IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 June 2024, accepted 24 July 2024, date of publication 1 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3436813

RESEARCH ARTICLE

ReDeformTR: Wildlife Re-Identification Based on Light-Weight Deformable Transformer With Multi-Image Feature Fusion

ZITONG LI, ZHENGMAO YAN, WEIHONG TIAN, DETIAN ZENG, YI LIU, AND WEIMIN LI
School of Information, Hunan University of Humanities, Science and Technology, Loudi 417000, China

Corresponding author: Weimin Li (weiminli@huhst.edu.cn)

This work was supported in part by Hunan Provincial Natural Science Regional Joint Fund under Grant 2024JJ7237, in part by the Key Scientific Research Project of Hunan Provincial Department of Education under Grant 22A0617, in part by Hunan Natural Science Foundation-Regional Joint Fund under Grant 2023JJ50490, in part by Hunan Provincial Department of Education Scientific Research Project under Grant 22C0596, and in part by Hunan Provincial Innovation Foundation for Postgraduate under Grant CX20231276.

ABSTRACT Wildlife re-identification (Re-ID) techniques are key to animal tracking and preservation. However, the performance of current deep learning methods is unsatisfactory in cross-camera scenarios, especially in terms of mean average precision (mAP). This work introduces ReDeformTR, a novel model designed for wildlife Re-ID tasks, particularly focusing on the identification of individual animals' images captured by different cameras. ReDeformTR integrates a lightweight deformable transformer architecture capable of multi-image feature fusion, which can extract and fuse features from multiple images and scales, facilitating efficient representation of individual animals and enhancing performance during queries. A convolutional neural network (CNN) backbone is adopted for feature extraction, while a deformable transformer is used for feature refinement and fusion. The deformable attention mechanism reduces computation overhead by selectively sampling features, thereby enhancing efficiency. The experiments show that ReDeformTR demonstrates superior performance in terms of mAP on a cross-camera wildlife dataset in ATRW. The mAP is 84.98%, which represents a significant improvement of 12.29% compared to the state-of-the-art model (PPGNet). Furthermore, our model achieves a significant reduction in model parameter size, positioning it as a promising solution for wildlife Re-ID tasks.

INDEX TERMS Animal re-identification, deformable transformer, multi-image, wildlife re-identification.

I. INTRODUCTION

Re-identification (Re-ID) is a common task for object retrieval [1]. The main goal of this task is to identify the queried object or instance in images sampled by different visual acquisition devices or the same device at different times [2], [3]. Generally, the query is represented by an image, while other forms like video or text are also optional. With the rapid increase in demand for public safety and smart management, Re-ID technology is widely used in different aspects, especially for the Re-ID of persons, vehicles, and animals [2], [4], [5], [6]. Nonetheless, there are great challenges in this task due to varying viewpoints and image characteristics, like hue, saturation, and lightness [2], [6], [7], [8], which leads to variation and ambiguity when re-identifying.

To achieve better performance, various models have been proposed, many of which are built with CNNs only. Techniques in contrastive learning and representation learning are introduced, like pseudo labels [9], [10], contrastive mechanisms [11], and memory bank [12]. Meanwhile, since the attention mechanism and transformers [13] are widely applied in visual tasks, transformer-based Re-ID models have become common [15], [16], [19]. Some research focuses on the attention mechanism and tries to quantitatively analyze the attention [14]. Many efforts are also made to support

The associate editor coordinating the review of this manuscript and approving it for publication was Su Yan.

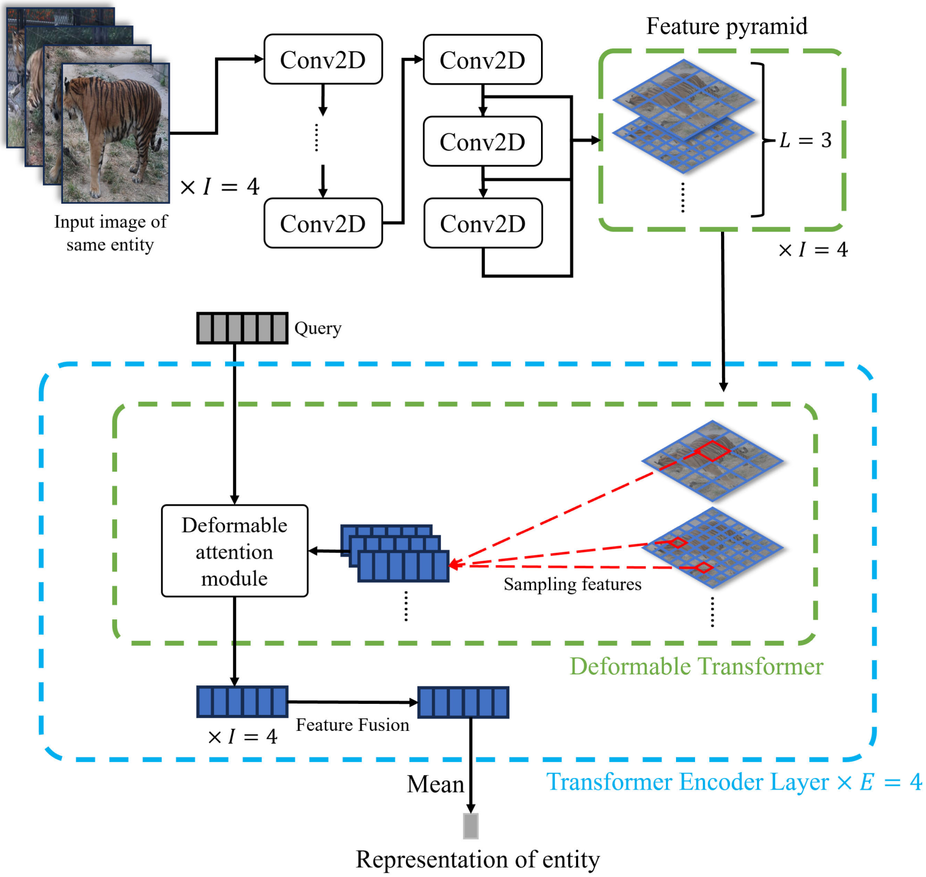


FIGURE 1. **ReDeformTR framework.**

self-supervised learning or unsupervised learning [16], [17], [18], [19]. However, most of the above works are based on human datasets, and the parameter size of models has become larger, especially in recent years.

Notably, the DETR [20] and its variants achieve great success in object detection. Deformable DETR [21] provides a deformable attention module that only attends to a small set of key sampling points around a reference, enabling it to extract features on different scales. Moreover, different improvements have been proposed to solve problems related to small feature resolution, slow training convergence, and computational efficiency [22], [23], [24]. However, the deformable attention mechanism has not been applied to the Re-ID task yet.

Inspired by above models for object detection, especially the DETR series, this work proposes a new DETR-based model for the Re-ID task. Our model adopts the deformable attention mechanism in the encoder layers and can fuse features from different images with varying viewpoints or time shifts for better performance.

There are two parts to our model. The first part is a common CNN-based backbone for feature extraction. The second part is the deformable attention and transformer for feature refinement and fusion. The backbone outputs a feature pyramid consisting of the last 3 backbone layer outputs. The deformable transformer refines the features of different scales by selecting key features according to learnable positions and fuses features from different feature pyramids, functioning as an encoder. Since there is no decoder, the parameter size of the entire model is quite small. The loss function is an improved triplet loss, which regards the nearest feature of a non-identical instance as the negative and the farthest feature of an identical instance as the positive.

Our contributions can be summarized as follows:

1. This work designs a DETR-based model for the Re-ID task with several novel techniques or mechanisms, like deformable attention and transformer, and the improved triplet loss. Moreover, our model can extract and fuse features from multiple images, enabling it to represent instances with multiple images by a single representation vector and alleviating performance expenses when querying.

2. The model scale is significantly reduced compared to mainstream baselines. By focusing on selected features in the whole pyramid and using a simple backbone, parameter size and memory requirements are significantly lower, while the performance of our model remains at a top level.

3. We conduct an ablation study to analyze the effectiveness of different numbers of images for fusion in our model.

II. RELATED WORKS

A. RE-IDENTIFICATION AND SIMILAR TASKS

Object classification and detection are the most mainstream tasks in the visual deep learning field. With the increasing deployment of visual artificial intelligence, other visual tasks are also receiving more attention in recent years, including Re-ID, instance-level image retrieval, and fine-grained classification. These three tasks are similar to each other, and we will discuss them in this section.

1) RE-IDENTIFICATION

In the visual machine learning domain, most major studies on Re-ID focus on humans and vehicles. However, in recent years, various Re-ID demands have emerged, such as finding lost pets, livestock locating, and wildlife tracking. To satisfy these requirements with visual approaches, many deep learning methods designed for vehicles and human faces are applied to animals or other objects, and different new models are proposed.

The SOLIDER [10] model incorporates prior knowledge derived from human images to generate pseudo semantic labels, thereby enriching the learned representation with additional semantic information. MSINet [11] introduces a Twins Contrastive Mechanism within its architecture, offering a more appropriate framework for Re-ID architecture search. This mechanism operates within a Multi-Scale Interaction search space, enabling the identification of rational interaction operations between features across different scales. Furthermore, the utilization of attention mechanisms and transformer architectures has become increasingly prevalent in visual tasks, including Re-ID. Rao et al. [14] explore the impact of learned visual attention on network predictions through counterfactual intervention, seeking to optimize attention mechanisms for fine-grained image recognition. Additionally, the TransReID [15] model pioneers a transformer-based framework for object Re-ID, achieving competitive performance across several benchmark datasets traditionally dominated by CNN-based methods. In another approach, Li et al. [16] suggest integrating self-supervised representation learning to facilitate the discovery of geometric features within Re-ID tasks. This model introduces an interpretable attention module centered around local maxima aggregation, offering a transparent mechanism for understanding model behavior and producing physically reasonable response maps. Finally, an inter-instance contrastive encoding (ICE) method enhances class-level contrastive Re-ID techniques by leveraging inter-instance pairwise similarity scores to boost performance [19]. However, these methods consume significant memory and require high-performance devices for training and testing.

2) INSTANCE-LEVEL IMAGE RETRIEVAL

Instance-level image retrieval (IIR) is a similar task, aiming to retrieve images containing a certain instance from a gallery with millions or even billions of images. The query images may also be captured in different environments, such as varying viewing angles, distances, illuminations, and backgrounds [1]. The Reranking Transformers represent a lightweight model capable of integrating both local and global features to re-rank matching images in a supervised manner [25]. The adversarial instance-level image retrieval method employs a redesigned generator and discriminator utilizing a 1 × 1 one-layer convolutional network, resulting in significant improvements in retrieval accuracy without a corresponding increase in time costs [26]. Moreover, the deep-seated features histogram method is proficient in extracting low-level features by simulating human orientation selection and color perception mechanisms [27].

3) FINE-GRAINED RECOGNITION

Another similar task is fine-grained recognition. Humans can not only perform simple classification of different animals, like dogs or birds, but also identify specific species, such as pigeons or eagles. This kind of visual task is called fine-grained recognition [28]. TransFG [29] introduces an augmented transformer-based approach aimed at improving image patch selection and feature representation generation by integrating raw attention weights into an attention map. CLIP-Art [30], on the other hand, concentrates on recognizing fine-grained attributes of artworks by leveraging Contrastive Language-Image Pre-Training (CLIP) during training. Notably, its zero-shot capability enables the prediction of relevant natural language descriptions for images without direct optimization for the task. P2P-Net [31] is designed to incorporate discriminative features related to specific parts while promoting object representation discrimination through pose-insensitive feature regularization. Lastly, the Attention Pyramid Convolutional Neural Network is characterized by a pyramidal hierarchy structure comprising both top-down feature pathways and bottom-up attention pathways. As a result, it can learn representations encompassing both high-level semantic information and low-level detailed features [32].

B. DEFORMABLE DETR AND RELATED MODELS

The DETR model [20] achieves outstanding performance in traditional classification with transformer structures. Moreover, variants of the DETR approach achieve improvements in different aspects, such as faster training speed or lower memory costs. This series of models inspires us to apply related technologies in the Re-ID task. We will discuss the DETR-based model series in this section.

1) DETR

This end-to-end model adopts the transformer encoder-decoder after a CNN-based backbone, more specifically,

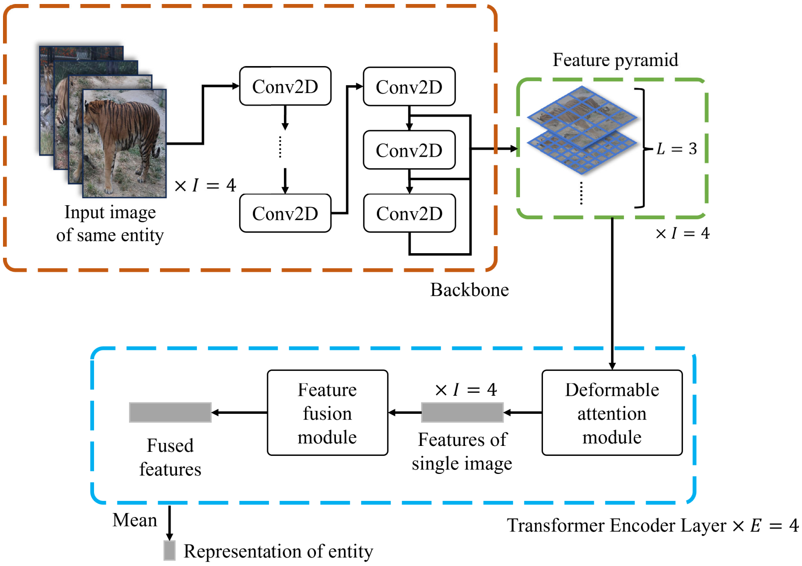


FIGURE 2. Model structure overview.

the ResNet. The backbone extracts feature maps from each image. After that, a 1 × 1 convolution is applied to reduce the channel of image features for less computation in subsequent parts, and fixed positional encoding is added. The feature map is then collapsed into one dimension as a sequence input for the transformer encoder. Each encoder layer has a standard architecture including the multi-head self-attention module and a feed-forward network (FFN). Furthermore, the transformer decoder also follows the standard architecture and decodes the N object queries at each decoder layer. Each object query is decoded into box coordinates and class labels by the prediction feed-forward network to indicate the class and boundary box of each object.

2) DEFORMABLE DETR

DETR is a novel work that introduces the transformer mechanism into the object detection task, achieving outstanding performance. Besides, it is suitable for many downstream tasks, like panoptic segmentation. However, the original DETR requires high memory and is not good at small object detection. Thus, the deformable DETR is proposed to solve those problems and speed up the convergence of training [36].

The framework of deformable DETR is similar to the original DETR; however, the transformer module is entirely different and replaced by the deformable transformer encoder-decoder [21]. The core of the deformable transformer encoder is the multi-scale deformable attention module. Unlike standard attention, deformable attention does not calculate the correlation of each feature from the backbone. Instead, several features are selected by location on each head, and then the selected features perform the attention process and are input into the FFN.

Multi-scale means the source of selected features should be from different levels of the feature map, which are the outputs of different blocks in the backbone network, benefiting from multi-level information. Moreover, since the attention part has limited input, the memory and computation costs are less than those of the original DETR [24], [36].

The encoder with multi-scale input produces multi-scale feature maps whose sizes are exactly the same as the input. After that, decoder decodes N object queries based on the multi-scale feature maps with cross-attention and self-attention modules. In the cross-attention modules, object queries extract features from the feature maps, where the key elements are the output feature maps from the encoder. In the self-attention modules, object queries interact with each other, where the key elements are the object queries.

Finally, N object queries are available after every decoder layer and fed to the FFN to predict the class and boundary box of each object. In this paper, we introduce the multi-scale deformable attention module into the Re-ID domain and show its effectiveness in terms of accuracy, mAP, parameter size, and computing overhead.

III. METHODOLOGY

A. MODEL OVERVIEW

This work is based on the deformable DETR and introduces a Re-ID model. The model is capable of extracting and fusing features from wildlife photos captured at different angles and with different cameras. It can then identify individual animals based on these features and retrieve corresponding unique IDs from the database. Additionally, the model does not impose limitations on the number of images for fusion as long as the hardware allows and can also accept input of a single image.

As shown in Figure 2, the Re-ID model consists of two main parts: the first part involves the backbone network

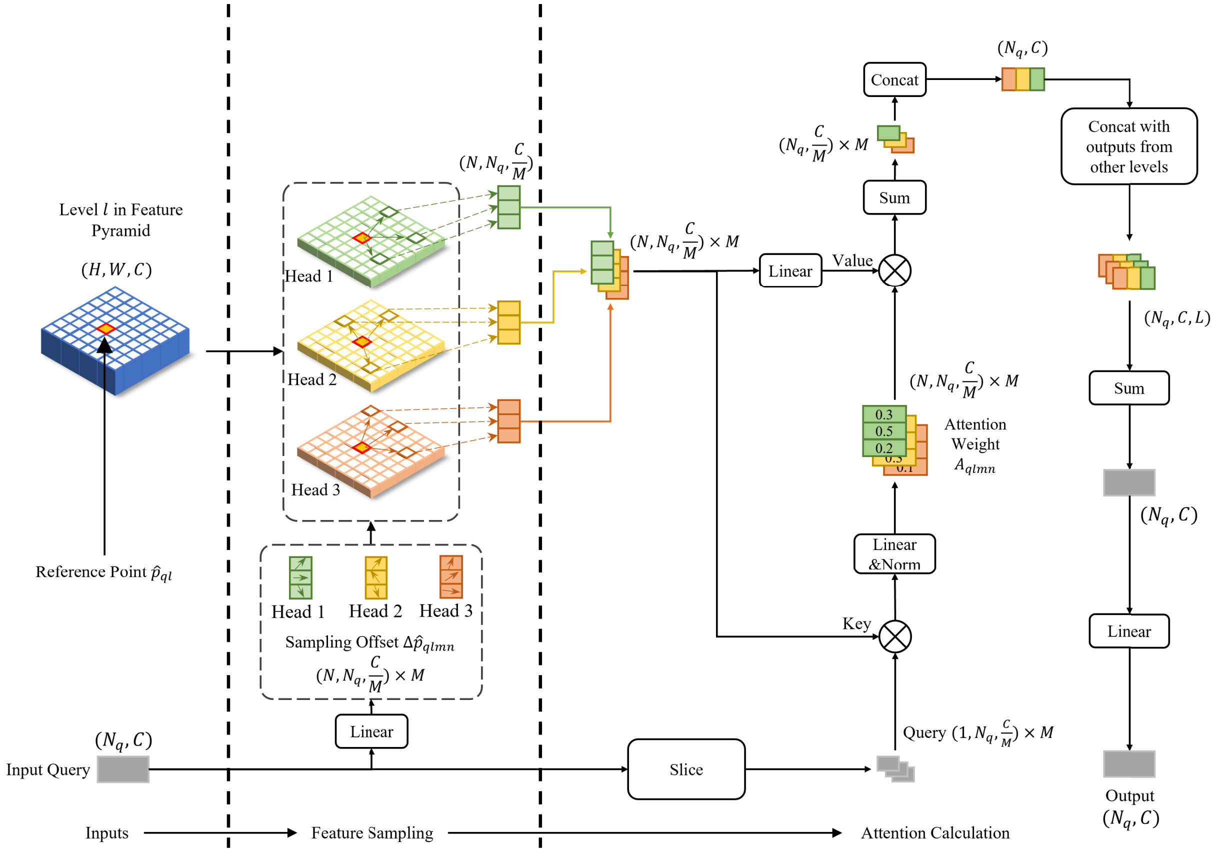

Level l in Feature Pyramid
(H, W, C)
Reference Point $\hat{p}_{ql}$
$(N, N_q, \frac{C}{M})$
Head 1
Head 2
Head 3
Head 1
Head 2
Head 3
Sampling Offset $\Delta\hat{p}_{qlmn}$
$(N, N_q, \frac{C}{M}) \times M$
$(N, N_q, \frac{C}{M}) \times M$
Linear
Value
Concat
(N_q, C)
$(N_q, \frac{C}{M}) \times M$
Sum
Concat with outputs from other levels
(N_q, C, L)
$(N, N_q, \frac{C}{M}) \times M$
0.3
0.5
0.2
Attention Weight A_{qlmn}
Linear &Norm
Key
Sum
(N_q, C)
Linear
(N_q, C)
Input Query
Linear
Slice
Query $(1, N_q, \frac{C}{M}) \times M$
Output (N_q, C)
Inputs
Feature Sampling
Attention Calculation

FIGURE 3. Deformable attention module in mean fusion method for single feature pyramid level.

for extracting the feature pyramid, and the second part, the deformable transformer, employs a deformable attention module and a feature fusion module from multiple images in each layer and consists of $E \in \mathbb{N}$ layers in total. Firstly, the backbone extracts feature pyramids for input images which contain the same animal. Secondly, the deformable transformer refines and fuses features from the feature pyramids, then outputs a feature sequence. Finally, averaging the feature sequence results in the representation of that animal.

B. BACKBONE AND FEATURE PYRAMID

To achieve a lightweight model, unlike common Re-ID models, our backbone is just a 7-level convolutional neural network (CNN). The output feature maps of the last $L \in \mathbb{N}$ levels will be selected to form a feature pyramid. Since feature maps from different levels are output, subsequent structures are enabled to extract representations of entities at different scales. The process is shown in the backbone part in Figure 2.

C. DEFORMABLE ATTENTION

Generally, the multi-head attention module in transformers calculates the attention of every input embedding, which can be represented by the following if the key is the same as the value:

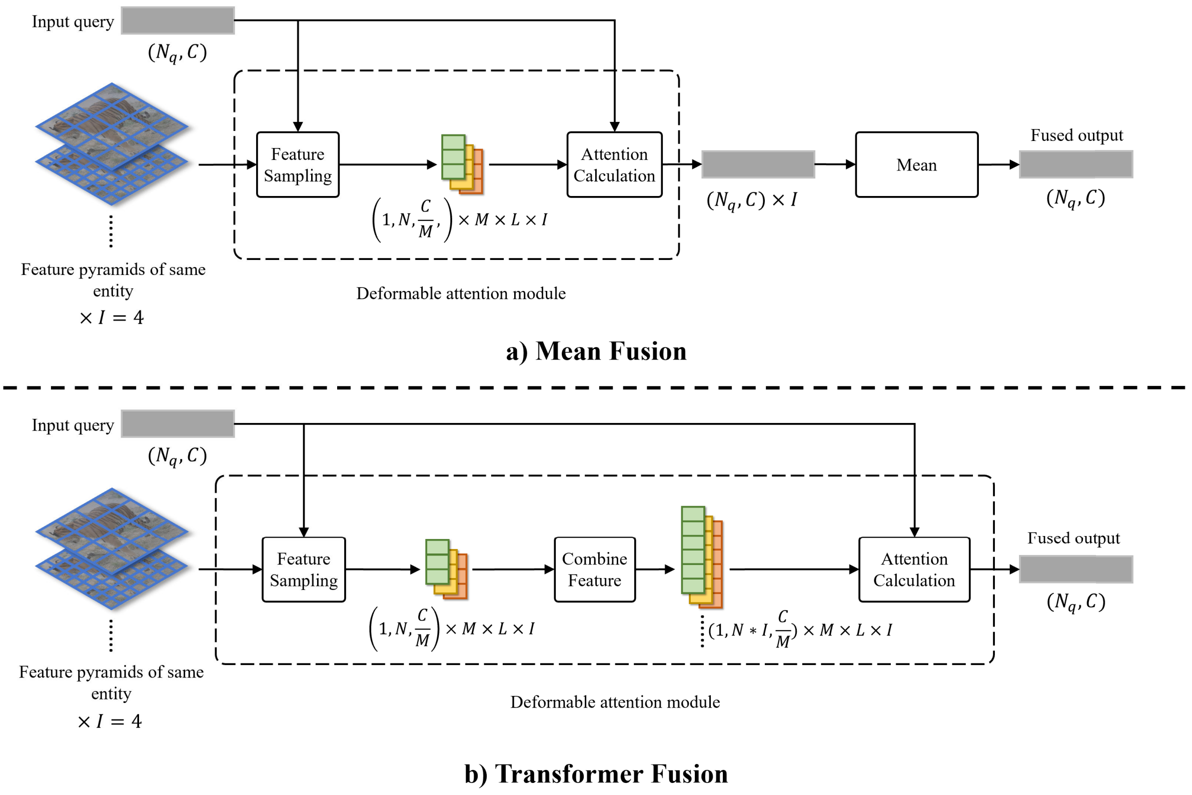
$$MultiHeadAttn(z_q, \boldsymbol{x}_k) = \sum_{m=1}^{M} \boldsymbol{W}_m^O \sum_k A_{qkm} \cdot \boldsymbol{W}_m^V \boldsymbol{x}_k \tag{1}$$

And $A_{qkm} \propto \exp[(z_q \boldsymbol{W}_m^Q)^T \boldsymbol{W}_m^K \boldsymbol{x}_k / \sqrt{d_k}]$. In this case, computation complexity is $O(N_q C^2 + N_k C^2 + N_q N_k C)$, where N_q and N_k are the number of query and key features, C is the channel number of features.

However, in the visual domain, query and key usually are the flattened feature map with size of $[H, W]$ extracted from the input image, which means $N_k = HW$. But the size of the feature map can be large for common backbone networks, which increases N_k and causes the computation complexity to become $O(HWC^2 + N_q HWC)$ for cross-attention.

Therefore, to decrease the cost of computation, inspired by the deformable convolution [37], the deformable attention module is proposed. This module will not handle every input feature; instead, only a few key points on the feature pyramid will be chosen. Then a series of features will be sampled according to key points and input to an attention module as shown in Figure 3.

The first step is to find those sampling point locations, which consist of reference position and sampling offset. The normalized reference position of each query z_q on each pyramid level is given as $\hat{\boldsymbol{p}}_{ql}$, and normalized sampling offset


Input query
(N_q, C)
Feature pyramids of same entity
$\times I = 4$
Feature Sampling
$\left(1, N, \frac{C}{M}, \right) \times M \times L \times I$
Attention Calculation
Deformable attention module
$(N_q, C) \times I$
Mean
Fused output
(N_q, C)

a) Mean Fusion

Input query
(N_q, C)
Feature pyramids of same entity
$\times I = 4$
Feature Sampling
$\left(1, N, \frac{C}{M}\right) \times M \times L \times I$
Combine Feature
$(1, N * I, \frac{C}{M}) \times M \times L \times I$
Attention Calculation
Deformable attention module
Fused output
(N_q, C)

b) Transformer Fusion

FIGURE 4. **Encoder layer structure and feature fusion methods.**

is generated by applying a linear transformation to each query, which is $\Delta\hat{\boldsymbol{p}}_{qlmn}$. For each attention head and each feature pyramid level, there are N points for sampling features and n indexes the sampling point. M is the number of attention heads and m indexes the attention head. L is the number of pyramid levels and l indexes the level of the feature pyramid. $\boldsymbol{x}$ is the feature pyramid with H_l height and W_l width at level l. Therefore, the sampled feature of each query can be written as:

$$\boldsymbol{x}(\hat{\boldsymbol{p}}_{ql} + \Delta\hat{\boldsymbol{p}}_{qlmn}) \tag{2}$$

Then, calculating the attention weight A_{qlmn} of each query and their sampled features and multiplying the sampled features with the attention weight gives:

$$A_{qlmn} = \boldsymbol{W}_m^A (z_q)^T \boldsymbol{x}(\hat{\boldsymbol{p}}_{ql} + \Delta\hat{\boldsymbol{p}}_{qlmn}) \tag{3}$$

Here, we merge $\boldsymbol{W}_m^Q$ and $\boldsymbol{W}_m^K$ into $\boldsymbol{W}_m^A$ to reduce the parameter count and operations. To avoid very large values in the process, a SoftMax operation can be applied to the attention weight before summing or normalizing the A_{qlmn} with other methods so that $\sum_{l=1}^{L} \sum_{n=1}^{N} A_{qlmn} = 1$. After that, we could aggregate output components in each level, each point, and each head, and get the final output of each query as:

$$\sum_{m=1}^{M} \boldsymbol{W}_m^O \sum_{l=1}^{L} \sum_{m=1}^{N} A_{qlmn} \cdot \boldsymbol{W}_m^V \boldsymbol{x}(\hat{\boldsymbol{p}}_{ql} + \Delta\hat{\boldsymbol{p}}_{qlmn}) \tag{4}$$

Compared with the common attention module, the deformable attention does not consider all input. Let N_q be the number of queries, and N can be quite small if the sampling points are limited. The complexity of sampling offset $\Delta\hat{\boldsymbol{p}}_{qlmn}$ and attention weight A_{qlmn} is $O(N_q LMNC)$ and the total complexity of deformable attention is $O(N_q C^2 + N_q NC^2 + N_q LNC + N_q LMNC)$. By default, $L = 3, M = 8, N = 4$ and $C = 256$. Obviously, $L(M + 1)N < C$ and the final complexity is $O(2N_q C^2 + N_q NC^2)$. While the complexity of the common cross-attention module is $O(N_q[\sum_l H_l W_l]C + [\sum_l H_l W_l]C^2)$ where $N_k = \sum_l H_l W_l$. Because $2N_q + N \ll \sum_l H_l W_l$, the deformable attention module consumes less computational overhead than the standard attention module.

D. TRANSFORMER ENCODER AND FEATURE FUSION

For every transformer encoder layer, before the deformable attention module, the position embedding will be added to the query. Then the output features from the deformable attention module of a single image are available and there is a feed-forward network (FFN) for processing. Like DETR, the FFN contains a 2-layer MLP, and a normalization layer, and the output size is exactly the same as the input at the beginning of the layer.

After the transformer section in the encoder layer is completed, the final step is the feature fusion. Our model is capable of fusing features from images of the same entity.

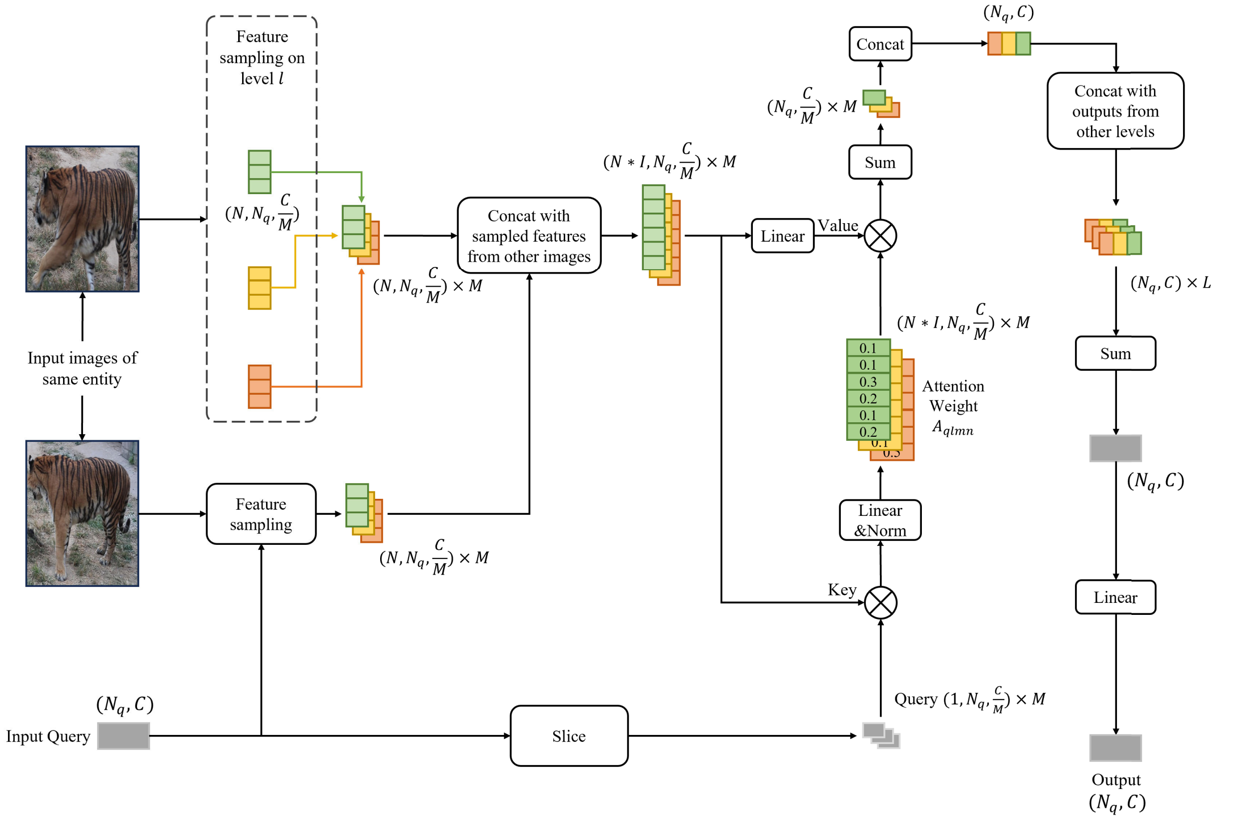

Feature sampling on level l
$(N, N_q, \frac{C}{M})$
$(N, N_q, \frac{C}{M}) \times M$
Concat with sampled features from other images
$(N*I, N_q, \frac{C}{M}) \times M$
Linear
Value
Sum
$(N_q, \frac{C}{M}) \times M$
Concat
(N_q, C)
Concat with outputs from other levels
$(N_q, C) \times L$
Sum
(N_q, C)
Linear
Output
(N_q, C)
$(N*I, N_q, \frac{C}{M}) \times M$
0.1
0.1
0.3
0.2
0.1
0.2
Attention Weight A_{qlmn}
Linear &Norm
Key
Query $(1, N_q, \frac{C}{M}) \times M$
Input images of same entity
Feature sampling
$(N, N_q, \frac{C}{M}) \times M$
(N_q, C)
Input Query
Slice

FIGURE 5. Deformable attention module in transformer fusion.

Assume that the number of images for fusion is I. In this section, two fusion methods are provided: a) mean fusion and b) transformer fusion, shown in Figure 5.

The mean fusion is relatively direct, which just gets the output of each image with the deformable attention module as shown in Figure 4, then calculates the mean of all output features belonging to the identical entity.

Transformer fusion contains a different deformable attention module as shown in Figure 6. Before attention calculation, sampled features will be combined into a longer feature sequence along the dimension of the sampling point N with shape $(1, N * I, C/M)$ in every head, every level, and every image. This is equivalent to more sampling points from other images belonging to the identical individual. Then, the attention output of the extended sequence and the query features would be determined and the result, which has the same length as the input query, is available.

In the Re-ID task, images in the gallery may be captured in different situations, and representation from a single image is not capable of representing the whole entity. The image feature fusion methods contribute an approach to integrate variant features in different circumstances, which means fewer representations in the gallery and less computation when querying an image. Thus, our model not only performs better during training and inference, but also consumes fewer resources when searching images in the gallery database.

E. LOSS FUNCTION

In many Re-ID tasks, the triplet loss function is frequently chosen as the loss function. In this paper, an improved triplet loss function with hard mining is adopted. This loss function does not consider other representations but focuses on the farthest representation of an identical entity and the nearest representation of a non-identical entity. They are treated as positive and negative respectively and the loss function can be denoted as:

$$\begin{aligned} Loss(a, pos, neg) = \ & max\{d(a_i, pos_{farthest}) \\ & -d(a_i, neg_{nearest}) + 1, 0\} \\ where\ d(x_i, y_i) = & \| \boldsymbol{x}_i - \boldsymbol{y}_i \|_2 \end{aligned} \tag{5}$$

The other part is the same as the common triplet loss and the margin is 1. In this case, only the most important images from both identical entities and non-identical entities are considered and the model is able to converge quickly [38].

IV. EXPERIMENT

A. TEST ENVIRONMENT AND CONFIGURATION

1) DATASET

The datasets are the basis of machine learning. In the early years, datasets for animal Re-ID were not common. But recently, animal reservation has become an important issue, and gradually, more related datasets have become available.

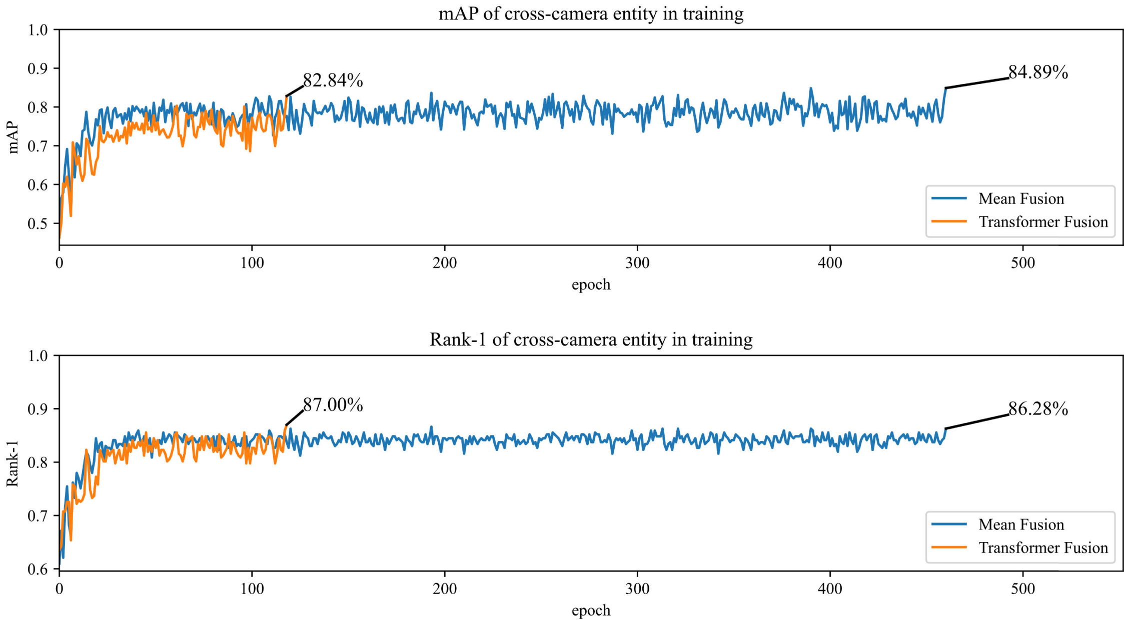


FIGURE 6. Training process of ReID-MF and ReID-TF on ATRW cross-camera testing set.

ATRW (Amur Tiger Re-identification in the Wild) contains over 8,000 video clips from 92 Amur tigers, with bounding box, pose key point, and tiger identity annotations [33].

MakerCollider collects over 8,000 video clips from about 10 zoos in China. The ATRW team annotates these video frames with bounding boxes, key-point-based poses, and identities to formulate the dataset. All the images were taken in zoos, so many sample images contain man-made objects such as fences and concrete floors. However, some pictures were captured from higher vantage points, resulting in backgrounds that feature only grass or trees without any artificial items. Additionally, the camera angles are not fixed, generally capturing the tigers from above or from the sides. Some sample images come from the same short video clip, thus having very similar angles and content.

For the Re-ID data, the left-side images and right-side images are treated as different entities of the tiger, which means the number of entities is greater than the number of tigers.

In the training set, there are 107 entities from 75 tigers with a total of 1,887 images. The ratio of single-camera entities to cross-camera entities is 6:4. Besides, 75 entities from 58 tigers are available in the testing set, which consists of 1,762 images. Detailed information about cross-camera and single-camera cases can be found in Table 1. Our paper focuses on the cross-camera case and achieves better mAP of cross-camera data in later experiments.

2) IMPLEMENTATION DETAILS AND MODEL CONFIGURATION

Our experiments were accomplished in Docker with 3 RTX 2080Ti GPUs hosted on a GPU server. The backbone consists of 7 2D-convolution layers with batch normalization and ReLU activation function, which is implemented by the function “Conv2dNormActivation” from the “torchvision” module. The detailed configuration of each layer is provided in Table 2.

TABLE 1. Statistics of testing set about the camera view.

Camera View	#images	#entities	#tigers
Single Camera	701	47	42
Cross Camera	1061	28	20

TABLE 2. Backbone configuration.

Layer	Kernel Size	Stride	Output	Output to Pyramid
1	[3, 3]	2	[32, 128, 128]	No
2	[3, 3]	2	[256, 64, 64]	No
3	[3, 3]	2	[1024, 32, 32]	No
4	[3, 3]	2	[256, 16, 16]	No
5	[3, 3]	2	[256, 8, 8]	Yes
6	[3, 3]	2	[256, 4, 4]	Yes
7	[3, 3]	2	[256, 2, 2]	Yes

The input shape of the backbone is [3, 256, 256], and the output feature pyramid contains the output feature maps of the last 3 layers with sizes of [256, 2, 2], [256, 4, 4], and [256, 8, 8]. For the convenience of subsequent calculations, all channel numbers are set to 256 ($C = 256$). The number of parameters in the backbone is around 6.57M.

In a common transformer, the input usually involves query, key, and value. The key and value are from the feature pyramid, while the queries are initiated by feeding the feature map of layer 5 in the backbone to a linear transformation and flattening it as a sequence and the final shape is [256, 64] ($N_q = 64$).

The deformable transformer part includes 4 encoder layers but no decoder layer to achieve a lightweight model. In the deformable attention module, no channel number variation occurs, and each query produces 4 sampling offsets ($N = 4$) on every feature pyramid level and every attention head by linear transformation, while the sampling basis position is fixed by linear interpolating the normalized plane with a size of 8 × 8. The number of feature pyramid levels is 3 ($L = 3$), the number of attention heads is 8 ($M = 8$), and the number of images for fusion is 4 ($I = 4$). The FFN after the deformable attention module is almost the same as Deformable DETR, except the depth is 1024.

All parameters of the encoder are shared among different images for fusion. The model with the mean fusion method is denoted as ReID-MF, and the transformer fusion method is denoted as ReID-TF. The optimizer in training is SGD with an initial learning rate of 0.05 ($lr_0 = 0.05$), momentum of 0.2, dampening of 0.2, and weight decay of 10^{-6}. A dynamic learning rate is adopted, and the learning rate in the k-th epoch can be calculated by:

$$lr_k = 0.93^k \cdot lr_0 \cdot (0.2 \cdot sin(\frac{2\pi k}{5}) + 1) \tag{6}$$

3) IMAGE AUGMENTATION

We apply image augmentation with a 50% probability. This image augmentation includes color jitter, random horizontal and vertical flip, random affine and perspective transformation, and all functions are provided by the "torchvision.transforms" package. Specifically, this process follows these steps:

Step 1. ColorJitter with brightness variance of [−0.2, +0.2], contrast variance of [−0.2, +0.2], saturation variance of [−0.1, +0.1], and hue variance of [−0.1, +0.1].

Step 2. RandomHorizontalFlip with 50% probability.

Step 3. RandomVerticalFlip with 50% probability.

Step 4. RandomApply a RandomAffine module with 50% probability. RandomAffine is configured with a random rotation degree in the range of [−90, 90], maximum absolute fraction for horizontal and vertical translations of 0.3, and a random scaling factor in the range of [0.8, 3].

Step 5. RandomPerspective with distortion scale of 0.5 and probability of 50%.

B. RESULTS AND COMPARISON

1) MAIN RESULTS

The training process of the two methods is shown in Figure 6. Compared with ReID-TF, ReID-MF requires more epochs to achieve the best mAP but has better performance in terms of mAP. Specifically, ReID-MF gains +2.05% mAP, while ReID-TF gains +0.72% Rank-1. Moreover, ReID-MF achieves the best result at epoch 460, whereas ReID-TF achieves the best result at epoch 118.

TABLE 3. Comparison with other methods on ATRW cross-camera testing set.

Method	Backbone	Cross-Cam Rank-1(%)	Cross-Cam mAP(%)	Params	GFLOPs
PPbM	ResNet50	77.1	47.8	≥28M	≥16
PPGNet	ResNet101	93.6	72.6	192M	63.07
TransReID	ViT-B/16	99.39	67.5	101M	1.06
MBR4B-LAI	ResNet50	98.78	67.13	59M	0.34
MSINet	-	**99.9**	58.66	2.4M	8.75
ReID-MF(ours)	7×Conv2D	86.28	**84.89**	10.6M	13.55
ReID-TF(ours)	7×Conv2D	87	82.84	8.4M	13.55

2) COMPARISON

Since there are limited models designed for animal Re-ID, we select some state-of-the-art models in the vehicle Re-ID task for comparison. As shown in Table 3, PPbM [33] is the model proposed by the ATRW team, and PPGNet [39] is the champion model of Re-ID in Computer Vision for Wildlife Conservation 2019. TransReID [15], MBR4B-LAI [40], and MSINet [11] are recent vehicle Re-ID models that achieve high accuracy in the ATRW cross-camera testing set. However, their mAP is not satisfactory compared to previous animal Re-ID models.

Our model yields significant improvement in the cross-camera mAP, and the model parameter size is relatively lightweight since our backbone consists of just 7 Conv2D layers with a size of 6.57M. Each encoder layer of the deformable transformer contributes around 1M parameters, resulting in a total parameter size of around 10M. Meanwhile, the mAP of our model gains +12.29% compared to the SOTA method.

3) SAMPLING POSITIONS IN DEFORMABLE ATTENTION MODULE

As mentioned earlier, the deformable attention module samples features on the pyramid according to the reference position and the sampling offset. We grabbed the sampling position of every deformable attention module when the trained ReID-MF model was inferencing and plotted those points on the original image. Only the queries with the top 20% attention weight were pointed out, as shown in Figure 7(a). Most sampling positions are located on the tiger body, especially in layer 3. In Figure 7(b), we provide more sampling positions in layer 3 of another image.

C. ABLATION STUDY

1) NUMBER OF IMAGES FOR FUSION

As shown in Table 4, we test the impact of different image numbers for fusion in the ReID-MF model. Our model has no limitation on the number of images for fusion as long as the hardware allows. However, a larger number of images for fusion does not mean better performance.

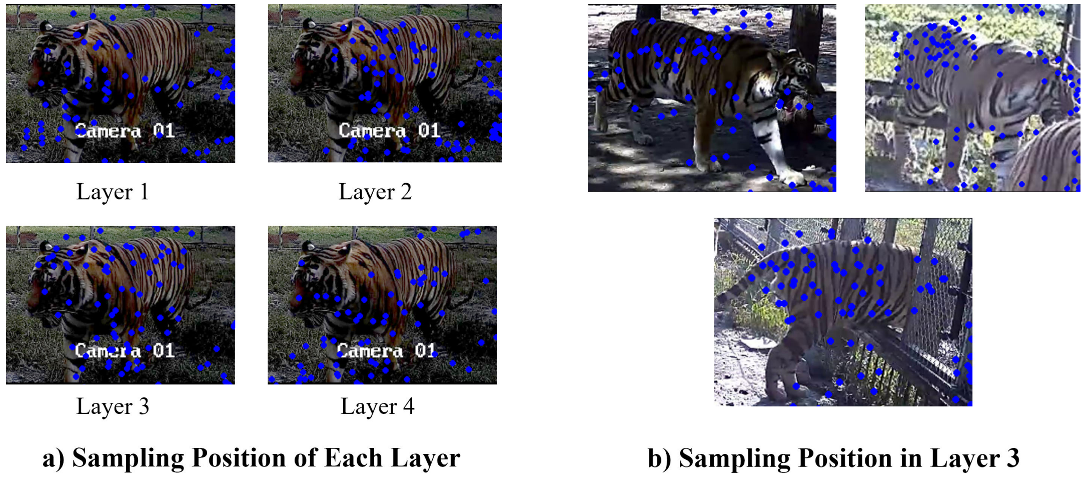

Camera 01
Layer 1
Camera 01
Layer 2
Camera 01
Layer 3
Camera 01
Layer 4
a) Sampling Position of Each Layer
b) Sampling Position in Layer 3

FIGURE 7. Sampling position study a) Sampling position of each layer. Due to the low brightness of the original image, we increased the exposure and reduced the contrast of the picture for a clearer view. b) Sampling position in layer 3 of another image.

TABLE 4. Ablation study for the number of fusion images.

#Images for Fusion	Cross-Cam Rank-1(%)	Cross-Cam mAP(%)
2	83.03	80.53
3	84.1	78.7
4	**86.28**	**84.89**
5	84.5	83.4
6	80.5	78.5
7	81.2	78.1
8	70.37	66.38

When $I = 4$, the result is the best. But further increases show worse performance. This situation may be caused by the mean fusion method. Since images in the dataset are captured from random viewpoints, more images for fusion mean more variance of viewpoint and representation of each image. However, the mean fusion method just calculates the mean feature and cannot prevent the distortion of representation caused by the different viewpoints. Once too many images are input for fusion, the final representation will be distorted and the result will be worse. Conversely, a limited number of images for fusion result in less entity feature in the final representation, which also leads to a worse result.

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

In conclusion, we present a novel approach, ReDeformTR, for wildlife re-identification based on a lightweight deformable transformer with multi-image feature fusion. The re-identification task, crucial for object retrieval, has garnered significant attention due to its applications in public safety and smart management across various domains, including wildlife tracking and conservation.

Traditional re-identification methods face challenges such as varying viewpoints and image characteristics, leading to ambiguity during identification. To address these challenges, we introduce a DETR-based model that incorporates deformable attention mechanisms and feature fusion from multiple images. This approach enables the extraction of representative features from wildlife photos captured under different conditions and angles.

The methodology section outlines the model architecture, including the backbone network for feature extraction, deformable attention modules for refining and fusing features, and the loss function used for training. We also provide details on the experimental setup, including dataset information, model configurations, and training parameters.

Experimental results demonstrate the effectiveness of the proposed approach, with the ReID-MF method outperforming ReID-TF in terms of mAP. Comparison with state-of-the-art models in vehicle re-identification tasks highlights the superior performance of the proposed method on the ATRW cross-camera testing set, achieving a significant improvement in mAP while maintaining a lightweight parameter size.

Further analysis includes a study of sampling positions in the deformable attention module, showcasing the model's ability to focus on key features in wildlife images. Additionally, ablation studies explore the impact of varying the number of images for fusion, revealing optimal performance with a balanced approach.

B. FUTURE WORK

In the future, we plan to improve the feature fusion method of this model. As shown in the experiment section, the mean fusion method performs better than the transformer fusion method. However, the simple mean operation cannot highlight which part of the feature is more significant to the final representation. On the contrary, theoretically, the transformer fusion method is capable of this task. Also, it is an option to introduce a decoder-like structure to refine the features from different images belonging to the identical entity. Furthermore, there is much extra information about tigers, like boundary boxes and key points of entity skeletons. This data can be utilized to improve the training process. For example, adopting a random mask of tiger body parts in image augmentation.

Overall, the ReDeformTR model presents a promising solution for wildlife re-identification tasks, offering improved accuracy, reduced computational overhead, and scalability for real-world applications in wildlife conservation and management.

REFERENCES

[1] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7270–7292, Jun. 2023.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[3] H. Wang, J. Hou, and N. Chen, "A survey of vehicle re-identification based on deep learning," *IEEE Access*, vol. 7, pp. 172443–172469, 2019.

[4] S. D. Khan and H. Ullah, "A survey of advances in vision-based vehicle re-identification," *Comput. Vis. Image Understand.*, vol. 182, pp. 50–63, May 2019.

[5] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2022, pp. 1142–1160.

[6] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2897–2906.

[7] H. Ge, K. Zhang, L. Sun, and G. Tan, "Exploring latent information for unsupervised person re-identification by discriminative learning networks," *IEEE Access*, vol. 8, pp. 44748–44759, 2020.

[8] Z. Cao and H. J. Lee, "Learning multi-scale features and batch-normalized global features for person re-identification," *IEEE Access*, vol. 8, pp. 184644–184655, 2020.

[9] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11921–11930.

[10] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, "Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Sep. 2023, pp. 15050–15061, doi: 10.1109/CVPR52729.2023.01445.

[11] J. Gu, K. Wang, H. Luo, C. Chen, W. Jiang, Y. Fang, S. Zhang, Y. You, and J. Zhao, "MSINet: Twins contrastive search of multi-scale interaction for object Reid," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19243–19253.

[12] Z. Deng, Y. Zhong, S. Guo, and W. Huang, "InsCLR: Improving instance retrieval with self-supervision," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 516–524.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010. [Online]. Available: https://arxiv.org/abs/1706.03762

[14] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1005–1014.

[15] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.

[16] M. Li, X. Huang, and Z. Zhang, "Self-supervised geometric features discovery via interpretable attention for vehicle re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 194–204.

[17] Z. Yang, X. Jin, K. Zheng, and F. Zhao, "Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14278–14287.

[18] K. Zhu, H. Guo, T. Yan, Y. Zhu, J. Wang, and M. Tang, "PASS: Part-aware self-supervised pre-training for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 198–214.

[19] H. Chen, B. Lagadec, and F. Bremond, "ICE: Inter-instance contrastive encoding for unsupervised person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14940–14949.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[21] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[22] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, and L. Zhang, "Dynamic DETR: End-to-end object detection with dynamic attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2968–2977.

[23] D. Zheng, W. Dong, H. Hu, X. Chen, and Y. Wang, "Less is more: Focus attention for efficient DETR," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6674–6683.

[24] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query DeNoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13609–13617.

[25] F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12085–12095.

[26] C. Bai, H. Li, J. Zhang, L. Huang, and L. Zhang, "Unsupervised adversarial instance-level image retrieval," *IEEE Trans. Multimedia*, vol. 23, pp. 2199–2207, 2021.

[27] G.-H. Liu and J.-Y. Yang, "Deep-seated features histogram: A novel image retrieval method," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107926.

[28] X.-S. Wei, Y.-Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8927–8948, Dec. 2022.

[29] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "TransFG: A transformer architecture for fine-grained recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 852–860, doi: 10.1609/aaai.v36i1.19967.

[30] M. V. Conde and K. Turgutlu, "CLIP-Art: Contrastive pre-training for fine-grained art classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3951–3955.

[31] X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, "Fine-grained object classification via self-supervised pose alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7389–7398.

[32] Y. Ding, Z. Ma, S. Wen, J. Xie, D. Chang, Z. Si, M. Wu, and H. Ling, "AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2826–2836, 2021.

[33] S. Li, J. Li, H. Tang, R. Qian, and W. Lin, "ATRW: A benchmark for amur tiger re-identification in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2590–2598.

[34] E. Nepovinnykh, T. Eerola, V. Biard, P. Mutka, M. Niemi, M. Kunnasranta, and H. Kälviäinen, "SealID: Saimaa ringed seal re-identification dataset," *Sensors*, vol. 22, no. 19, p. 7602, May 2022, doi: 10.3390/s22197602.

[35] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Khl, and J. Denzler, "Chimpanzee faces in the wild: Log-Euclidean CNNs for predicting identities and attributes of primates," in *Proc. Pattern Recognit. 38th Ger. Conf. GCPR 2016*, Germany, 2016, pp. 51–63.

[36] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[37] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[38] H. Xuan, A. Stylianou, and R. Pless, "Improved embeddings with easy positive triplet mining," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2463–2471.

[39] C. Liu, R. Zhang, and L. Guo, "Part-pose guided amur tiger re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 315–322.

[40] E. Almeida, B. Silva, and J. Batista, "Strength in diversity: Multi-branch representation learning for vehicle re-identification," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 4690–4696.

ZITONG LI received the B.Sc. degree in electronic and information engineering from Hong Kong Polytechnic University, Hong Kong, China, in 2017. He is currently pursuing the master's degree in agricultural information technology with Hunan University of Humanities, Science and Technology, Loudi, China. His research interests include computer vision and blockchain.

ZHENGMAO YAN received the B.Sc. degree in computer science and technology from Hunan University of Humanities, Science and Technology, Loudi, China, in 2023, where he is currently pursuing the master's degree in agricultural information technology. His research interests include blockchain and edge computing.

WEIHONG TIAN received the B.Sc. degree in information management and systems from Xinjiang Agricultural University, Urumqi, China, in 2001. She is currently pursuing the master's degree in agricultural information technology with Hunan University of Humanities, Science and Technology, Loudi, China. Her research interests include blockchain and artificial intelligence.

DETIAN ZENG received the M.S. degree in information and communication engineering from Central South University, in 2018. He is currently an Assistant Professor with Hunan University of Humanities, Science and Technology, Loudi, China. His research interests include neural networks and intelligent optimization.

YI LIU received the master's degree in computer science from Hunan University, in 2010. She is currently pursuing the Ph.D. degree in software engineering with Hunan University of Humanities, Science and Technology. Her research interests include code big data and computer vision.

WEIMIN LI received the M.Sc. degree in computer technology from Yunnan University, Kunming, China, in 2010, and the Ph.D. degree in software engineering from Central South University, Changsha, China, in 2018. He is currently an Associate Professor with the School of Information, Hunan University of Humanities, Science and Technology, Loudi, China. His research interests include edge computing and computer vision.