Part-Pose Guided Amur Tiger Re-Identification

Cen Liu, Rong Zhang, Lijun Guo [*]
Faculty of Electrical Engineering and Computer Science, Ningbo University, China
liucen05@163.com, {zhangrong, guolijun}@nbu.edu.cn

Abstract

In this paper, we present our solution to tiger re-identification (re-ID) in both the plain and the wild tracks in the 2019 Computer Vision for Wild life Conservation Challenge (CVWC2019). We introduce a novel part-pose guided framework for the tiger re-ID task, which consists of two part streams and one full stream based on the pose characteristics of tiger. Considering missing and inaccurate pose annotations, the two part streams are used as a regulator to guide the full stream in learning and aligning the local features in the training stage. We only use the learnt full stream for the tiger re-ID task in the inference stage. The proposed model has the advantage that despite requiring pose information at training time it is not needed during inference, so it is particularly suitable for tiger re-ID in the wild. Our proposed method outperforms the state-of-the-art and finished top in both the PlainID and WildID competitions at CVWC2019. The source of code will be public available at https://github.com/LcenArthas/CVWC2019-Amur-Tiger-Re-ID

1. Introduction

Object re-ID is a challenging task in the field of computer vision because of the changes in illumination, camera viewpoint, background and occlusions. Specifically, re-ID means given a query image containing the target individual, a re-ID algorithm will retrieve all the images of the same identity from a large gallery and return ranking results. Previous studies on re-ID mainly focus on person re-ID and have achieved fruitful research results in the field of pedestrian tracking and public safety [30]. In recent years, researchers have noticed that the re-ID can also be used to protect wildlife because it can help to obtain accurate population counts and track wildlife trajectory [11, 14, 1]. In this paper, we are concerned about wild Amur tiger protecting using re-ID technique.

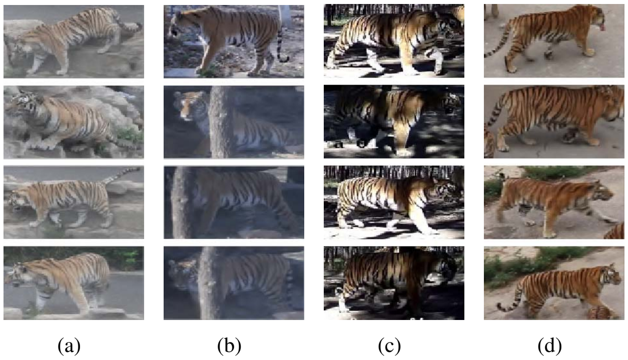


Figure 1: Examples of Amur tiger samples in the ATRW. (a) A tiger with different poses. (b) A tiger with different degrees of occlusion. (c) A tiger under unconstrained illumination conditions. (d) Different tigers with very similar stripes.

Conceptually, both the re-identification of person and tiger are image retrieval problems. Thus, some common person re-ID strategies are also useful in the tiger re-ID. But compared to the person re-ID, there are some different challenges in the tiger re-ID field, as shown in Figure 1. First, wildlife data have a wide range of pose variations due to unrestricted four-limbed movement. Second, because Amur tigers have a wide range of activities and images are captured from the wild environment, there will be more occlusion, more complex natural background and unconstrained illumination conditions. In addition, the existing re-identification method of Amur tiger is mainly based on body stripes [10], which are more ambiguous because of the similarities among different individuals than a person's appearance.

In recent years, many person re-ID methods used local features extracted from various parts of the human body to improve the global representation ability. For such methods, alignment is the key factor to make local features work. A simple alignment is to partition the person image into a few fixed horizontal stripes straightforwardly [21]. However, such a partition cannot well align the human body parts well. Recent works have attempted the use of pose

[*] Corresponding author. Email: guolijun@nbu.edu.cn

(keypoints of body parts) to localize body parts for learning part-aligned features. However, posture-based alignment of body parts easily fails due to missing or inaccurate pose information, especially in tiger re-ID task.

To address the above challenges, we introduce the part-pose guided network (PPGNet) for the Amur tiger re-ID task. In our method, we use local image features based on pose data to drive the main network to learn alignment features from the original image. As shown in Figure 2, our network consists three streams with one global stream (full stream) and two local streams (part streams) according to the inputs of different components. Meanwhile, considering missing and inaccurate pose annotations, the two local streams are used as regulators to guide the original full image stream in learning and aligning the local features. In the inference stage, we only use the full image stream, which have learned about local part information and alignment features. Experiments show that our PPGNet performs well with high efficiency and accuracy in the Amur tiger re-ID task. Moreover, we found it is also very suitable for Amur tiger re-ID in the wide since it doesn't need the pose estimation step.

In summary, our major contributions are:

- We propose a novel part-pose guided network (PPGNet) tailored for tiger re-ID. The proposed framework uses limited pose annotation data as regulator to impose an alignment constraint on global feature learning.
- The proposed model can greatly reduce the amount of calculation because it only use the full stream in the inference stage. Meanwhile, it is also very suitable for tiger re-ID in the wild since it doesn't need the pose estimation step.
- Our algorithm achieved the first place both in the plain and wild tiger re-ID tracks at the 2019 Computer Vision for Wildlife Conservation Challenge.

2. Related Work

2.1. Person and Wildlife Re-ID

Person re-identification is an active research topic which has been paid more and more attention by academia and industry in recent years. The success of deep learning methods for person re-ID is well documented with the improved computational power and the availability of large datasets [29, 16]. Most of these methods use deep metric learning [17, 5, 4], while some others regard person re-ID as a classification problem [31, 24, 27]. Recent studies have proposed an optimization model that combines classifying loss and measuring learning loss to improve the re-identification rate and accelerate the convergence of the model [3, 22].

Inspired by the success of deep learning methods for person re-ID, researchers are aware that re-ID can also be used for wildlife conservation because of its ability to obtain accurate population size and track wildlife trajectories [15, 11, 14, 1, 23]. Li *et al*. [15] introduced a large-scale the Amur Tiger Re-identification in the Wild (ATRW) dataset and proposed a novel method for Amur tiger re-identification which introduces precise pose parts modeling in deep neural networks to handle large pose variation of tigers. Weideman *et al*. [23] introduced novel combinations of integral curvature representation and two matching algorithms for identifying individual cetaceans from their fins. Körschens *et al*. [14] successfully implemented a system to assist biologists to identify elephants they encounter in the wild field. In this paper, we focus on Amur tigers re-ID based on end-to-end network.

2.2. Pose Alignment

How to deal with pose variations of pedestrian is a key factor in person re-ID. Among the existing person re-ID models based on deep learning, the alignment method based on keypoints is used to eliminate the adverse effect of pose variance. Zheng *et al*. [28], proposed a PoseBoxes method to correct the misalignment. Su *et al*. [20] proposed a Pose-driven Deep Convolutional (PDC) model based on the pose cue to learn improved feature extraction. Zhao *et al*. [26] introduced a body part specific attention modeling for person re-ID. There is also a lot of work on image misalignment from the masks and semantics [18, 12, 28]. All of the above works are based on the combination of pose information and global information for training and inferencing.

Compared with person re-ID, wildlife data have a wide range of pose variations due to unrestricted four-limbed movements. Additionally, the models cannot obtain accurate joints information or miss some joints information because the data are taken from the wild environment with more complex natural backgrounds, unconstrained lighting conditions and more occlusion. So it is not feasible to transfer the alignment method from the above person re-ID models to a tiger re-ID model directly. Inspired by [25], we leverage the body pose parts, which play the role of regulators, to guide the feature learning from the original image during the training stage. Thus, we design a triplet-branched network which highlight the feature extraction of tiger trunk and use the limbs symmetry characteristics according to the peculiarities of tiger. Different from the triplet-branch network in [2], we only use the learnt full image stream which have learnt local part information and alignment features.

3. Methodology

This work proposes the PPGNet which can fully exploit aligned part-pose representations to guide the learning of

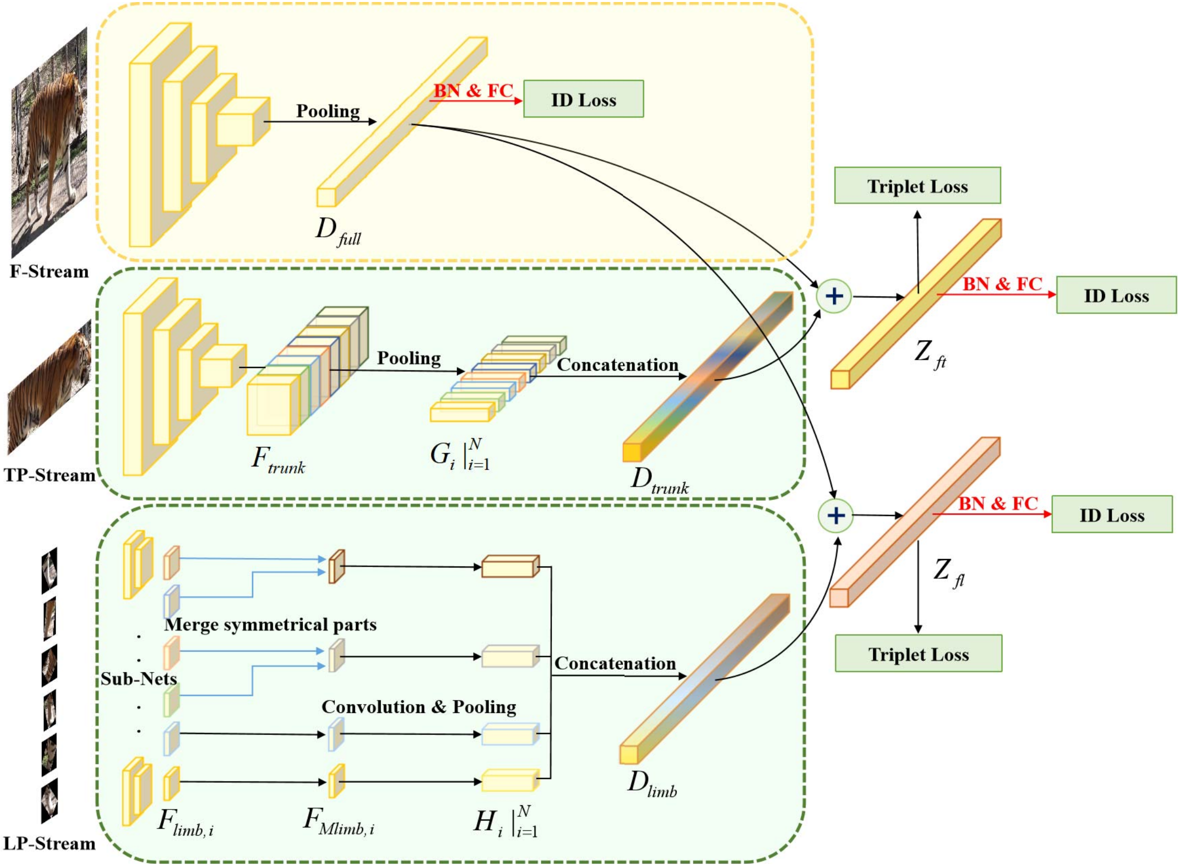


Figure 2: The PPGNet Structure. It consists of three streams: one F-Stream and two part streams. The F-stream is the main stream and used to extract the global feature from the full tiger image, while the two part streams, the TP- and LP- streams, are used to extract local features from the trunk image blocks and limb image blocks, respectively. In the training stage, the local features play the role of regulator for the global feature learning. In the inference stage, only the F-stream is used for tiger re-ID.

global feature for the tiger re-ID. Figure 2 gives the pipeline of our method. We construct two part streams and one full stream based on the pose characteristics of tiger. The part streams includes TP-Stream which takes trunk part of tiger image as input, and LP-Stream, which takes all the limbs of tiger image as input based on the pose skeleton. The input of the full stream (F-Stream) is the full original image. Through final feature fusion and loss design, the both part streams can be used as a regulators to constrain F-Stream feature learning from the original image. We elaborate the PPGNet design as follows.

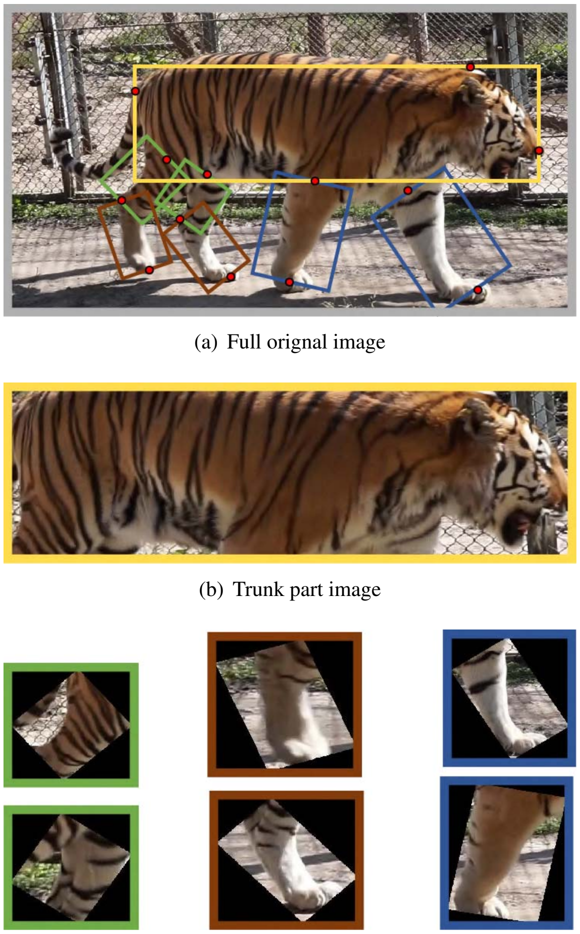
3.1. Part Images

Based on the pose skeleton annotations, 7 body parts can be cropped as shown in Figure 3. These parts include trunk, left and right of front legs, hind thighs, and hind shanks, covering almost the entire body. Except for the rectangle of the trunk, the remaining limb quadrilateral parts are combined with an outer rectangle in a black background(the RGB values are all zero).

As far as the details are concerned, the trunk is confined by four body joints, the ear, nose, shoulder and tail, so we simply draw a quadrangle for the trunk. The front legs are defined as the joints of the shoulder and front paw. A hind thighs are confined by the hip and knee joints, and the hind shanks are confined by the knee and back paw joints. We manually set the width of each front leg, hind thigh and hind shank bounding box to $\frac{1}{6}$, $\frac{1}{6}$ and 1 of its height respectively. It is worth mentioned that we add some small random disturbances when cropping part images to improve the robustness of the model.

3.2. Part-Pose Guided Network

Because tigers are non-rigid quadrupeds and have more abundant postures than person, coupled with the complex background and occlusion in the wild, it is impossible to


(a) Full orignal image

(b) Trunk part image

(c) Limbs part image

Figure 3: Examples of the original image and the part images. (a) The full original image, in which skeleton annotations are given. The colorful rectangles are the part image block cropped based on the skeleton annotations. (b) The trunk part image. (c) The limbs part images, including the hind thighs, the hind shanks and the front legs from the left to the right.

obtain accurate tiger pose information in most cases. Even in some cases, pose data is unreliable because some joint points might be missed. So the effect of tiger re-ID will not be improved if we adopt the same strategy as the previous work [12, 28], in which the final features are obtained fusing the global and the local features directly based on pose data. Inspired by [25], in order to make full use of local features and solve the above challenges, we propose to use limited pose annotation data to guide the features learning from original images in training stage. That means to impose a alignment regularization constraint on global feature learning. In the inference stage, we only use the original full stream image and the pose information is no longer needed. This can greatly reduce the amount of calculation. So the PPGNet is outstandingly effective since it not only is in line with the needs of the Amur tiger re-ID system in the wild but also doesn't need the pose estimation except for detection.

Part Streams. In our design, the part pose stream includes TP-Stream and LP-Stream with trunk and limb part images as input. The main considerations are as follows. 1) The trunk part occupies the main part of the tiger image. The stripes from the trunk are abundant and contain more fine-grained features, so it has stronger ability to represent the individual tiger. 2) Different from the trunk image, the feature representation of the limbs image is more easily affected by occlusion because of its small relative size. 3) Due to the symmetry of limbs, we will consider this symmetry in network design. The details of these two part stream structure is shown in Table 1.

Trunk Part Stream. The backbone of the network is pre-trained on ImageNet dataset [19] to extract feature map F_{trunk} from the trunk part image. As mentioned above, the trunk part of a tiger has the richest stripe that can identify the individual tiger. here we use the method proposed by [21], the output F_{trunk} of the backbone network is divided into 8 vertically stripes, each stripe undergoes average pooling, and then generates 8 corresponding 512-dimension vector $G_i|_{i=1}^{N}$, where N is 8. The extracted fine-grained features are concatenated along channels and we have the final trunk feature vector D_{trunk}.

Limbs Part Stream. To learn the local details of individual parts area rather than mixing them all together, subnetworks with multiple branches are used to learn the limbs feature map. We use the symmetry of tiger limbs to extract features in two steps, and merge the features of the symmetrical limbs to reduce the computational complexity by reducing the number of branches. First, we use the 6 parallel independent pre-trained networks to learn the local feature $F_{limb,i}$, $i = 1, 2..., 6$. According to the image symmetry of left and right limbs, we merge the feature maps of the hinds legs left-right separately finally obtain 4 new local feature $F_{Mlimb,i}$, $i = 1, 2..., 6$. In the second stage, similarly, we merge the two branches corresponding to the left and right symmetrical parts of the hind legs, and finally get 4 branches, as shown in Table 1. After an average pooling, generate 4 corresponding 1024-dimension vector $H_i|_{i=1}^{N}$, where N is 4, and then similar to trunk-Stream, concatenate the H_i along channels. Finally, we obtain the limb feature vector D_{limb}.

Full Stream. In this stream, we mainly use a classification network to extract rich feature representations. Specifically, we also use the pre-trained series network of ResNet [8] as the backbone, followed by an average pooling layer to get the global feature vector D_{full}. It's worth mentioning that we set the last stride of ResNet from 2 to 1, there are two advantages to doing so. 1) We can get larger feature maps, so that the higher spatial resolution of features, the stronger the feature representation. 2) Make the obtained global features D_{full} dimension consistent with the local feature D_{trunk} and D_{limb} obtained above, convenient next feature fusion. Note that only the D_{full} extracted from the F-Stream are used for the final re-ID in the inference phase.

Table 1: The details of the TP-Stream and LP-Stream structure. We use a structure similar to ResNet-34 [8]. #Bran. denotes the number of sub-branches.

Layer name		**Parameters**	**Output size**	**#Bran.**
TP-Stream	conv1	7×7, 64, stride 2	32×64	1
	conv2_x	3×3, Max pool, stride 2	16×32	
		$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times2$		
	conv3_x	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times2$	8×16	
	conv4_x	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times2$	4×8	
	Average pool	4×1	1×1	8
LP-Stream	conv1	7×7, 64, stride 2	32×32	6
	conv2_x	3×3, Max pool, stride 2	16×16	
		$\begin{bmatrix}3\times3, 64\\3\times3, 64\end{bmatrix}\times 2$		
	conv3_x	$\begin{bmatrix}3\times3, 128\\3\times3, 128\end{bmatrix}\times 2$	8×8	
	conv4_x	$\begin{bmatrix}3\times3, 256\\3\times3, 256\end{bmatrix}\times 2$	4×4	
	merging	add element	4×4	6 → 4
	conv5_x	$\begin{bmatrix}3\times3, 512\\3\times3, 512\end{bmatrix}\times 2$	2×2	4
	Average pool	2×2	1×1	

3.3. Feature Fusion

In order to use local part information to guide learning and alignment of global features, we fuse part and global features of the three streams by adding corresponding elements:

$$Z_{ft} = D_{trunk} + D_{full} \tag{1}$$

$$Z_{fl} = D_{limb} + D_{full} \tag{2}$$

where Z_{ft}, Z_{fl} denote the fused feature from the TP-Stream and LP-Stream with F-Stream, respectively.

Table 2: Different performance with different F-Stream backbones on the plain re-ID test set.

Backbones	**mmAP**	**Single-Cam**			**Cross-Cam**		
		mAP	**Rank-1**	**Rank-5**	**mAP**	**Rank-1**	**Rank-5**
ResNet-50	0.779	0.896	0.994	0.994	0.663	0.908	**0.977**
ResNet-101	**0.810**	**0.907**	**0.995**	0.991	**0.712**	**0.931**	0.971
ResNet-152	0.802	0.902	0.991	**0.997**	0.703	0.931	0.965

3.4. Loss Function

For the tasks of person [3, 22] and wildlife [1, 15] re-identification, both the ID loss(cross entropy loss) and the triplet loss [9] have been widely used for optimizing the network. In our design, we also use the combination of these two losses to optimize our network in training phase.

Among all the learned features, we supervise the feature D_{full} from the F-Stream and the fusion features, i.e. Z_{ft} and Z_{fl}, as shown in Figure 2. Specifically, a ID loss is computed over the the original full image feature vector D_{full} extracted by the F-Stream, and in the two part streams a ID loss and a triplet loss are separately calculated using the fused features, Z_{ft} and Z_{fl}.

Z_{ft} and Z_{fl} are made up of the globel feature and the local feature which are extracted from the F-Stream and Part-Streams, respectively, so the generation of loss depends not on a single branch but on all the three streams. When gradient back-propagation is performed in training phase, the F-Stream will suffer gradient loss computed by two fusion features from the two pose part streams. So the F-Stream can always be affected by local features to adjust network parameters. That is to say, these two branches play a similar regularization role in guiding learning features for the main stream in the training phase.

For the ID loss, each feature vector is followed by a batch normalization (BN) layer and a fully connected (FC) layer. And we use the method proposed in [7] to initialize the FC layers.

4. Experiments

4.1. Dataset

The ATRW dataset is proposed by [15] and is a benchmark dataset for Amur tiger re-identification. The dataset consists of 3649 bounding box annotations of 182 entities of 92 tigers. In this dataset, not all entities appear cross camera but 50 entities are cross-camera, the remaining are different frames from a single camera. There are 60% entities from single-camera and 40% entities from cross-camera in the training set and the remaining images are in the test set, i.e., 1887 images are in the training set and 1762 images in the testing set. The testing set is the query set and also the gallery set. This dataset also provides the key-points an-

Table 3: The top10 teams in the Plain Tiger ReID track. Our results are in the first place.

Team	mmAP	Single-Cam			Cross-Cam		
		mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Bestfitting_NBU	**0.816**	**0.906**	**0.977**	**0.991**	**0.726**	**0.936**	**0.967**
BRL-RedPanda	0.770	0.898	0.966	0.977	0.643	0.913	0.958
NWPU_ASGO	0.751	0.879	0.969	0.983	0.622	0.925	0.951
DeepBlueAI	0.704	0.865	0.956	0.983	0.543	0.889	0.929
DelPro	0.696	0.836	0.973	0.981	0.556	0.872	0.948
SDL	0.672	0.857	0.940	0.960	0.488	0.783	0.867
zdi	0.658	0.846	0.954	0.984	0.470	0.841	0.904
NDWild	0.658	0.763	0.907	0.967	0.553	0.851	0.944
Batiary	0.634	0.757	0.900	0.967	0.511	0.824	0.935
aaa	0.631	0.758	0.906	0.964	0.505	0.806	0.917

notations for each image, but the shaded joints on the body are not marked.

4.2. Evaluation Protocol

We use the mean average precision (mAP) and the rank-k accuracy as the metric to evaluate the algoritm. And according to the situation of the query image appearing in camera, separate each query image into two fileds: ‘single camera’, where the target tiger appears only in single camera, and ‘cross camera’ where the target appears in the multiple cameras [15]. And in this challenge, for the re-ID task, the final ranking metric is the average of mAP on both the single-cam case and the cross-cam case (mmAP).

4.3. Implementation Details

Data Augmentation. To enlarge the dataset, we horizontally flip the images in the training set to create more ’new entities’ since different sides of the same Amur tiger are regarded as different entities [15]. So we have double the number of the original training set. During the training for re-ID, each original image is resized into 256 × 512, each trunk part image 64 × 128 and each limb part image 64 × 64. Three types of data augmentations are applied to each image input the network: 1) Rotate with a degree randomly sampled from -5 to 5 degrees. 2) Randomly change the brightness, contrast and saturation range from 0.8 to 1.2 respective. 3) Random affine transformation of image.

Backbones. For the TP-Stream and multi-branch subnets in the LP-Stream, we use the ResNet-34 [8] as the backbone to obtain the feature map. And for the F-Stream, we know from experiments that different backbones of F-Stream may get different performance. As shown in Table 2, ResNet-101 [8] arrives the best performance in the plain re-ID testing set.

Table 4: Our method compeared with the baseline.

Setting	Method		Single-Cam			Cross-Cam		
			mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Plain	Baseline [15]	CE	0.591	0.786	0.927	0.381	0.697	0.878
		Triplet loss	0.713	0.866	0.960	0.472	0.776	0.906
		Aligned-reID	0.648	0.812	0.924	0.442	0.738	0.905
		PPbM-a	0.741	0.882	0.964	0.517	0.768	0.910
		PPbM-b	0.728	0.894	0.956	0.478	0.771	0.907
	Ours(ResNet101 + Rerank)		**0.906**	**0.977**	**0.991**	**0.726**	**0.936**	**0.967**
Wild	Baseline [15]	CE	0.588	0.787	0.925	0.345	0.685	0.868
		Triplet loss	0.707	0.865	0.951	0.452	0.776	0.905
		Aligned-reID	0.587	0.748	0.907	0.410	0.701	0.872
		PPbM-a	0.710	0.874	0.966	0.503	0.772	0.907
		PPbM-b	0.692	0.889	0.953	0.462	0.766	0.912
	Ours(ResNet101 + Rerank)		**0.889**	**0.956**	**0.974**	**0.724**	**0.929**	**0.953**

Optimization. In our experiments, the ID loss for the F-Stream, the ID loss and triplet loss for the fused features are weighed by 1.0, 1.5 and 2.0 respectively. The training is optimized by Adam [13] optimizer using 500 epochs and with a batch size of 32. Meanwhile, we adopt the warmup strategy to bootstrap the network for better performance. We spent 25 epochs linearly increasing the learning rate from 2.5×10^{-4} to 2.5×10^{-3}. Then, the learning rate is decayed by a factor of 0.5 for every 80 epochs. In our experiments, we use the Pytorch framwork to train and test our model.

Table 5: The rank results in the Wild Tiger ReID track. Our results are in the first place.

Team	mmAP	Single-Cam			Cross-Cam		
		mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Bestfitting_NBU	**0.807**	**0.889**	**0.956**	**0.974**	**0.724**	**0.929**	**0.953**
DeepBlueAI	0.696	0.845	0.944	0.967	0.548	0.908	0.941
Batiary	0.666	0.789	0.909	0.959	0.543	0.856	0.942
zdi	0.644	0.823	0.932	0.967	0.465	0.849	0.910

4.4. Tiger Re-ID Performance

In the 2019 Computer Vision for Wildlife Conservation Challenge (CVWC2019), the re-ID task is divided into two tracks, the Plain Tiger Re-ID and the Wild Tiger Re-ID. In the plain track uses the the cropped Amuer tiger bounding-box and keypiont annotations as introduced in the Section 4.1 for re-ID task. And in the wild track we aim to design a complete tiger re-identification system based on automatic detection and tiger pose estimation [15].

Here we briefly show our results of the Tiger re-ID in the plain and Tiger re-ID in the wild at this challenge.

Plain Tiger ReID. Table 3 shows the performances of top10 teams in the Plain Tiger ReID track. It is seen that our team achieves the top-1 in this task and overtake other teams by a large margin, which shows advancement

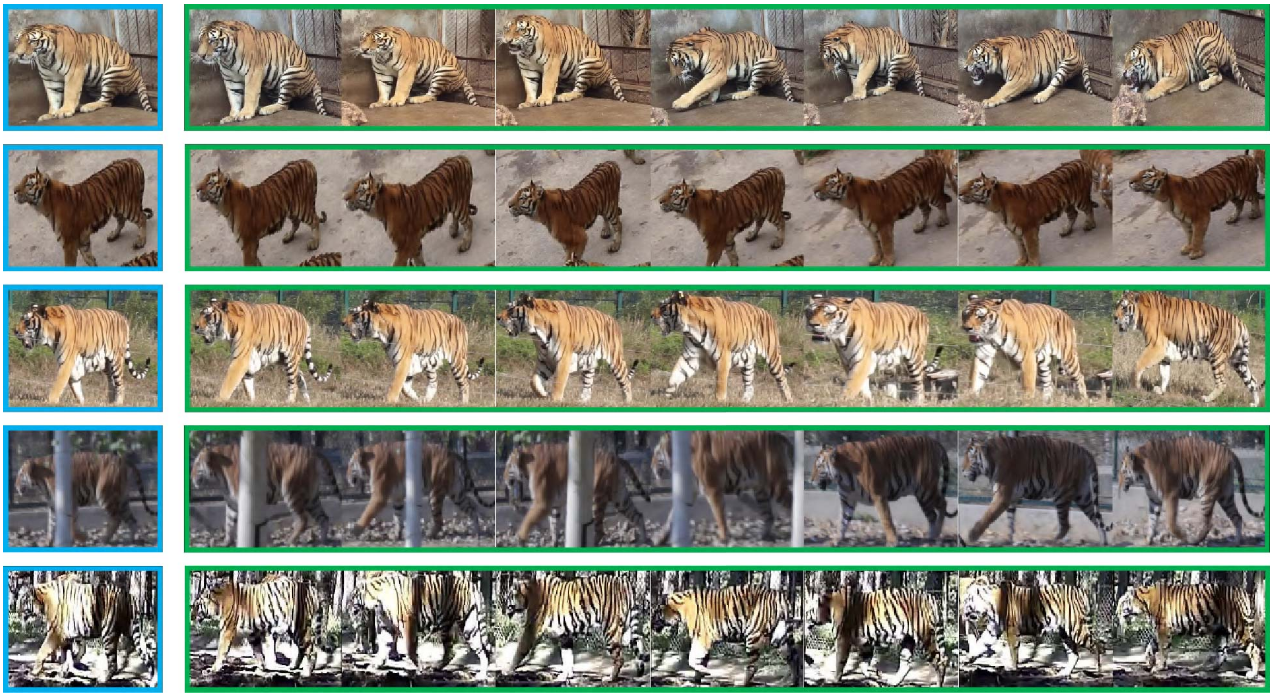


Figure 4: Example results of our method on ATRW. From left to right in each row, the first with the blue border is the query image and the remaining are the top-7 ranking results.

and high performance of PPGNet. Table 4 lists the mAP and top-k results for compared with baseline method in [15]. Examples re-ID results of our method are shown in Figure 4.

Wild Tiger ReID. In this track, for the detection module, we adopt Mask RCNN [6] as the detector since they are open-sourced with the state of the art results in the field of object detection. Then put the detected tigers images into PPGNet for the re-ID purpose. With the PPGNet, we also won the championship of this track. Table 5 shows the performances teams in the Wild Tiger ReID track.

5. Conclusions

In this paper, we present our solution to plain re-ID and wild re-ID Challenges on CVWC2019. We build a novel part-pose guided network (PPGNet) for the tiger re-ID task. Some data augmentations based on the characterizes of tiger re-ID are adopted. We arrive at 81.6% mmAP and 80.7% mmAP on plain re-ID and wild re-ID testing set respectively. Our proposed method ranks the first on both re-ID tracks. In the future, we will continue to focus on the research of computer vision for wildlife conservation.

6. Acknowledgement

We would like to thank the Maker Collider and WWF Amur tiger and leopard conservation programme team for their contribution to the ATRW dataset and the organizers of CVWC2019. This work is supported by the Zhejiang Provincial Natural Science Foundation of China (No. LY17F030002), the Zhejiang Provincial Public Welfare Technology Research Project (No. LGF18F020007) and the Ningbo Municipal Natural Science Foundation of China (No.2018A610057, 2018A610163)

References

[1] S. Bouma, M. D. Pawley, K. Hupman, and A. Gilman. Individual common dolphin identification via metric embedding learning. In *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2018. 1, 2, 5

[2] S. Branson, G. V. Horn, P. Perona, and S. Belongie. Improved bird species recognition using pose normalized deep convolutional nets. In *British Machine Vision Conference*, 2014. 2

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. A multi-task deep network for person re-identification. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2, 5

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE con-*

ference on computer vision and pattern recognition, pages 1335–1344, 2016. 2

[5] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[7] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 5

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 6

[9] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[10] L. Hiby, P. Lovell, N. Patil, N. S. Kumar, A. M. Gopalaswamy, and K. U. Karanth. A tiger cannot change its stripes: using a three-dimensional model to match images of living tigers and tiger skins. *Biology letters*, 5(3):383–386, 2009. 1

[11] B. Hughes and T. Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, 122(3):542–557, 2017. 1, 2

[12] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018. 2, 4

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] M. Körschens, B. Barz, and J. Denzler. Towards automatic identification of elephants in the wild. *arXiv preprint arXiv:1812.04418*, 2018. 1, 2

[15] S. Li, J. Li, W. Lin, and H. Tang. Amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586*, 2019. 2, 5, 6, 7

[16] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 2

[17] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015. 2

[18] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*, 2018. 2

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 4

[20] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3960–3969, 2017. 2

[21] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 1, 4

[22] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 274–282. ACM, 2018. 2, 5

[23] H. J. Weideman, Z. M. Jablons, J. Holmberg, K. Flynn, J. Calambokidis, R. B. Tyson, J. B. Allen, R. S. Wells, K. Hupman, K. Urian, et al. Integral curvature representation and matching algorithms for identification of dolphins and whales. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2831–2839, 2017. 2

[24] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016. 2

[25] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019. 2, 4

[26] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017. 2

[27] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 2

[28] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 2019. 2, 4

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2

[30] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[31] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 2