Overcoming Fine-Grained Visual Challenges in Animal Re-Identification via Semantic Feature Alignment - Supplementary Material

A. CARE Toolkit

In the pre-processing phase of the CARE toolkit, we leverage the pre-trained YOLO11 (You Only Look Once) [5] model to develop a species detector, enabling an autonomous process and annotation of the camera-trap data prior to identifying individual animals. We used the Labeled Information Library of Alexandria: Biology and Conservation (LILA BC) [10] dataset, supplemented with additional stoat data donated from five organizations, to fine-tune the pre-trained YOLO11 model. Specifically, we customized a species detector with 15 categories (*e.g.*, Bird, Cat, Deer, Ferret, Kiwi, Possum, Rodent, Stoat) by fine-tuning the YOLO11 model with approximately 350,000 images.

B. Algorithms

The pseudo-code for the VDTDG training in Phase 1 of our proposed CARE framework is demonstrated in Algorithm 1. Algorithm 2 shows the pseudo-code for the individual-specific textual prototypes fusion. The pseudo-code for semantic feature alignment within individuals to enhance the discrimination between individual animals is illustrated in Algorithm 3.

Algorithm 1 CARE Phase 1: VDTDG Training

Input: $\mathcal{D} = \{(\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D})\}$: training dataset with a batch size B.
Output: $\mathcal{F}$: optimized VDTDG.
Initialize $\mathcal{I}(\cdot), \mathcal{T}(\cdot)$, and $s(\cdot,\cdot)$ from the pre-trained CLIP.
Initialize VDTDG $\mathcal{F}(\cdot)$.
Initialize $v_1, v_2, \cdots, v_w$ randomly. ▷ initialize learnable textual tokens
Initialize static textual tokens $\omega_{y_i^c}$ with the identity label y_i^c.
while in training VDTDG **do**
  for (x_i, y_i^c) in $(\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D})$ **do**
    $v_{x_i} = \mathcal{I}(x_i)$
    $\omega'_{x_i} = \mathcal{F}(v_{x_i})$ ▷ generate image-specific textual tokens
    $v_1, v_2, \cdots, v_w \mathrel{+}= \omega'_{x_i}$
    $t_{x_i} = \omega_{y_i^c} \oplus \omega'_{x_i}$
    Compute $\mathcal{L}_{img2txt}$ with Equation (5)
    Compute $\mathcal{L}_{txt2img}$ with Equation (6)
    $\mathcal{L}_{VDTDG} = \mathcal{L}_{img2txt} + \mathcal{L}_{txt2img}$
    Update $\mathcal{F}(\cdot)$ with $\mathcal{L}_{VDTDG}$
  end for
end while

C. Experimental Settings

Datasets. We evaluate our proposed CARE framework on seven Animal ReID benchmarks: **(1) ATRW** [7] includes 4,675 images from 182 Amur Tiger individuals; **(2) FriesianCattle2017** [1] includes 934 images from 84 Holstein Friesian Cattle individuals; **(3) LionData** [2] includes 740 images from 94 Lion individuals; **(4) MPDD** [4] includes 1,657 images from 191 Dog individuals; **(5) IPanda50** [14] includes 6,874 images from 50 Panda individuals; **(6) SeaStarReID2023** [13] includes 2,187 images from 95 Sea Star individuals; **(7) NyalaData** [2] includes 1,942 images from 237 Nyala individuals, and a newly collected **Stoat** dataset, including 5,302 images from 73 Stoat individuals.

Baselines. We compare our approach against five representative baselines: **(1) TransReID** [3] is a Transformer-based Object ReID framework with the jigsaw patches module and side information embedding to mitigate feature bias from the data for robust re-identification; **(2) AdaFreq** [6] leverages ViTs to enhance high-frequency feature learning via a data augmentation strategy and high-frequency information extraction for wildlife ReID; **(3) CLIP-ReID** [8] exploits the power of CLIP with prompt learning, outperforming previous CNN- and ViT-based ReID methods; **(4) CLIP-FineTuning (CLIP-FT)** is a baseline variant, which involves refining the visual concepts learning for Animal ReID through CLIP's image encoder adaptation; and **(5) CLIP-ZeroShot (CLIP-ZS)** relies on the pre-trained CLIP's image encoder for the inferences, exploring the zero-shot capabilities of CLIP for Animal ReID.

Evaluation Metrics. We employ two metrics for ReID tasks to evaluate CARE: mean Average Precision (mAP) [15] and Cumulative Matching Characteristics (CMC) [9]. Here, mAP assesses the model's overall

Dataset	$w=2$		$w=4$		$w=8$		$w=16$	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Tiger	54.3±.4	93.8±.3	**56.8±.2**	**95.3±.1**	53.8±.4	93.9±.1	53.5±.9	94.5±.6
Cattle	93.3±.2	98.8±.3	**95.8±.2**	**99.7±.3**	93.4±.1	98.8±.3	92.4±.4	98.8±.1
Lion	39.0±.1	53.7±.4	**40.3±.3**	55.3±.2	39.2±.1	54.9±.2	39.2±.1	**54.6±.2**
Dog	87.6±.3	95.9±.5	**89.0±.2**	95.4±.4	87.4±.2	95.5±.1	87.1±.2	95.3±.2
Panda	35.8±.1	47.7±.3	**37.3±.2**	48.2±.4	37.1±.2	47.9±.7	37.1±.5	**49.0±.3**
Sea Star	83.9±.4	98.3±.2	**88.6±.4**	**99.6±.1**	83.5±.5	98.1±.2	83.3±.6	98.1±.2
Nyala	16.5±.3	28.4±.9	16.6±.1	29.0±.2	16.0±.8	27.3±.9	15.8±.0	25.9±.5
Stoat	90.6±.3	97.5±.0	90.6±.1	96.9±.5	90.5±.4	96.9±.9	90.1±.3	95.0±.0

Table 1. Analysis of the number of learnable textual tokens (w) in image-specific textual descriptions generated from VDTDG.

Dataset	$\gamma=0.1$		$\gamma=0.3$		$\gamma=0.5$		$\gamma=0.7$		$\gamma=0.9$	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Tiger	55.4±.1	**95.1±.1**	53.8±.2	94.5±.4	53.0±.2	93.5±.2	**56.8±.2**	95.3±.1	53.1±.2	94.1±.3
Cattle	94.0±.2	100.0±.0	95.1±.3	100.0±.0	94.7±.2	100.0±.0	**95.8±.2**	99.7±.3	94.9±.2	99.7±.3
Lion	39.6±.2	54.1±.6	40.2±.5	52.9±.7	39.8±.2	52.9±.9	40.3±.3	**55.3±.2**	40.6±.3	56.2±.9
Dog	88.6±.2	96.4±.4	88.5±.2	95.9±.4	89.1±.1	95.9±.2	89.0±.2	95.4±.4	88.9±.1	95.4±.2
Panda	36.6±.2	**49.5±.5**	36.4±.2	47.3±.3	36.8±.1	47.9±.6	**37.3±.2**	48.2±.4	36.7±.1	47.7±.7
Sea Star	85.9±.1	97.6±.1	87.5±.1	98.0±.1	**88.5±.1**	99.1±.2	88.6±.4	**99.6±.1**	88.7±.2	99.1±.1
Nyala	16.0±.1	27.5±.5	16.2±.2	28.6±.4	16.4±.0	28.2±.2	**16.6±.1**	29.0±.2	16.3±.1	29.0±.2
Stoat	90.2±.1	**95.6±.5**	90.1±.3	96.2±.6	89.9±.1	95.0±.0	**90.6±.1**	96.9±.5	89.4±.2	95.0±.0

Table 2. Analysis of the effect of the margin (γ) in triplet loss $\mathcal{L}_{tri}$.

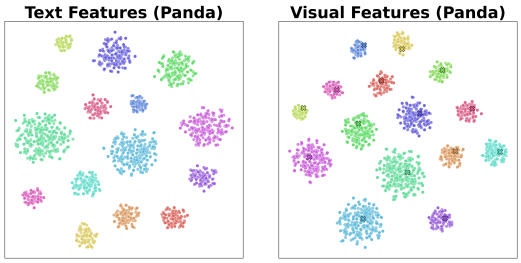
Algorithm 2 Individual-Specific Textual Prototypes Fusion

Input: $\mathcal{D} = \{(\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D})\}$: entire training dataset with a size of N_D; $\mathcal{F}$: optimized VDTDG.
Output: $T_{Bank} \in \mathbb{R}^{\Phi \times d}$: memory bank to store the textual prototypes of all individuals.
Initialize $\mathcal{I}(\cdot)$ and $\mathcal{T}(\cdot)$ from the pre-trained CLIP.
Initialize a memory bank T_{Bank}.
for each distinct identity ϕ in $\mathcal{D}$ **do**
    Retrieve all samples, $\{(x_i, y_i^c)\}_{i=1}^n$, with the identity ϕ from $\mathcal{D}$.
    Compute the image-specific textual descriptions of these retrieved n samples with the optimized $\mathcal{F}$.
    Compute the textual prototype for the individual with the identity label ϕ through a mean pooling fusion and store it into the memory bank T_{Bank}.
end for

Algorithm 3 CARE Phase 2: Semantic Feature Alignment for Animal ReID

Input: $\mathcal{D} = \{(\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D})\}$: training dataset with a batch size B; T_{Bank}: memory bank.
Output: $\mathcal{I}$: optimized CLIP's image encoder.
Initialize $\mathcal{I}(\cdot), \mathcal{T}(\cdot)$, and $s(\cdot,\cdot)$ from the pre-trained CLIP.
while in fine-tuning CLIP's image encoder **do**
    for (x_i, y_i^c) in $(\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D})$ **do**
        $v_{x_i} = \mathcal{I}(x_i)$
        $t_{y_i^c} = T_{Bank}[i]$ ▷ extract the individual-specific textual prototype from the memory bank T_{Bank}
        Compute $\mathcal{L}_{id}$ with Equation (8)
        Compute $\mathcal{L}_{tri}$ with Equation (9)
        Compute $\mathcal{L}_{i2pCon}$ with Equation (10)
        $\mathcal{L}_{FA} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \mathcal{L}_{i2pCon}$
        Update $\mathcal{I}(\cdot)$ with $\mathcal{L}_{FA}$
    end for
end while

retrieval performance through the average precision computation across all query images from the Query Set $\mathcal{Q}$, highlighting the comprehensive capability of ReID models. CMC measures the model's ability to identify the correct match within the top-k ranked gallery images from the Gallery Set $\mathcal{G}$, revealing the number of images that have to be examined to achieve a desired level of performance. We report the mAP and CMC accuracy at Rank-1. All performance measures are averaged over ten runs, along with their corresponding 95% confidence intervals.



(a) Visualization of text features for each individual after training the VDTDG in Phase 1.

(b) Visualization of visual features for each individual after finetuning CLIP's image encoder in Phase 2.

Figure 1. **t-SNE [12] visualizations of text and visual features for 15 individuals from the IPanda50 [14] dataset.** Each color represents a distinct individual. The cross shape in (b) depicts the individual-specific textual prototype of each individual.

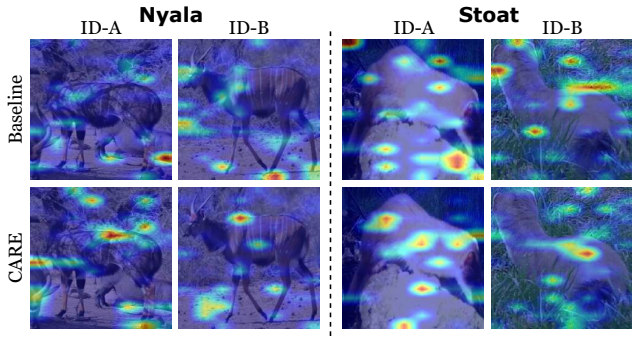


Figure 2. **Grad-CAM [11] visualizations for the failure cases on the NyalaData [2] and Stoat datasets.** We show the Grad-CAM failure samples with a comparison between the baseline method (top row) and our proposed framework (bottom row). Two distinct individuals of each species are displayed.

D. Further Analysis

Sensitivity Analysis. Table 1 presents an analysis of how the number of learnable textual tokens (w) in image-specific textual descriptions generated by VDTDG impacts the performance of the CARE framework on Animal ReID. We investigate the sensitivity to the expressiveness of the generated prompts and evaluate the performance across eight Animal ReID datasets for w values ranging from 2 to 16. The results suggest that learning 4 textual tokens is sufficient to capture the semantic features from the visual streams.

Table 2 shows a sensitivity analysis of the margin (γ) in triplet loss, evaluating its impact on CARE's performance across various animal datasets. The study reveals that a margin of $\gamma = 0.7$ yields the highest overall performance (mAP) for five datasets. This indicates that enforcing a greater separation between positive and negative samples is crucial for learning highly discriminative features in Animal ReID, due to the inherent visual challenge of low inter-identity variations in animals.

t-SNE. Figure 1a illustrates the text features of the learned image-specific textual descriptions from VDTDG for each individual. The images within an individual are associated with slightly different textual tokens, capturing the individual's semantic features across various visual streams. Furthermore, the clusters of textual tokens across individuals are clearly separated, mirroring the structured patterns observed in the visual features shown in Figure 1b. Figure 1b highlights the relationship between the individual-specific textual prototypes and the visual features. The visual embeddings are optimized to cluster around the corresponding textual prototypes, where each individual cluster exhibits clear margins. This demonstrates the effectiveness of individual-specific textual prototypes in distinguishing individuals, thereby enhancing Animal ReID performance.

Analysis of Grad-CAM Failure Cases. We analyze Grad-CAM failure cases using samples from the NyalaData [2] and Stoat datasets. Figure 2 presents visualizations from both the baseline method and CARE for two distinct individuals within the same species. The results indicate that, in some instances, models attend to spurious features, such as background elements, rocks, grass, and surrounding vegetation, when re-identifying individual animals. This limitation could degrade the model's generalizability, especially when images of the same individual are captured in different environments or when distinct individuals appear in visually similar surroundings. Compared with the baseline, our proposed framework directs most of its attention to the animals themselves (*e.g.*, the Nyala's upper torso and the Stoat's body), though some residual focus on contextual background remains.

References

[1] William Andrew, Colin Greatwood, and Tilo Burghardt. Visual localisation and individual identification of holstein friesian cattle via deep learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017. 1

[2] Nkosikhona Dlamini and Terence L van Zyl. Automated identification of individuals in wildlife population using siamese neural networks. In *2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2020. 1, 3

[3] 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 (ICCV)*, 2021. 1

[4] 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)*, 2023. 1

[5] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO. https://github.com/ultralytics/ultralytics, 2024. Version 11.0.0, License AGPL-3.0. 1

[6] Chenyue Li, Shuoyi Chen, and Mang Ye. Adaptive high-frequency transformer for diverse wildlife re-identification. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*. Springer, 2024. 1

[7] 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*, 2020. 1

[8] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[9] Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 2001. 1

[10] Dan Morris. Trail Camera Images of New Zealand Animals. https://lila.science/datasets/nz-trailcams, 2025. 1

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[12] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. 3

[13] Oscar Wahltinez and Sarah J. Wahltinez. An open-source general purpose machine learning framework for individual animal re-identification using few-shot learning. *Methods in Ecology and Evolution*, 2024. 1

[14] Le Wang, Rizhi Ding, Yuanhao Zhai, Qilin Zhang, Wei Tang, Nanning Zheng, and Gang Hua. Giant panda identification. *IEEE Transactions on Image Processing*, 2021. 1, 3

[15] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1