep Learning Library. In: *CoRR* abs/1912.01703 (2019). – URL http://arxiv.org/abs/1912.01703

[12] Phillott, Andrea ; Skerratt, Lee ; Mcdonald, Keith ; Lemckert, Frank ; Hines, Harry ; Clarke, John ; Alford, Ross ; Speare, Rick: Toe-Clipping as an Acceptable Method of Identifying Individual Anurans in Mark Recapture Studies. In: *Herpetological Review* 38 (2007), 09, S. 305–308

[13] Quispe, Rodolfo ; Pedrini, Hélio: Top-DB-Net: Top DropBlock for Activation Enhancement in Person Re-Identification. In: *CoRR* abs/2010.05435 (2020). – URL https://arxiv.org/abs/2010.05435

[14] Schindler, Frank ; Steinhage, Volker: Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. In: *Ecological Informatics* 61 (2021), S. 101215. – URL https://www.sciencedirect.com/science/article/pii/S1574954121000066. – ISSN 1574-9541

[15] Schneider, Stefan ; Taylor, Graham W. ; Linquist, Stefan S. ; Kremer, Stefan C.: Similarity Learning Networks for Animal Individual Re-Identification - Beyond the Capabilities of a Human Observer. In: *CoRR* abs/1902.09324 (2019). – URL http://arxiv.org/abs/1902.09324

[16] Vidal, Maxime ; Wolf, Nathan ; Rosenberg, Beth ; Harris, Bradley P. ; Mathis, Alexander: Perspectives on Individual Animal Identification from Biology and Computer Vision. In: *Integrative and Comparative Biology* (2021), 05. – URL https://doi.org/10.1093/icb/icab107. – ISSN 1540-7063

[17] Wang, Xun ; Han, Xintong ; Huang, Weiling ; Dong, Dengke ; Scott, Matthew R.: Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. In: *CoRR* abs/1904.06627 (2019). – URL http://arxiv.org/abs/1904.06627

[18] Yadav, Ankit ; Vishwakarma, Dinesh K.: Person Re-Identification using Deep Learning Networks: A Systematic Review. In: *CoRR* abs/2012.13318 (2020). – URL https://arxiv.org/abs/2012.13318

[19] Zheng, Liang ; Shen, Liyue ; Tian, Lu ; Wang, Shengjin ; Wang, Jingdong ; Tian, Qi: Scalable Person Re-identification: A Benchmark. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, S. 1116–1124

[20] Zheng, Liang ; Yang, Yi ; Hauptmann, Alexander G.: Person Re-identification: Past, Present and Future. In: *CoRR* abs/1610.02984 (2016). – URL http://arxiv.org/abs/1610.02984

[21] Zhou, Kaiyang ; Xiang, Tao: Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch. In: *CoRR* abs/1910.10093 (2019). – URL http://arxiv.org/abs/1910.10093

[22] Zhou, Kaiyang ; Yang, Yongxin ; Cavallaro, Andrea ; Xiang, Tao: Omni-Scale Feature Learning for Person Re-Identification. In: *CoRR* abs/1905.00953 (2019). – URL http://arxiv.org/abs/1905.00953