An improved architecture for part-based animal re-identification through semantic segmentation distillation

Eugênio Dias Ribeiro Neto[1,3], Marc Chaumont[1,2], Gérard Subsol[1],
Michel de Garine-Wichatitsky[3], Hélène Guis[3]

[1]LIRMM, Univ Montpellier, CNRS, France
[2]IRISA, Univ Bretagne Sud, France
[3]CIRAD, France

ediasribeiro@lirmm.fr, marc.chaumont@irisa.fr, gerard.subsol@lirmm.fr,
michel.de_garine-wichatitsky@cirad.fr, helene.guis@cirad.fr

Abstract

Wildlife re-identification (Re-ID) is critical for non-invasive monitoring. Yet, animal Re-ID performances remain far behind person Re-ID due to limited datasets and a greater fine-grained appearance variability between individuals. One strategy is to adopt part-based methods in order to more precisely attend to distinct anatomical regions. To adapt to animal Re-ID, we propose PAW-ViT (Part-AWare animal re-identification Vision Transformer), a ViT that replaces the standard classification token with K learnable part tokens, each specialized to a specific anatomical region of the animal. Spatial specialization is achieved via feature-based knowledge distillation by training each token's attention to image patches to produce a semantic segmentation mask. An additional aggregation token fuses the part embeddings into a single part-aware descriptor. Trained with a multi-task loss, PAW-ViT outperforms state-of-the-art methods in animal Re-ID on ATRW (Amur tigers) and YakREID-103 (yaks), particularly in scenarios of strong viewpoint variations like the cross-camera setting.

1. Introduction

With the rapid evolution of video monitoring systems, object re-identification (Re-ID) has emerged as an important challenge in computer vision. Given a *query* image, the goal is to develop a model to automatically retrieve images of the same individual from a *gallery* of different images, despite the variations in viewpoint, illumination or occlusions [7]. Figure 1 illustrates this task.

Deep learning has driven most recent advances, relying mainly on convolutional neural networks (CNNs) and, more recently, vision transformers [35, 37]. Deep learning is also being adopted for animal Re-ID [13, 15, 22, 46]. Automated wildlife identification enables non-invasive monitoring and tracking, consequently allowing a better understanding of animal population dynamics and efforts on wildlife conservation. However, animal Re-ID still lags far behind person Re-ID. Some reasons to explain this gap are the fine-grained variability between individuals, the broad range of camera viewpoints, natural habitats, and large pose variations. Additionally, the available animal benchmarks are both fewer and smaller than those for humans.

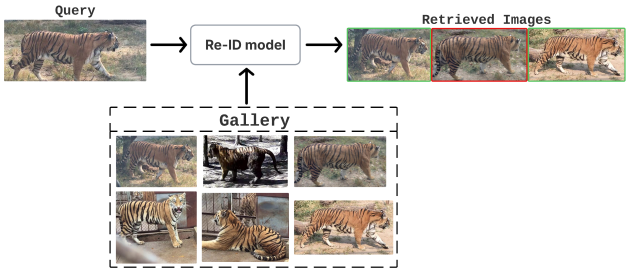

Figure 1. Overall schema of the re-identification task: given a query image, the Re-ID model must find images of the same individual in the gallery.

Many methods exploit pose or part annotations to focus on discriminative regions of animals [9, 14, 16]. For example, Liu *et al.* [16] achieves the best performances on the ATRW (amur tiger Re-identification) [15] dataset by decomposing each tiger image into pose-aligned rectangles. Despite their strong performance, pose-based methods depend on metadata that are difficult to annotate manually and are often unavailable.

To overcome this, we introduce PAW-ViT (Part-AWare animal re-identification Vision Transformer), a part-based

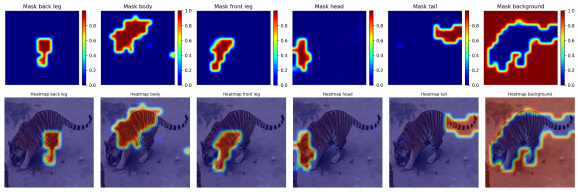

Figure 2. Illustration of attention masks generated by PAW-ViT's part tokens, our approach, on an image of the ATRW dataset. By column: 1. Hind legs, 2. Body, 3. Front legs, 4. head, 5. tail, 6. background.

ViT [4] that leverages pose information in learning phase without manual pose annotations. Rather than a single `[CLS]` token, PAW-ViT prepends K learnable part tokens to the patch embeddings, each guided to specialize in an anatomical region (e.g. head, body, tail) of the animal. The K tokens are trained jointly to predict identity, orientation (e.g. left, right), and part masks. Fig. 2 shows attention masks of PAW-ViT's extra tokens on an image of the ATRW dataset, illustrating that PAW-ViT successfully decomposes re-identification into localized anatomical parts, supporting both strong accuracy and interpretability.

In the next section, we review works on object Re-ID, focusing on animals. Section 3 details our proposed architecture PAW-ViT, and section 4 reports benchmarks on two well known animal Re-ID datasets, showing that our method achieves strong performances when compared to the current *state-of-the-art* (SOTA) methods.

2. Related Work

Person Re-Identification. Object Re-ID has seen its most rapid progress in the person Re-ID domain with the proposal of large benchmarks such as Market-1501 [43], MSMT17 [33], and DukeMTMC-reID [25]. Early methods relied on CNN backbones trained with identity classification (cross-entropy) loss [44], and metric learning losses like triplet loss [10], ArcFace [3], and circle loss [29]. More recently, Vision Transformers have been employed, exploiting the global self-attention that allows transformers to attend to the entire image at once, rather than the local receptive fields of CNNs, enabling long-range dependencies [1, 8, 45].

To deal with the fine-grained aspects of person Re-ID, some part-based methods were proposed. PCB [28] uniformly partitions convolutional feature maps into horizontal stripes, learning a separate descriptor for each part. In occluded person Re-ID, where visible regions must be matched while occluded parts ignored, part-based methods are crucial [23, 32]. Using a ViT, Somers *et al*. [27] proposes KPR, applying a token-wise part classification loss to ignore occluded tokens.

Animal Re-Identification. As stated by Ravoor *et al*. [24], "the shape or form of the animal varies significantly compared to that of humans, since human movements involve smaller changes". This underscores the need for dedicated animal Re-ID methods rather than directly applying those developed for human Re-ID.

Some studies explore the development of general Re-ID methods to handle multiple species. Cermák *et al*. [46] presents WildlifeDatasets, a toolkit compiling various public wildlife re-identification datasets. They introduce MegaDescriptor, a Swin Transformer trained with ArcFace [3]. Jiao *et al*. [13] developed UniReID (Universal ReID), a model capable of identifying wildlife animals unseen in training phase. A problem with such general methods is that they can underperform when compared to specialized approaches, as each species relies on distinct visual cues.

Species-specific Re-Identification. Many studies focus on a single species under constrained settings (controlled environments, fixed viewpoints, or focusing on specific parts like the animal's face) [2, 20, 21]. Although these methods generally achieve good performances in those scenarios, they may struggle when applied in the wild.

For full-body animal Re-ID in uncontrolled settings, Li *et al*. [15] proposed the ATRW (Amur tiger Re-identification in the Wild) dataset. Many solutions have been proposed to the re-identification of amur tigers [17, 38, 39], Liu *et al*. [16] uses ground-truth pose keypoints to extract part features, Yu *et al*. [38] focused on enhancing the performances by computing the shortest path between corresponding local parts. Zhang *et al*. [42] created the YakReID-103, benchmarking both animal-specific methods (e.g. PGCFL [17]) and person Re-ID models (e.g. PCB [28]). They propose RERP [14], which boosts performances via random erasure and region-visibility prediction. These methods highlight the effectiveness of part-aware models, which sometimes require costly manual annotations.

PAW-ViT benefits from part-based feature learning without ground-truth pose labels. Beyond introducing learnable part tokens to the ViT, we use an off-the-shelf pose estimator and a segmentation ensemble to generate pseudo semantic segmentation masks. Through knowledge distillation [5, 11], each part token's attention over image patches is reconstructed into an attention map that serves as its predicted mask. This transfers spatial knowledge to the model and reduces reliance on spurious background cues [19, 22, 31, 36].

3. Method

In this section, we detail our proposed PAW-ViT (Part-AWare animal re-identification Vision Transformer). We begin by describing the PAW-ViT overall architecture,

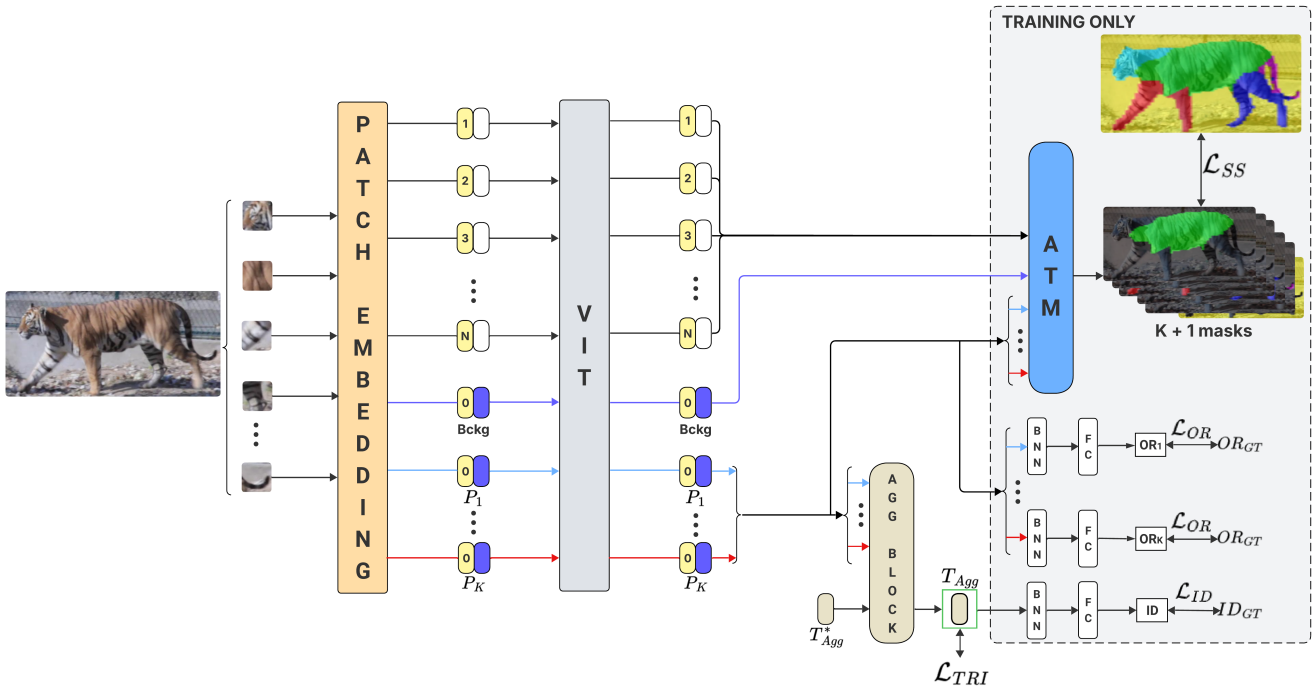

Figure 3. PAW-ViT overview. We prepend K part tokens plus one background token to the ViT. An Attention-to-Mask (ATM) decoder converts each token's attention over image tokens into a mask, supervised via semantic segmentation distillation. An aggregation block pools the K part features (background excluded) into a single descriptor trained with triplet and ID losses. Part tokens are supervised with individual orientation losses. At inference, only T_{Agg} is used, all modules inside the gray box are removed.

which adapts the vision transformer to learn from distinct animal parts. We then explain how we apply knowledge distillation from semantic segmentation masks to transfer spatial knowledge to PAW-ViT, and introduce an aggregation block that fuses the part-specific features into a single robust Re-ID descriptor. We conclude by discussing the loss functions used in our multi-task learning framework. The complete architecture is illustrated in Fig. 3.

3.1. PAW-ViT

PAW-ViT is an adaptation of the ViT, given an input image $I \in \mathbb{R}^{3\times H\times W}$, I is first divided into a sequence of flattened patches. A patch-embedding layer (here, a single 2D convolution layer with stride d and C output channels) projects these patches into $N = \frac{H}{d} \times \frac{W}{d}$ patch tokens, $T_{Patches} \in \mathbb{R}^{N\times C}$, where $d = 16$ and $C = 768$ for the default ViT-Base.

The core architectural change in PAW-ViT is the replacement of the single `[CLS]` token by K learnable *part tokens*, each corresponding to a predefined anatomical part of the animal, and one *background* token, resulting in the set of tokens $T \in \mathbb{R}^{(N+K+1)\times C}$. The tokens are added to positional embeddings and processed by a series of standard transformer blocks (Multi-Head Self-Attention and Feed-Forward Network). We use ViT weights pretrained on ImageNet[1], but the additional tokens are initialized randomly.

3.2. Distilling masks with the ATM decoder

To encourage each part token to attend a specific anatomical region, we supervise it with pseudo semantic segmentation masks given by a frozen teacher (an ensemble of segmentation and animal pose estimation models). We realize this spatial supervision with an *Attention-to-Mask (ATM) decoder*, a lightweight decoder that uses the spatial information in transformer's attention maps to generate mask predictions.

ATM was originally introduced for SegViT [41], a vision transformer for semantic segmentation. We adapt ATM to PAW-ViT with some major changes: we apply a single ATM block on the final Transformer layer rather than at multiple layers, we keep only the cross-attention that measures similarity between part tokens and image tokens, and we aggregate the similarity of multiple heads by calculating their average. Fig. 4 illustrates the block.

Let $T_{\text{Parts}} \in \mathbb{R}^{(K+1)\times C}$ be the K part tokens plus one background token, and $T_{\text{Patches}} \in \mathbb{R}^{N\times C}$ the image patches tokens. We obtain queries and keys via linear projections through Feed-Forward Neural Networks (FFN):

[1] https://github.com/huggingface/pytorch-image-models

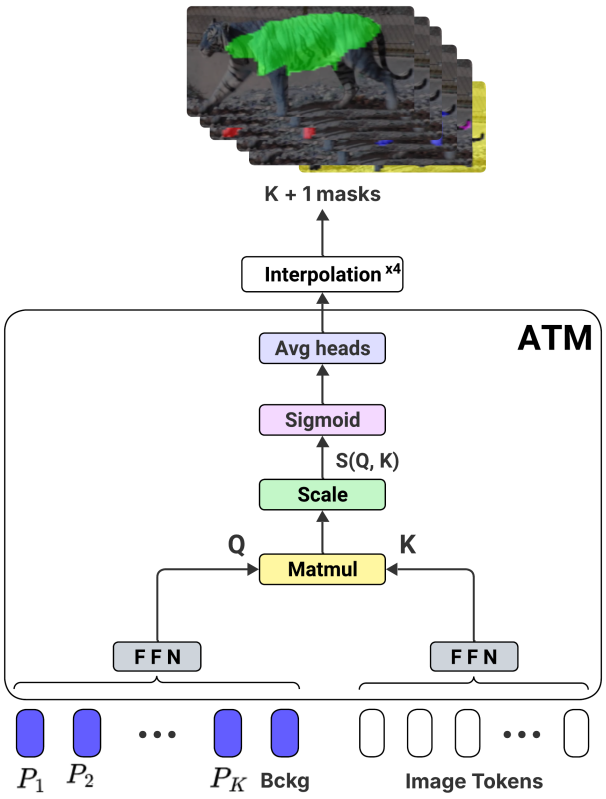

Figure 4. Illustration of the adapted ATM decoder.

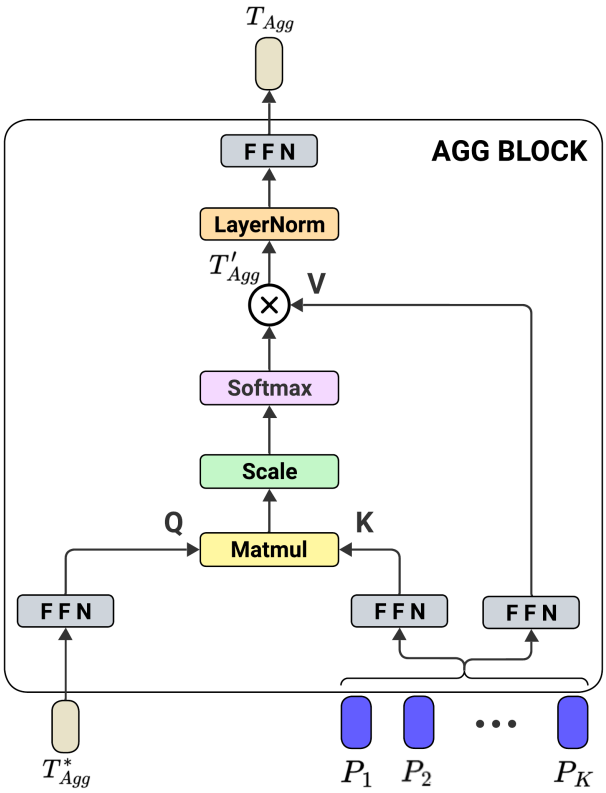

Figure 5. Illustration of the aggregation block.

$$Q = W^Q T_{\text{Parts}},$$
$$K = W^K T_{\text{Patches}},$$

where W^Q and W^K are the weight matrices of the FFNs. We then calculate the scaled dot-product that gives us the similarity between the parts and the image tokens:

$$S(Q, K) = \frac{QK^\top}{\sqrt{C/H}} \in \mathbb{R}^{H\times(K+1)\times N},$$

where H is the number of attention heads fixed to 12 in standard ViT-base. The ATM decoder converts these similarities into masks M_S by applying a sigmoid. The softmax, commonly used in attention mechanism, is not used here because it normalizes scores to sum to 1, which can alleviate strong similarities. The sigmoid treats each location independently, preserving high-attention regions. Finally, to encourage all cross-attention heads to focus on the same region, we average the per-head masks.

Given the teacher masks $M_T \in \mathbb{R}^{(K+1)\times\frac{H}{f}\times\frac{W}{f}}$, we reshape the averaged maks $M_S \in \mathbb{R}^{(K+1)\times N}$ to a spatial grid $(K+1)\times\frac{H}{d}\times\frac{W}{d}$, and then upsample by a factor $r = \frac{d}{f}$ using bilinear interpolation to obtain the final mask M_S. By default, we set $f = 4$, so for $d = 16$ the upsampling factor is $r = 4$.

As shown in Fig. 3, the ATM decoder is not present at inference.

3.3. Aggregation block

Two problems may arise when decoupling the animal re-identification into parts. First, to match two different individuals, we compare their feature vectors, if each individual is now represented by K part tokens, we slow down our inference by K times. Furthermore, not all parts are equally visible, and some parts are more discriminative than others. For example, for amur tigers the most important part is the body, from where we extract the stripes patterns that allow us to distinguish different individuals. To address this, we introduce a lightweight aggregation block that learns to fuse the part tokens into a single optimized descriptor. The aggregation block is described in Fig. 5.

This block introduces a new learnable aggregation token $T^*_{Agg} \in \mathbb{R}^C$, and employs a simple cross-attention mechanism. The aggregation token provides the query via a learned linear projection, while the set of K part tokens provides the keys and values:

$$Q = W^Q T^*_{Agg},$$
$$K = W^K T_{Parts},$$
$$V = W^V T_{Parts}.$$

The attention weights are computed and used to infer the final descriptor:

$$\alpha_i = \text{softmax}\Big(\frac{Q \cdot K_i}{\sqrt{C/H}}\Big),$$

$$T'_{Agg} = \sum_{i=1}^{K} \alpha_i \, V_i,$$

which is then processed by a LayerNorm and a FFN, to form the final Re-ID descriptor:

$$T_{Agg} = \text{FFN}\big(\text{LayerNorm}(T'_{agg})\big).$$

Differently from the classic cross-attention, we do not apply residual connection. At inference, only T_{Agg} is used as the final Re-ID descriptor, ensuring matching speed is independent of the number of parts K, and capturing which parts are more descriptive. As with the part tokens, T^*_{Agg} is randomly initialized. Notice that, in order to ignore background cues, the background token is not an input of this aggregation block.

3.4. Loss functions

The overall training objective is a weighted sum of four distinct loss functions.

The primary re-identification losses are aplied to the aggregation token T_{AGG}. We apply an identity loss, which is a cross-entropy loss:

$$\mathcal{L}_{ID} = -\sum_{c=1}^{C} y^c_{id} \log\big(\hat{y}^c_{id}\big), \tag{1}$$

where y^c_{id} is the ground-truth identity and C is the number of training identities. T_{Agg} is first passed through a BNNeck layer [18] before the fully connected classifier (see Fig. 3).

To learn a discriminative embedding space, we also apply metric learning to the aggregation token via triplet loss [26]. The loss is defined as:

$$\mathcal{L}_{\text{TRI}} = max(0, d(T_{Agg}, T^+_{Agg}) - d(T_{Agg}, T^-_{Agg}) + m), \tag{2}$$

where d denotes a distance function, m is a scalar defining the margin, T_{Agg} is a given *anchor* feature, T^+_{Agg} is a *positive* sample of the same identity as the anchor, and T^-_{Agg} is a *negative* sample of a different identity. The triplet loss pulls positives closer to the anchor and pushes negatives apart for at least the given margin m. We use the online batch hard triplet mining [10], which involves, to each anchor, selecting the positive sample with the highest distance and the negative sample with the lowest distance at each epoch of training.

Following prior work [14, 17, 39], we use a binary cross-entropy loss for orientation (e.g. left or right). However, in contrast to these methods, we apply this loss to each part token individually rather than to a global feature, improving viewpoint robustness.

$$L_{OR} = -\frac{1}{|\mathcal{V}|}\sum_{i=1}^{|\mathcal{V}|}\Big[y_{or}\ \log(\hat{y}^i_{or}) + (1 - y_{or})\ \log(1 - \hat{y}^i_{or})\Big], \tag{3}$$

where $\mathcal{V}$ is the set of visible parts $\mathcal{V} = \{\, i \mid \exists\, x, y : M^{(i)}_T(x, y) = 1\}$, $y_{\text{or}} \in \{0, 1\}$ is the ground-truth orientation label for token P_i, and $\hat{y}^i_{or}$ the predicted orientation from token i. We also apply a BNNeck before passing the tokens through the classifier. We test our method on datasets that provide y_{or}, but notice that on datasets without such labels, our pseudo semantic segmentation masks make it possible to deduce the orientation.

The distillation loss $\mathcal{L}_{\mathcal{SS}}(M_T, M_S)$ is a combination of the exponential logarithmic Dice loss (L_{Dice}) and the exponential cross-entropy (L_{Cross}) proposed by Wong *et al.* [34], that is denoted by:

$$\mathcal{L}_{\mathcal{SS}} = L_{Exp} = w_{Dice} L_{Dice} + w_{Cross} L_{Cross}. \tag{4}$$

Given that:

$$L_{\text{Dice}} = \mathbb{E}[(-\ln(\text{Dice}_i))^{\gamma_{\text{Dice}}}],$$

$$L_{\text{Cross}} = \mathbb{E}[w_l\,(-\ln p_l(\mathbf{x}))^{\gamma_{\text{Cross}}}],$$

with

$$\text{Dice}_i = \frac{2\sum_{\mathbf{x}}(\delta_{i\,l(\mathbf{x})}\, p_i(\mathbf{x})) + \epsilon}{\sum_{\mathbf{x}}(\delta_{i\,l(\mathbf{x})} + \sum_{\mathbf{x}} p_i(\mathbf{x})) + \epsilon},$$

where $\mathbf{x}$ represents the pixel, i the labels (including background), l the ground-truth label, $\mathbb{E}[\cdot]$ the mean value, $p_i(\mathbf{x})$ the softmax probability for class i, δ_{il} is the Kronecker delta that is 1 when $l(x) = i$, ϵ is a small constant, and $w_l = \big(\sum_k f_k / f_l\big)^{1/2}$ weights classes by inverse frequency. We use this loss because it is adapted to handle objects of unbalanced sizes. The animal parts sizes are highly unbalanced, for example, the body of animals appears far more often and is much larger than the tail.

The full multi-task training loss is defined as:

$$\mathcal{L} = \lambda \mathcal{L}_{SS} \ + \ \alpha \mathcal{L}_{TRI} \ + \ \mathcal{L}_{ORI} \ + \ \mathcal{L}_{ID} \tag{5}$$

Here, λ and α are constants that we applied to the distillation and to the triplet loss. More details about the chosen hyper-parameters will be discussed in section 4.2.

4. Experiments and Results

We begin by describing the datasets used to evaluate our method, then detail the hyperparameters used in our experiments, and benchmark PAW-ViT against SOTA animal Re-ID methods, discussing the results. Finally, we analyze the importance of each component of our multi-task loss.

4.1. Datasets

We evaluate PAW-ViT on two challenging animal re-identification datasets: the Amur Tiger Re-Identification in the Wild (ATRW) [15], and YakReid-103 [42]. Both datasets treat left-side and right-side views of the same animal as distinct identities (entity). For fair comparison, like previous works [17, 38, 39, 42], we augment the training data by horizontally flipping all images. The flipped images are then treated as separate individuals, doubling the size of the training set.

The ATRW dataset is a widely used benchmark, it contains in the training set 1,887 images, 107 entities, and 75 tigers. For our experiments, we follow Liu *et al*. [17] and use 1,824 images for training. The test set is divided in 701 images of 47 entities, and 42 tigers in the 'single-camera' subset, where the query tiger appears in only one camera, and 1061 images, 28 entities, and 20 tigers in the 'cross-camera' subset, where the tigers appear in multiple cameras. The cross-camera set represents a significant challenge due to larger variations in viewpoint, illumination, and background across different cameras.

The YakReid-103 training data consists of 1,404 images of 121 entities and 103 yaks. The testing set is divided into simple-testing and hard-testing. The simple-testing is composed of 843 images and 61 entities. The hard-testing is a subset of the simple set where images of the same individual with highly similar poses and backgrounds are manually excluded. It contains 433 images and 61 entities.

4.2. Implementation details

We use ViT-Base with 16×16 patch size and a stride of 16, input resolution of 256×256, and pseudo-masks of 64×64. We use models pre-trained on ImageNet, but the additional tokens and modules are initialized randomly. Input images were normalized using mean of [0.5, 0.5, 0.5] and standard deviation of [0.5, 0.5, 0.5]. We trained each model for 50 epochs using AdamW optimizer with a maximum learning rate of 8.5×10^{-5} and weight decay of 5×10^{-4}. We employ a one-cycle learning rate scheduler with cosine annealing, 3 epochs of linear warm-up, initial learning rate of 1.7×10^{-7} and minimum of 6.8×10^{-5}.

For data augmentation, we applied a random rotation between -15 and 15 degrees to both the input image and its corresponding mask. Following PK sampling [40], each mini-batch is composed of 128 images containing 8 individuals and 16 images per individual.

For the triplet loss, we employed the cosine distance as distance metric with a margin $m = 0.4$. We set the triplet loss constant to $\alpha = 0.5$. The weighting constant for the semantic segmentation loss is set to $\lambda = 1$ for ATRW and $\lambda = 0.5$ for the YakReID-103. For the semantic segmentation loss, we set $\gamma_{Cross} = \gamma_{Dice} = 1.2$, $w_{Cross} = 0.5$ and $w_{Dice} = 0.5$. Furthermore, we apply label smoothing [30] with smoothing of 0.2 to the cross-entropy loss to mitigate overconfident predictions. The hyper-parameters are defined empirically.

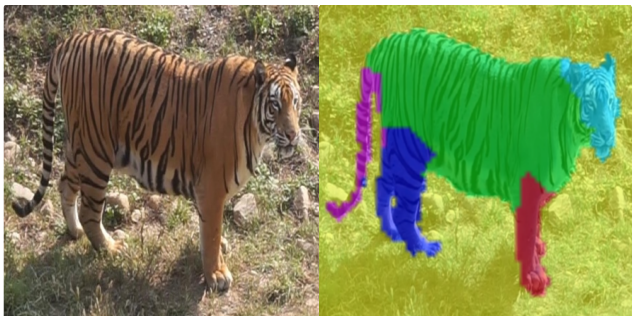

Figure 6. Example of amur tiger pseudo-masks using $K = 5$.

Model performance was evaluated every 10 epochs using standard Re-ID metrics such as the *rank-k* and *mean Average Precision* (*mAP*) [6, 43]. We select the best model based on the highest mAP. For both datasets, PAW-ViT achieves its best performances at epoch 40.

Our experiments are run on a Nvidia A100.

4.3. Results

To ensure a fair comparison, we report the results as published in the original papers of the respective methods. In addition, we compare against a baseline, the pre-trained ViT-B trained under the same settings as PAW-ViT. Since most SOTA approaches rely on SE-ResNet50 [12], we also train a SE-ResNet50 pre-trained on ImageNet to demonstrate that the two backbones have comparable performances.

4.3.1. ATRW Dataset

For the ATRW dataset, we define $K = 5$ semantic parts: head, body, tail, hind legs, and front legs (see Fig. 6). For fair comparison, we follow prior works on the ATRW and conduct evaluation by concatenating the features extracted from the original and horizontally flipped images.

Table 1 presents the results. Here, mmAP is the mean between single and cross-camera mAP. Treating PPGNet as an upper bound, as it uses ground-truth pose and higher resolution, our method reaches the best mmAP. In the harder cross-camera setting, it improves mAP by almost 1% over the strongest SOTA method PGCFL. Note that the SE-ResNet50 baseline outperforms ViT-B, yet PAW-ViT surpasses SOTA methods that rely on SE-ResNet50, which makes this result particularly compelling as our approach can be adapted to more performant transformer backbones.

PAW-ViT substantially outperforms the ViT-B baseline, raising the single-camera mAP by 3.8% and the cross-camera by 3.2%. The single-camera rank-1 is improved by 2.3% and the cross-camera of 1.5%. Figure 7 shows one

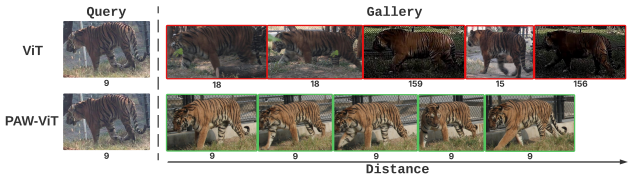

Figure 7. Top-5 retrieval comparison between ViT and PAW-ViT. Below each image, the corresponding ID of the tiger.

Figure 8. Example of yak pseudo-masks using $K = 4$.

retrieval ranking example for a query tiger image of ID 9. PAW-ViT ranks the correct identity within the top-5 nearest gallery images, whereas the ViT baseline returns 4 incorrect identities.

4.3.2. YakReID-103 Dataset

For the YakReID-103 dataset, we define $K = 4$ parts: head, body, hind legs, and front legs. Since the tail of the yaks is almost attached to the body, we do not represent it as a separate part. Fig. 8 shows one example of pseudo-masks.

Table 2 reports our final results on YakReID-103 alongside AER [39], and SOTA baselines as reported by Zhang *et al*. [42]. PAW-ViT significantly outperforms all prior methods in the hard-testing split, our base model exceeds PCN-RERP by 4.6% of mAP. Notice that the baseline already surpasses existing methods, suggesting that our training setup itself may constitute an additional contribution to this dataset. Relative to the baseline, PAW-ViT reaches 1.5% better mAP. Figure 9 shows a hard-split retrieval example, the ViT baseline finds the correct identity at rank-1 but matches different identities within the other images of the top-4, whereas PAW-ViT returns correct matches among the top-5 rank.

In the simple-testing split, PAW-ViT achieves 79.1% mAP, 2.8% better than PCN-RERP, increasing the gap over the baseline to 1.9%.

4.4. Ablation Study

We evaluate the contribution of each multi-task loss component (Section 3.4) using PAW-ViT 256×256 model on the ATRW dataset. Table 3 reports single- and cross-camera

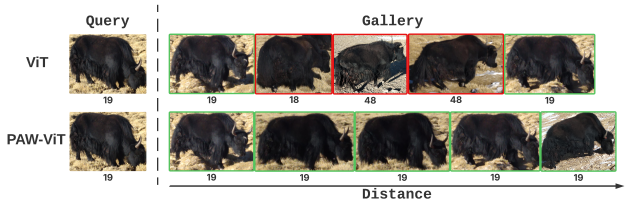

Figure 9. Top-5 retrieval comparison between ViT and PAW-ViT. Below each image, the corresponding ID of the yak.

mAP and Rank-1.

First, training only with cross-entropy loss results in performances that are very similar to the baseline, showing that merely adding extra tokens brings no consistent benefit. Adding triplet loss yields a clear gain of 3.4% single-camera mAP and 1.0% cross-camera, demonstrating the importance of metric learning. With the distillation loss, single-camera metrics improve slightly, while cross-camera mAP rises by 0.9% and Rank-1 by 1.5%, indicating that the semantic segmentation loss helps PAW-ViT to learn more robust, view-invariant features, which is critical for real-world animal Re-ID. Finally, the orientation loss on individual tokens further boosts performances, especially the single-camera that sees a gain of 1.3% of mAP. The orientation loss is especially important on the two tested datasets, where left/right views are treated as distinct individuals, helping the model to differentiate left and right features. On datasets where identity is viewpoint-invariant, its benefit may be limited.

5. Conclusion

We introduced PAW-ViT, a part-based Vision Transformer for efficient animal re-identification. By replacing the standard `[CLS]` token with K learnable part tokens and an aggregation token, PAW-ViT combines fine-grained, part-specific embeddings into a single descriptor, both reducing computational costs at inference time and automatically weighting the most discriminative regions. During training, each part token is guided to specialize on a specific part via a multi-task loss that includes Re-ID, orientation, and semantic segmentation objectives. Using pseudo masks generated from off-the-shelf segmentation and pose models, the spatial knowledge is distilled through an attention-to-mask decoder, that is used only during training, and discarded at inference. Our approach stands out on complex scenarios, closer to real-world applications, achieving SOTA mAP on the amur tiger (ATRW) and YakReID-103 datasets, and significantly improving the performances over the ViT-Base baseline.

PAW-ViT's main limitation is its reliance on teacher-generated masks. Some animals assume uncommon poses, which degrade the pseudo annotations causing poor performances. Furthermore, the number of parts, K, is currently

Method	Backbone	Resolution	mmAP	Single-Camera			Cross-Camera		
				mAP	R-1	R-5	mAP	R-1	R-5
PPbM-a [15]	ResNet50	256×128	62.9	74.1	88.2	96.4	51.7	76.8	91
PPbM-b [15]	ResNet50	256×128	60.3	72.8	89.4	95.6	47.8	77.1	90.7
NWPU-ASGO [38]	DenseNet	256×256	75.1	87.9	**96.9**	98.3	62.2	**92.5**	95.1
AER [39]	SE-ResNet50	224×224	76.2	–	95.7	–	–	88.0	–
PGCFL [17]	SE-ResNet50	224×448	77.0	**89.8**	96.6	97.7	64.3	91.3	**95.8**
Baseline	SE-Resnet50	256×256	74.2	84	95.9	**98.6**	64.4	86.7	94.2
Baseline	ViT-Base	256×256	73.5	85.2	94.4	97.1	61.9	89.1	94.9
PAW-ViT (ours)	ViT-Base	256×256	**77.1**	89	96.7	**98.6**	**65.2**	90.6	95.2
PPGNet* [16]	SE-ResNet50	256×512	77.9	89.6	99.4	99.4	66.3	90.8	97.7

Table 1. Comparison of performance on the ATRW dataset.
* Uses ground-truth pose annotations.
— Indicates that the result was not provided by the authors.

Method	Backbone	Resolution	Simple-testing			Hard-testing		
			mAP	R-1	R-5	mAP	R-1	R-5
PGCFL [17]	SE-ResNet50	224×448	69.4	94.2	96.1	59	85.1	91.8
PCB [28]	ResNet50	384×128	70.9	94.7	97.1	64.8	89.9	94.2
AER [39]	SE-ResNet50	224×224	—	—	—	66.1	92.3	—
PCN-RERP [14]	SE-ResNet50	288×488	76.3	**97.6**	98.1	68.6	91.8	93.3
Baseline	SE-Resnet50	256×256	77.7	**97.6**	**98.3**	71.9	92.8	94.2
Baseline	ViT-Base	256×256	77.2	95.6	98.1	71.7	**93.7**	96.1
PAW-ViT (ours)	ViT-Base	256×256	**79.1**	96.1	96.8	**73.2**	92.8	**96.6**

Table 2. Comparison of performances on the YakReID-103 dataset.
— Indicates that the result was not provided by the authors.

$\mathcal{L}_{CE}$	$\mathcal{L}_{TRI}$	$\mathcal{L}_{SS}$	$\mathcal{L}_{OR}$	Single-Camera		Cross-Camera	
				mAP	R-1	mAP	R-1
✓				84	94.7	62.5	88.8
✓	✓			87.4	95.6	63.5	88.5
✓	✓	✓		87.7	96	64.4	90
✓	✓	✓	✓	**89**	**96.7**	**65.2**	**90.6**

Table 3. Ablation study of PAW-ViT on the ATRW dataset.

selected as a fixed hyperparameter prior to training. In future studies, we plan to explore more robust mask generation and develop a method to automatically determine the optimal K.

A second limitation is the joint optimization of distillation and re-ID from the very start, when part tokens are random and unspecialized. For now, we balance this with a single hard weight λ on the distillation loss. In future work, we should include a multi-task balancing mechanism. We already obtain strong results without careful multi-task weighting, suggesting room to explore rebalancing strategies that emphasize distillation early (while tokens specialize) and then shifts weight toward the Re-ID objective.

Finally, hyper-parameters related to the distillation loss could be explored, like γ_{Cross}, γ_{Dice}, w_{Dice}, and w_{Cross}, used in the distillation loss. We can also explore varying the number of parts K on different species.

We believe PAW-ViT represents a significant step toward accurate, explainable, and annotation-free animal re-identification, with clear applications in wildlife monitoring and conservation.

6. Declarations

Data availability. The datasets used in this study are third-party. The ATRW dataset is publicly available[2], and the YakReID-103 dataset can be obtained upon request to the authors.

Code availability. The code implementation of PAW-ViT, along with the scripts used to generate the seman-

[2] https://cvwc2019.github.io/challenge.html

tic segmentation pseudo-masks, are available on github: https://github.com/eugeniodias5/PAW-ViT.

Acknowledgments. This project was provided with computing AI and storage resources by GENCI at IDRIS thanks to the grant 2025-AD011015410R1 on the supercomputer Jean Zay's A100 partition.

References

[1] Shuoyi Chen, Mang Ye, and Bo Du. Rotation invariant transformer for recognizing object in uavs. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2565–2574, 2022. 2

[2] Xiaolang Chen, Tianlong Yang, Kaizhan Mai, Caixing Liu, Juntao Xiong, Yingjie Kuang, and Yuefang Gao. Holstein cattle face re-identification unifying global and part feature deep network with attention mechanism. *Animals*, 12(8): 1047, 2022. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021. 2

[6] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill., Oct.*, 2007. 6

[7] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, pages 1–7, 2007. 1

[8] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 2

[9] Zhimin He, Jiangbo Qian, Diqun Yan, Chong Wang, and Yu Xin. Animal re-identification algorithm for posture diversity. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 1

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2, 5

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[13] Bingliang Jiao, Lingqiao Liu, Liying Gao, Ruiqi Wu, Guosheng Lin, PENG WANG, and Yanning Zhang. Toward re-identifying any animal. In *Advances in Neural Information Processing Systems*, pages 40042–40053. Curran Associates, Inc., 2023. 1, 2

[14] Lei Li, Tingting Zhang, Da Cuo, Qijun Zhao, Liyuan Zhou, and Suonan Jiancuo. Automatic identification of individual yaks in in-the-wild images using part-based convolutional networks with self-supervised learning. *Expert Systems with Applications*, 216:119431, 2023. 1, 2, 5, 8

[15] Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. Atrw: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2590–2598, New York, NY, USA, 2020. Association for Computing Machinery. 1, 2, 6, 8

[16] Cen Liu, Rong Zhang, and Lijun Guo. Part-pose guided amur tiger re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2, 8

[17] Ning Liu, Qijun Zhao, Nan Zhang, Xinhua Cheng, and Jianing Zhu. Pose-guided complementary features learning for amur tiger re-identification. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 286–293, 2019. 2, 5, 6, 8

[18] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019. 5

[19] Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023. 2

[20] Thierry Moreira, Mauricio Perez, Rafael Werneck, and Eduardo Valle. Where is my puppy? retrieving lost dogs by facial features. *Multimedia Tools and Applications*, 76, 2017. 2

[21] Guillaume Mougeot, Dewei Li, and Shuai Jia. A deep learning approach for dog face verification and recognition. In *PRICAI 2019: Trends in Artificial Intelligence*, pages 418–430, Cham, 2019. Springer International Publishing. 2

[22] Eugênio Dias Ribeiro Neto, Cyril Barrelet, Marc Chaumont, Gérard Subsol, Muhammad Nur Faiz Mahfudz, Muhammad Najib Arung Petana Raja Bone, Barandi Sapta Widartono, Dyah Ayu Widiasih, Mia Nur Farida, Wayan Tunas Artama, Thibaut Langlois, Hélène Guis, Etienne Loire, and Michel de Garine-Wichatitsky. Background-invariant re-identification of dogs from camera-trap videos in non-controlled environments. *Ecological Informatics*, page 103547, 2025. 1, 2

[23] Enhao Ning, Changshuo Wang, Huang Zhang, Xin Ning, and Prayag Tiwari. Occluded person re-identification with deep learning: a survey and perspectives. *Expert systems with applications*, 239:122419, 2024. 2

[24] Prashanth C Ravoor and TSB Sudarshan. Deep learning methods for multi-species animal re-identification and

tracking–a survey. *Computer Science Review*, 38:100289, 2020. 2

[25] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 2

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 5

[27] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *European Conference on Computer Vision*, pages 216–233. Springer, 2024. 2

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin 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. 2, 8

[29] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020. 2

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[31] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5794–5803, 2018. 2

[32] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2540–2549, 2022. 2

[33] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2

[34] Ken CL Wong, Mehdi Moradi, Hui Tang, and Tanveer Syeda-Mahmood. 3d segmentation with exponential logarithmic loss for highly unbalanced object sizes. In *International conference on medical image computing and computer-assisted intervention*, pages 612–619. Springer, 2018. 5

[35] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019. 1

[36] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition, 2020. 2

[37] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, pages 1–31, 2024. 1

[38] Jiwen Yu, Haibo Su, Junnan Liu, Zhizheng Yang, Zhouyangzi Zhang, Yixin Zhu, Lu Yang, and Bingliang Jiao. A strong baseline for tiger re-id and its bag of tricks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 302–309, 2019. 2, 6, 8

[39] Yingxue Yu, Vidit Vidit, Andrey Davydov, Martin Engilberge, and Pascal Fua. Addressing the elephant in the room: Robust animal re-identification with unsupervised part-based feature alignment. *arXiv preprint arXiv:2405.13781*, 2024. 2, 5, 6, 7, 8

[40] Yao Zhai, Xun Guo, Yan Lu, and Houqiang Li. In defense of the classification loss for person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1526–1535, 2019. 6

[41] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al. Segvit: Semantic segmentation with plain vision transformers. *Advances in Neural Information Processing Systems*, 35:4971–4982, 2022. 3

[42] Tingting Zhang, Qijun Zhao, Cuo Da, Liyuan Zhou, Lei Li, and Suonan Jiancuo. Yakreid-103: A benchmark for yak re-identification. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2021. 2, 6, 7

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 2, 6

[44] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1):1–20, 2017. 2

[45] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4692–4702, 2022. 2

[46] Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5941–5951, 2024. 1, 2