

Visual-Prompt Guided Wildlife Instance-Level Recognition

Mufhumudzi Muthivhi¹, Jiahao Huo¹, Terence van Zyl¹, and Fredrik Gustafsson²

¹ University of Johannesburg, South Africa
 {mmuthivhi,216045414,tvanzyl}@uj.ac.za

² Linköping University, Sweden fredrik.gustafsson@liu.se

Abstract. Fine-grained wildlife re-identification remains a challenging area in research. Current state-of-the-art approaches apply a detection and re-identification pipeline. We propose a one-stage end-to-end detection and re-identification model that performs identity searching within the latent space. We adopt DINOv2 for robust spatial geometry and MegaDescriptor for wildlife re-identification. We enhance latent queries with prompt re-identification features. A detection decoder queries the scene latent space to establish object boundaries around the target identity. Preliminary findings reflect a competitive mean average precision score of 30.584% compared to the state-of-the-art two stage approach of 44.89%. Qualitative results depict effective bounding and identification of animal identities.

Keywords: instance recognition · re-identification · grounding · wildlife

1 Introduction

Generic object detection models categorize wildlife at the species level. Conversely, instance-level wildlife re-identification requires the discrimination of fine-grained visual patterns. The model is tasked with learning from the unique stripe configurations of a zebra or the spot patterns of a cheetah. Current state-of-the-art approaches deploy an isolated detection and re-identification pipeline using two separate models [8]. An object detection model isolates individuals in a scene, crops the bounding box pixels, resizes them to a fixed resolution, and passes them to an independent re-identification feature extractor for similarity matching against a database [2]. Recent advancements in promptable vision foundation models and visual grounding present a compelling alternative [5]. We could instead view the database as a prompt to query the scene directly to locate a target. However, adapting single-stage visual grounding architectures for instance-level re-identification introduces a training mismatch. A single transformer decoder layer is forced to simultaneously learn the global scenery to extract geometric boundaries of an object and the local texture patterns of a specific identity. To resolve this, we propose an end-to-end visual prompting task.

A frontier foundation model extracts robust spatial features and a specialist re-identification model extracts fine-grained features. Preliminary findings depict a competitive mean average precision score (mAP) over a two-stage approach. Qualitative results also depict tight bounding boxes around target identities in dense herd, occluded environments and distant subjects.

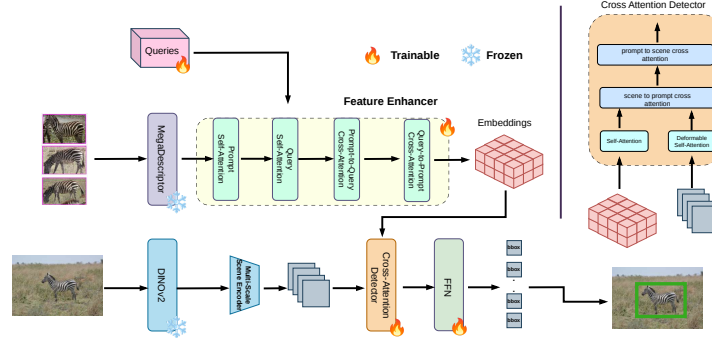


Fig. 1: Overview of the proposed framework. DINOv2 encodes the full scene image to capture spatial features and MegaDescriptor encodes reference crops into identity features used as visual prompts for target matching.

2 Methodology

The proposed framework in Figure 1 consists of three primary components. First, we extract scene and re-identification features from a frozen encoder. Next, we enhance the object queries with prompt re-identification features to enrich our region proposals. Lastly, we use a cross-attention detector to search for the target identity.

Multi-Scale Scene Encoder We use DINOv2 [6] to process the full scene image. To handle extreme scale variations in imagery, we adopt a similar strategy used by Grounding DINO [5]. We extract intermediate feature maps from layers 3, 6 and 10 within the transformer. These features are projected into a feature pyramid using varying convolutional strides, spatially interpolated to a uniform resolution, concatenated, and subsequently compressed via a 1×1 convolution.

Visual prompt The visual prompt is obtained through a pretrained fine-grained wildlife re-identification model, MegaDescriptor [2].

Feature Enhancer We initialize a set of learnable object queries as the region proposals. The feature enhancement layer ensures that the queries cross-attend directly with the visual prompt features. We use both self and cross attention to

enrich our region proposals (queries) with identity-related information from our visual prompts.

Cross-Attention Decoder The enriched region proposals attend to the multi-scale scene features via Deformable Attention [9]. The decoder outputs a set of N cross-attended queries. The cross-attended queries are passed through a feed-forward network to encapsulate the detected animal with a bounding box. An additional objectiveness score is also retrieved, indicating the presence of a valid foreground animal.

Loss Function This work proposes the optimization of four objectives, such that:

$$\mathcal{L} = \lambda_{obj}\mathcal{L}_{\text{FocalLoss}} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{GIoU}\mathcal{L}_{GIoU} + \lambda_{div}\mathcal{L}_{div} \quad (1)$$

is the detection loss. $\mathcal{L}_{FL}$ is the Sigmoid Focal Loss over the objectness score obtained from the N queries. $\mathcal{L}_{L1}$ and $\mathcal{L}_{GIoU}$ is the L1 loss and Generalized Intersection over Union of the predicted and ground truth bounding boxes. $\mathcal{L}_{div}$ encourages diversity of the embeddings of the queries produced by the feature enhancement layer.

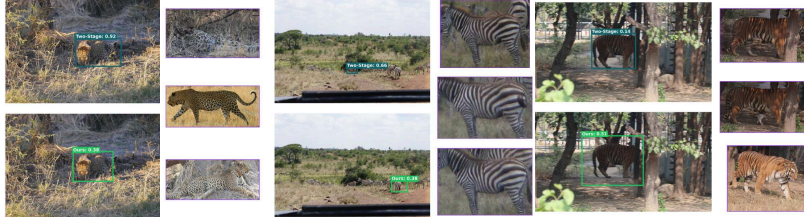


Fig. 2: Qualitative visual grounding results of the proposed architecture. For each scenario, the target identity prompts are displayed on the left, and the scene interrogation results are on the right.

3 Experimental Setup

Datasets We use a curated re-identification dataset from *WildlifeDatasets* [2]. We specifically utilize the *GiraffeZebraID* [7] and *ATRW* [4] datasets because they come with the full-scene image and the bounding box metadata.

Implementation We employ a ViT-S/14 variant of DINOv2 for the multi-scale scene encoder and a MegaDescriptor-T variant for the prompt encoder. Scene images are resized to 518×518 to align with the DINOv2 patch configuration, while prompt crops are resized to 224×224 . We selectively unfreeze the final transformer block and normalization layers of both encoders during training, allowing their distinct latent manifolds to synchronize. We use PyTorch Lightning for efficient training [3].

Hyperparameters: The network is optimized end-to-end using the AdamW optimizer with a base learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . We employ a MultiStepLR learning rate scheduler, which decays the rate by a factor of $\gamma = 0.1$. The objectness, identity matching, bounding box L1 distance, and Generalized IoU is set to 2.0, 2.0, 5.0, and 2.0, respectively.

4 Preliminary Results

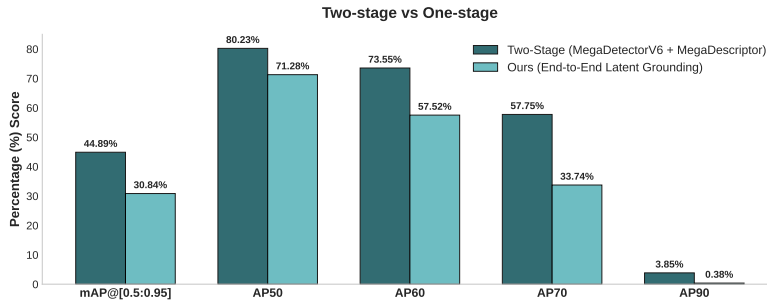


Fig. 3: Our End-to-End Latent Grounding method (one stage) compared to two stage method using MegadetectorV6 [1] with MegaDescriptor [2]. The results shown is evaluated on the combined test set from GiraffeZebraID [7] and ATRW [4]

We provide qualitative preliminary findings in Figure 2. The proposed framework produces bounding boxes that encapsulate the target animal. Our model can handle herds, domain shifts, occlusion, distance, and camouflage. We present results on retrieval performance (ReID) in Figure 3 using our end-to-end latent grounding method and compared it to the two-stage approach that uses MegadetectorV6 for detection, crop and followed by MegaDescriptor to embed the cropped images to perform retrieval. The results show the potential of the proposed approach even though it has not yet achieved state-of-the-art performance. However, the proposed approach is much more efficient.

5 Conclusion

This paper proposes a framework that combines visual prompt along with global scene features to perform end-to-end object detection together with re-identification. The preliminary findings suggest that visual identity prompts can reduce the reliance on traditional two-stage detect and crop followed by ReID pipelines by allowing the model to localize and match a target directly within the full scene image. We believe continued work in this direction could lead to state-of-the-art performance.

References

1. Beery, S., Morris, D., Yang, S.: Efficient pipeline for camera trap image review (2019)
2. Čermák, V., Pícek, L., Adam, L., Papafitsoros, K.: Wildlifedatasets: An open-source toolkit for animal re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5953–5963 (2024)
3. Falcon, W., team, T.P.L.: Pytorch lightning. <https://doi.org/10.5281/zenodo.3828935>, <https://github.com/Lightning-AI/pytorch-lightning>
4. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: Atrw: A benchmark for amur tiger re-identification in the wild. arXiv preprint arXiv:1906.05586 (2019)
5. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
6. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
7. Parham, J., Crall, J., Stewart, C., Berger-Wolf, T., Rubenstein, D.I.: Animal population censusing at scale with citizen science and photographic identification. In: AAAI spring symposium-technical report (2017)
8. Wu, Y., Zhao, D., Li, Y., Alajas, M., Glen, A.S., Zhang, J., Dobbie, G., Wilson, D., Koh, Y.S.: Overcoming fine-grained visual challenges in animal re-identification via semantic feature alignment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 371–381 (2026)
9. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4794–4803 (2022)