Karlsruher Institut für Technologie

Fakultät für Wirtschaftswissenschaften
Institut für Operations Research
Lehrstuhl Analytics and Statistics
Prof. Dr. Oliver Grothe

Effectiveness of Bee Reidentification using their Abdomen - A Reproduction and Analysis

Seminararbeit

Christopher Roth
Matrikelnr.: 2281943

July 15, 2024

Supervised by Parzival Borlinghaus

Contents

1. Introduction

Animal reidentification is a burgeoning field within computer vision that aims to address the challenge of recognizing and tracking individual animals over time and across different locations. This capability is critical for various applications, including wildlife conservation, behavioural studies, and ecological monitoring (Krebs 1989). The growing interest in this area is driven by the need for non-invasive and scalable methods to monitor animal populations, which traditional techniques, such as tagging and manual observation, often fail to meet.

Recent advancements in deep learning and computer vision have significantly enhanced the accuracy and feasibility of animal reidentification systems (Schneider, Taylor, Linquist, et al. 2019). Early methods relied heavily on handcrafted features (Kelly 2001, Ardovini, Cinque, and Sangineto 2008), and traditional machine learning techniques, often species-specific and requiring significant domain expertise to provide labelled datasets. Also, the so-called *observer-bias* (Tuyttens et al. 2014) impacted the quality of datasets. Besides, while identifying animals such as whales (Mizroch, Beard, and Lynde 1990) might be possible for human experts, there is still an observational bias that needs to be encountered (Foster and Harmsen 2012). The problem of identifying individuals to construct large labelled datasets, becomes even more prominent when trying to re-identify population-intensive animals such as bees or bumblebees (Borlinghaus, Tausch, and Rettenberger 2023). However, the advent of convolutional neural networks (CNNs) and other deep learning architectures has revolutionized the field. In pair with new data collection techniques, such as tagging individuals or collecting data self-supervised, it is possible to reach new heights in re-identifying performance.

This seminar will focus on reproducing the work of Chan et al. 2022. They investigated the impact of self-supervised learning for re-identifying honeybees and how bees can be re-identified using only the bee's abdomen. They developed a Convolutional Neural Network (CNN) to embed bees in a 128-dimensional Euclidian space. They used a triplet loss with semi-hard online mining to ensure small distances between images of similar bees and large distances between images of dissimilar bees. To show the impact of self-supervision, they pre-trained this network using a dataset that assumes the identity of bees based on the probability that no bee will enter the camera frame twice in a short period. The data was selected over a time frame of 12 days, so it captured variations such as lightning or the pose of the bee. However, the paper didn't examine the impact of selecting triplets in a way that the network learns to embed images with temporal differences close to each other. This could influence the hypothesis regarding the impact of self-supervised learning in their work.

2. Related Work

This section reviews the key contributions and methodologies in animal reidentification. An overview of traditional approaches and their limitations is provided first, followed by a discussion of contemporary deep learning-based techniques. Various datasets and benchmarks developed to facilitate research in this domain are then examined. Finally, some ongoing challenges and future directions in animal reidentification are highlighted, emphasizing the importance of interdisciplinary collaboration and the potential for cross-pollination with related fields such as human reidentification.

2.1. Dataset Collection

Collecting large, labelled datasets is critical for successfully applying deep learning techniques in reidentification tasks. However, acquiring such datasets poses significant challenges due to the inherent difficulty in capturing and labelling vast amounts of images. Various strategies have been employed to overcome these challenges.

Animal reidentification has historically relied on manual methods, including tagging and direct observation, which pose several limitations. Tagging, while useful, can be invasive and stressful for animals, potentially altering their natural behaviour. Although non-invasive, manual observation (Ardovini, Cinque, and Sangineto 2008) is labour-intensive, subject to human error and bias, and could introduce inaccuracies. For example, identifying individual whales from their unique markings and scars can be feasible for trained experts (Mizroch, Beard, and Lynde 1990), but this process is time-consuming and prone to subjective interpretation. These traditional methods struggle with scalability, especially for large populations or elusive species, and often fail to provide the continuous monitoring necessary for comprehensive behavioural studies and conservation efforts.

Recent approaches emphasize self-supervised data collection (Gao et al. 2021), which involves creating datasets that can be used for supervised learning without requiring manual labelling. These datasets hold potential for animal identification by enabling network pre-training. This allows the model to learn valuable features before fine-tuning with a manually labelled dataset. The effects of pre-training a network have been shown in numerous other fields, such as person re-identification (Fu et al. 2021) or animal tracking and detection (Yousra et al. 2023).

2.2. Re-Identification

The initial computational approaches in animal reidentification focused on species-specific methods utilizing handcrafted features and traditional machine learning algorithms. These techniques required substantial domain expertise to design effective features for different species, such as the nicks of elephants (Ardovini, Cinque, and Sangineto 2008) or the chest patterns of African penguins (Sherley et al. 2010). Handcrafted features, while effective to a degree, were limited by their reliance on specific, often subtle visual cues, making them less adaptable to varied environmental conditions and different species (Schneider, Taylor, Linquist, et al. 2019).

Moreover, the quality and consistency of labelled datasets significantly impacted the performance of these early methods. Observer bias (Tuyttens et al. 2014) and the inherent difficulty of accurately labelling large numbers of individuals further complicated the development of reliable models. For instance, manual labelling of honeybees in high-density populations is particularly challenging, often resulting in datasets comprising only a few individuals.

The emergence of deep learning, particularly CNNs, has dramatically transformed the field of animal re-identification. CNNs have demonstrated remarkable success in extracting hierarchical features directly from raw image data, reducing the need for handcrafted features and enabling more robust and scalable re-identification systems (Schneider, Taylor, Linquist, et al. 2019). State-of-the-art CNN architectures, such as ResNet (He et al. 2016a), have been effectively adapted for animal re-identification tasks, achieving significant improvements in accuracy and generalization across different species and environments (Schneider, Taylor, Linquist, et al. 2019).

More recent advancements in deep learning have introduced the use of Siamese networks (Schneider, Taylor, and Kremer 2020) and embedding architectures for identification (Schroff, Kalenichenko, and Philbin 2015. Siamese networks consist of two identical subnetworks that process two input images to produce comparable embeddings. The key idea is to minimize the distance between embeddings of the same individual and maximize the distance between embeddings of different individuals.

Embedding architectures have also been employed, transforming input images into a lower-dimensional space where similar images are closer and dissimilar images further apart. These architectures facilitate more efficient and accurate matching of individuals by comparing their embeddings using distance metrics rather than raw image data. Techniques such as triplet and contrastive loss are commonly used to train these networks, further enhancing their ability to distinguish between individuals.

3. Material and Methods

This section provides a detailed description of the study to be reproduced and examined in this seminar. First, the datasets, including the data collection process, are described and explained. Then, the training protocols and the evaluation procedure are detailed.

3.1. Dataset

The speciality of this dataset is the division into a 1) short-term dataset, which comprises a high number of individuals tracked during a short period, and a 2) long-term dataset, which comprises bees that were tagged using barcode tags and thus could be monitored for longer, as the barcode tags provided a ground truth for re-identifying the individual bees.

The extraction of individual images was conducted as follows. A video camera filmed the entrance of a colony at 20fps over a period of 12 days. Using the bee pose estimator published by Rodríguez et al. 2018, the skeleton of the bee, including the head, neck, waist and abdomen tip, was detected. As seen in figure 1 the detected reference points can be used to normalize the body of the bees and reduce the impact of the body position of bees, thus minimizing possible unwanted features that a Convolutional Neural Network would learn.

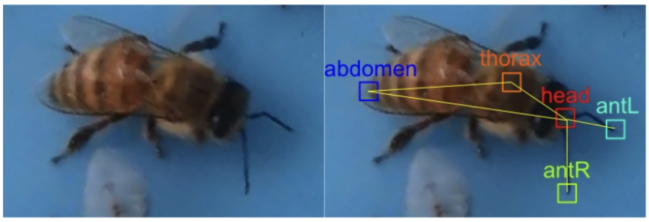

Figure 1: Estimation of the pose of the bee, in this case, abdomen, thorax, head, antL and antR, adopted from Rodríguez et al. 2018.

Based on the skeleton detection, individual bees were extracted from the images recorded by the camera. As seen in figure 2, some bees were occluded and could not be used for the dataset. Another challenge posed the curling of the bee's abdomen. Too much curling would lead to a deformation of the abdomen pattern. Thus, it would be hard to re-identify a bee if the abdomen is curled too much. Such occurrences were also filtered from the dataset.

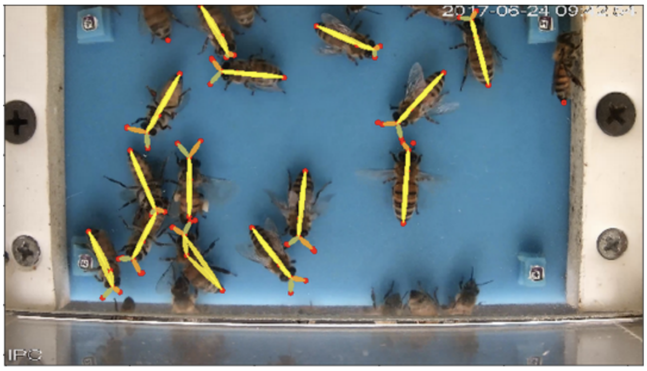

Figure 2: Beepose estimation of individual bees, adopted from Rodríguez et al. 2018.

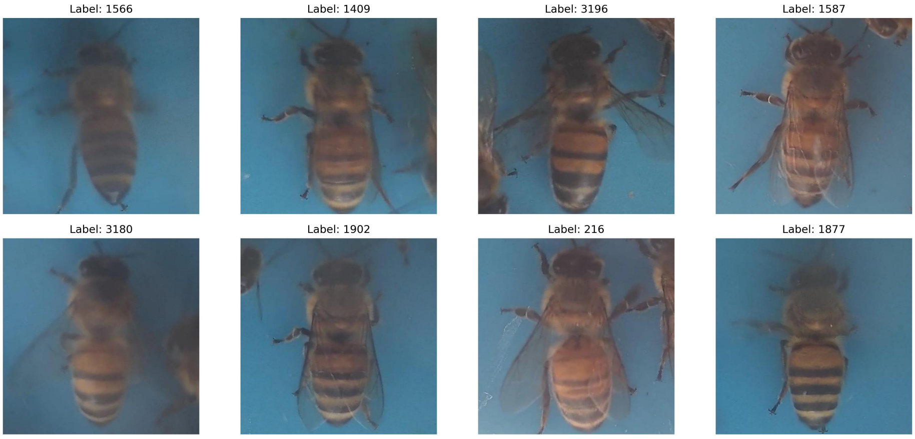

Figure 3: Randomly selected samples of the short-term dataset.

3.1.1. Short-Term Dataset

The short-term dataset comprises 109.645 images, divided into 4949 tracks with an average length of 22.15 images. These tracks were annotated with the respective track ID. The underlying assumption is that over a short period, in this case, 10 minutes, no bee will enter and exit the beehive again. This allows the collection of large amounts of training data, comprising many individuals. The low probability that two tracks are of the same bee is ignored in the construction of this dataset.

The dataset's purpose is solely training. Using many individuals during the network's training can help distinguish bees better in the feature space and help it learn more relevant features. Figure 3 shows an example of images in the dataset.

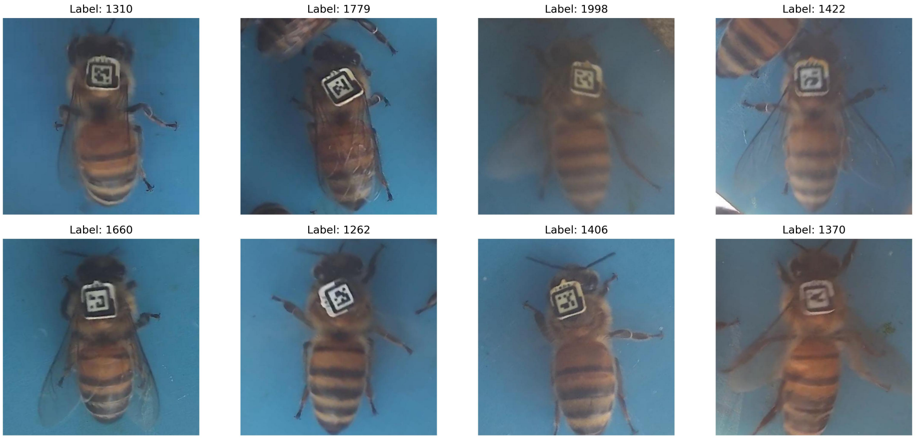

Figure 4: Randomly selected samples of the long-term dataset.

3.1.2. Long-Term Dataset

The long-term dataset is split into a training and validation set. The training set comprises 3.777 images of 181 individuals, and the validation set comprises 1.909 images of 66 individuals. The main difference from the short-term dataset is the missing 1:1 relationship between a track and an individual bee, meaning that this dataset contains multiple tracks of an individual bee. Thus, the variety of visual features is more expansive than in the short-term dataset, with differences in illumination, pose, and wings overlapping.

3.2. Network Architecture

Chan et al. 2022 employs a network that maps the provided cropped abdomen to a 128-dimensional feature vector. This feature vector will be referred to as the embedding in the rest of this work. Figure 5 displays the network's architecture. First, a 7x7 convolution is applied, followed by 3 ResNet full pre-activation units (He et al. 2016b). What is special about these units is that, in contrast to the common belief of applying the activation post calculations, they use the activation function, in this case, ReLU, before applying weights to the input. He et al. 2016b have shown that this helps to reduce the training loss faster. The L2 Normalization layer at the end ensures an embedding into a 128-dimensional Euclidean space. So the network learns a function $f : \mathbb{R}^f \rightarrow \mathbb{R}^d$, with $d = 128$. Hence, the similarity of vectors directly corresponds to the similarity of bees.

The network aims to output embeddings with a small Euclidian distance as defined in equation (1) into the embeddings from similar bees and a high distance to the embeddings of dissimilar

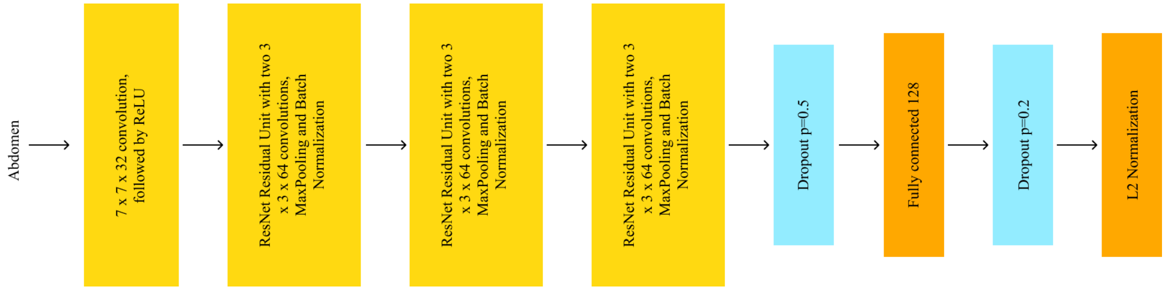
Abdomen
7 x 7 x 32 convolution, followed by ReLU
ResNet Residual Unit with two 3 x 3 x 64 convolutions, MaxPooling and Batch Normalization
ResNet Residual Unit with two 3 x 3 x 64 convolutions, MaxPooling and Batch Normalization
ResNet Residual Unit with two 3 x 3 x 64 convolutions, MaxPooling and Batch Normalization
Dropout p=0.5
Fully connected 128
Dropout p=0.2
L2 Normalization

Figure 5: Model architecture, adopted from Chan et al. 2022

bees.

$$EU_d = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{1}$$

This way, the network can re-identify bees, even if it hasn't seen them before. The objective loss function used by the network is the triplet loss. The goal is to minimize the distance between an anchor image x_i^a and a positive sample x_i^p and to maximize the distance between the anchor x_i^a and the negative sample x_i^n. Figure 6 shows this relationship. So, each triplet (x_i^a, x_i^p, x_i^n) should match the equation:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2, \quad \forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \tau \tag{2}$$

τ represents the set of all triplets that are possible in the training set. Based on equation 2, the objective function the network tries to minimize is:

$$\sum_{i}^{N} [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha], \quad N = \#\tau \tag{3}$$

The triplet selection technique used by Chan et al. 2022 was first described by Schroff, Kalenichenko, and Philbin 2015, and is called *semi-hard online mining*. Generally, given the anchor sample x_i^a the goal is to select the positive sample x_i^p such that $argmax_{x_i^p}\|f(x_i^a) - f(x_i^p)\|_2^2$ and the negative sample x_i^n such that $argmin_{x_i^n}\|f(x_i^a) - f(x_i^n)\|_2^2$. We refer to these samples as the hardest positive and the hardest negative. Selecting these samples from the whole training set is infeasible, as images with bad quality or falsely labelled images would have the highest chance of being the hardest negative. To mitigate this issue, we can select the triplets per batch, this is called online mining. It means that instead of sampling triplets from the complete training data, the triples are mined from mini-batches, in the case of this seminar, from a batch with size 256. Always selecting the hardest negatives in these batches can, in practice, lead to collapsing models and too early local minima in the training process (Schroff, Kalenichenko, and Philbin 2015). It can help to select the triplets, such that:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2 \tag{4}$$

These negatives need to lie inside the margin α and are called semi-hard, as the distance to the anchor image is close to the distance between the anchor and the positive image.

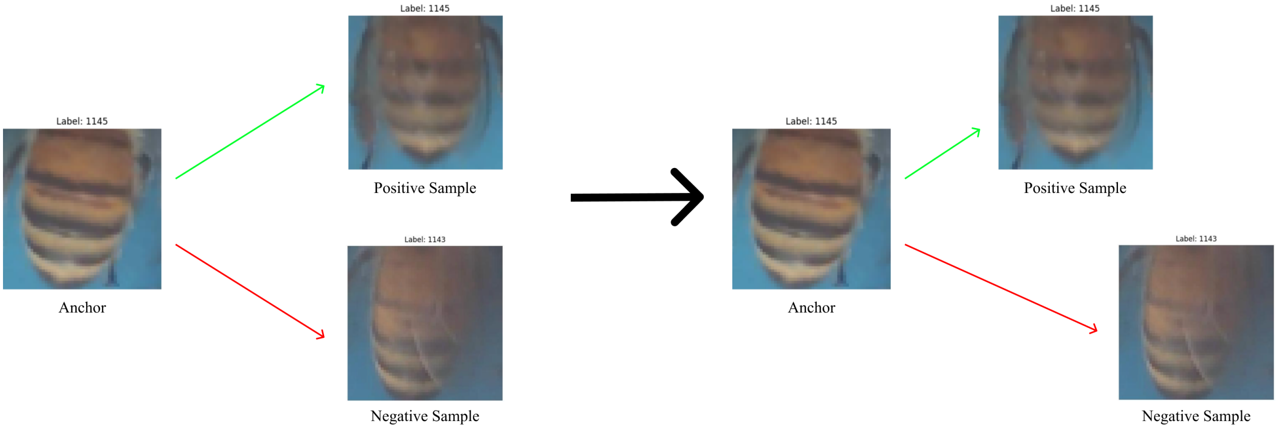

Figure 6: Triplet Loss - The goal is to minimize the norm that is used to calculate the distance between anchor and positive sample (green arrow) and to maximize the distance between anchor and negative sample (red arrow)

3.3. Training Protocols

To measure the impact of self-supervised learning in bee re-identification, Chan et al. 2022 used three training protocols:

1. *Fully-supervised*: Only the long-term dataset is used for training.
2. *Self-supervised*: Only the short-term dataset is used for training.
3. *Supervised + Pretraining Protocol*: The short-term dataset is used for pre-training and the long-term dataset for fine-tuning.

For each protocol, the Triplet Semi Hard Loss (He et al. 2016b) is the objective function to minimise, with a margin α of 0.2. For optimizing, Adam (Kingma and Ba 2014) with a learning rate of $1e^{-03}$ is used. The training was done for a maximum of 1000 epochs, using an early stop mechanism, monitoring the validation loss with a patience of 100 epochs. Early stopping ensures that the network is not overfitting on the training data. Thus, ensuring that the network is still able to generalize.

3.4. Evaluation Procedure and Metrics

3.4.1. Cumulative Matching Characteristic

The evaluation metric used by Chan et al. 2022 is the Cumulative Matching Characteristic (CMC). It is a comprehensive metric often used in human or animal identification as, unlike simple accuracy metrics, the CMC curve considers the ranking of matches. CMC rank k gives the probability that the desired sample was ranked at rank k within the gallery of samples. This is more practical than just having a single prediction because these rankings can also be used to understand the features that the classifying mechanism uses to discriminate samples. Constructing a CMC curve requires a dataset with labelled query and gallery samples. The respective procedure is explained in the next chapter.

3.4.2. Evaluation Setups

For the evaluation Chan et al. 2022 developed three scenarios with increasing difficulty. The *Same-day same-hour* evaluation setup samples track pairs with a margin of 15 minutes to 60 minutes of capture time. This is the easiest test setup, with less lighting and pose variation. Examples of sampled track pairs are displayed in figure 7.

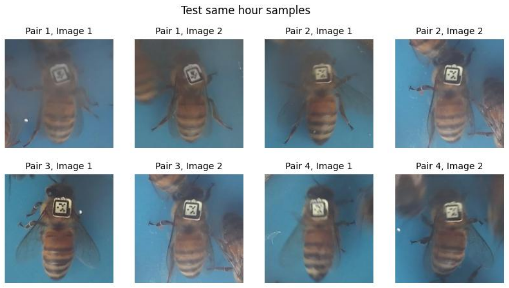

Figure 7: Test same-day same-hour: This is the easiest of the tests, as there is no lighting variation.

The *Different-day same-hour* evaluation setup samples track pairs from different days, but the time of the day is less than 60 minutes apart. Examples of sampled track pairs are displayed in figure 8.

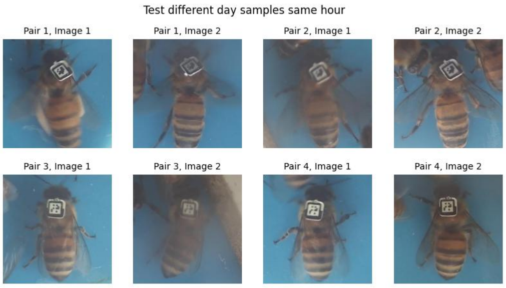

Figure 8: Test different-day same-hour

The *Different-day any-hour* evaluation setup samples track pairs from different days. No constraints for the time of the day difference are set. This is the most complex evaluation setup, capturing most variations, as seen in figure 9.

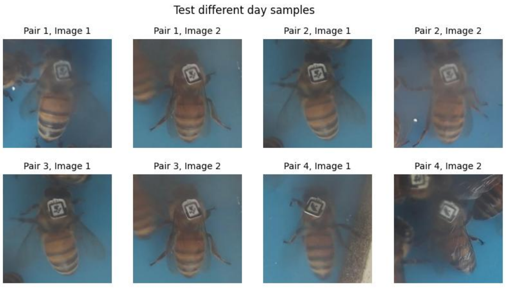

Figure 9: Test different day same hour

For each sampled track pair, an image was randomly sampled from the track. This process was repeated 100 times to generate an image query for evaluation. Additionally, image distractors were sampled from all negative IDs for each query.

4. Results

This section will present the results from reproducing the work of Chan et al. 2022. First, the baseline of their work is presented, and then the results obtained in this seminar are described. Results were obtained using an I5-9600k with an RTX 2070 Super GPU. The network was written using Pytorch Lightning instead of the author's choice of Tensorflow/Keras.

4.1. Baseline

The best-performing protocol was the Short-term + Long-term protocol, with a rank-1 score of 0.801 for same day, same hour and 0.659 and 0.624 for the other evaluation protocols. Unfortunately, they only present the rank-1 and rank-3 results, as seen in table 4.1. This makes it infeasible to compare the CMC curves. However, it's still a good indicator of performance. Due to limitations in computation capacity, this seminar will focus on reproducing the work without data augmentation.

Method	**Training Protocol**	**Same day, same hour**		**Diff day, same hour**		**Diff day, any hour**	
		Rank 1	**Rank 3**	**Rank 1**	**Rank 3**	**Rank 1**	**Rank 3**
Triplet loss, No Aug	Long-term	0.456	0.733	0.322	0.610	0.273	0.557
Triplet loss, Aug	Long-term	**0.529**	**0.781**	**0.391**	**0.709**	**0.362**	**0.664**
Triplet loss, No Aug	Short-term	0.682	0.888	0.508	0.775	0.436	0.720
Triplet loss, Aug	Short-term	**0.738**	**0.912**	**0.579**	**0.831**	**0.498**	**0.788**
Triplet Loss, No Aug	Short-term + Long-term	**0.801**	**0.932**	**0.659**	**0.889**	**0.624**	**0.868**
Triplet Loss, Aug	Short-term + Long-term	0.759	0.913	0.651	0.880	0.586	0.832

Table 1: Comparison of performance metrics for different training protocols and data augmentation techniques (Chan et al. 2022)

The reproduction results are nearly identical for the short-term and short-term + long-term training protocol. For the short-term protocol, the same-day same-hour test results were slightly better, with an improvement of 0.012 for CMC-rank 1. However, the different-day same-hour setup scored 0.052 less and the different-day any-hour setup 0.049 less for CMC-rank 1. The short-term + long-term training protocol achieved lower CMC-rank 1 scores for all evaluation setups, respectively 0.004, 0.014 and 0.037. Surprisingly, the results of the long-term protocol were better for all evaluation protocols by margins of 0.052, 0.066, 0.028. The respective results can be seen in figure 10. They don't indicate the big difference between short and long-term datasets reported by Chan et al. 2022. Thus making the short-term dataset superior in capturing relevant variations questionable.

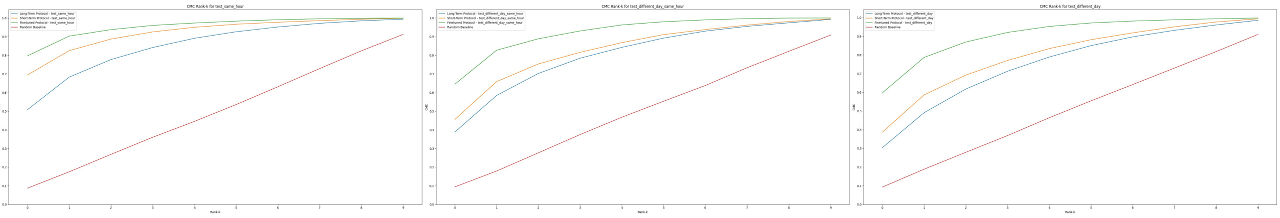

Figure 10: Baseline reproduction results, with a random evaluation protocol as a comparison. Zoom in for further details.

4.2. Triplet Selection

When applying the semi-hard triplet loss, triplets are sampled *online* per batch. So, based on the sorting of the original dataset, the respective positive sample could not be hard enough to learn effectively. It seems like Chan et al. 2022 do not shuffle their data before training the network. Thus keeping images of the same track in the same batch. This leads to the selection of easy triplets, as the images were taken in a short time period, thus capturing a low level of variance in pose and illumination. Figure 11 shows an example of positives selected in the same batch for the long-term training protocol.

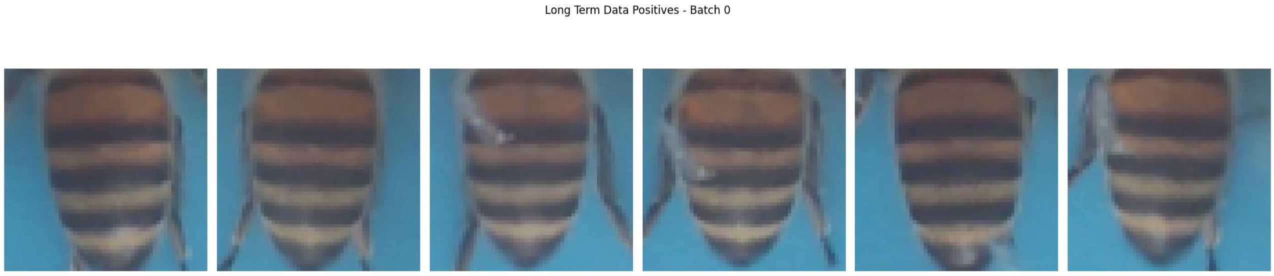

Figure 11: Example of images with the same label in a batch of the long-term training dataset. It can be seen that they capture a low level of variance in pose, illumination and abdomen curling.

It was found that shuffling the data before sampling pairs for the triplet loss could drastically improve the CMC scores. While this approach's randomness introduces missing interpretability and reproducibility, it can provide indicators for possible improvements in the triplet sampling technique. Figure 12 shows an example of positive samples within a batch if shuffling is applied. There are more variations in pose, illuminance and abdomen curling. This enables the network to better adapt to variations in the data.

Shuffling the data didn't drastically affect the short-term and short-term + long-term training protocol, so the following will focus on the long-term training protocol. The best result scored a

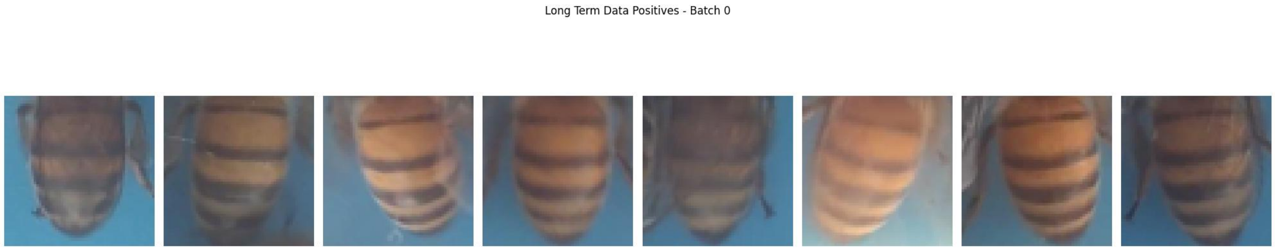

Figure 12: Example of images with the same label in a batch of the shuffled long-term training dataset. It can be seen that they include more variation in pose, illumination and abdomen curling.

CMC rank-1 of 0.569, 0.486, and 0.3995 for the respective evaluation setups, with CMC rank-3 scores of 0.8388, 0.768, and 0.695. Especially for the last two evaluation setups, the long-term training protocol even outperformed or was on par with the reproduction results of the short-term training protocol and outperformed the results with augmented data as presented in table 4.1.

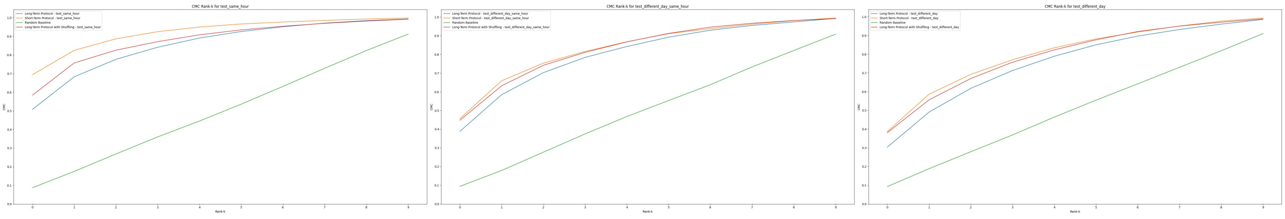

Figure 13: Test results when shuffling the data before creating batches. Thus ensuring that there is a longer time distance between captures. Zoom in for further details.

4.3. Impact of Number of Distractors

As explained in section 3.4, samples are ranked based on a query with a set of distractors. In the case of Chan et al. 2022, each query contained 10 distractors. However, the number of distractors has a significant impact on the results of the study. Therefore, the difference in CMC scores with an increasing number of distractors needs to be investigated. Due to computation and memory capacities, only 10 and 50 distractors were investigated, and only the evaluation protocols *same day same hour* and *different day same hour* can be compared. Still, the setup using 50 distractors showed a significant decrease in cmc-scores, indicating that the presented results might be glossed over. For the long-term protocol, the CMC rank-1 scores declined from 0.569 to 0.378 and from 0.486 to 0.2165, respectively. While the practice of using 10 distractors helps to reproduce the results and provides a common ground to evaluate for any dataset size, it can also leave false impressions, especially when comparing against other papers.

5. Discussion

The reproduction of the work by Chan et al. 2022 has led to several insights that will be further discussed in this chapter. Additionally, the limitations of the reproduction will be discussed.

The results obtained by shuffling the dataset, so increasing the temporal difference of track pairs, indicated that selecting triplets with more problematic positives improves the training performance. Always choosing the hardest positive is infeasible, as this introduces learning blurred images or images with a substantial wing overlap. Unfortunately, no systematic way of selecting triplets for this dataset can be presented; this makes the exact reproduction of the results in this seminar paper hard. However, it is a starting point for future work that could try to solve the triplet selection bias as presented by Yu et al. 2018. Also, the results contradict the paper's conclusion by Chan et al. 2022, which stated that short-term training ultimately outperforms long-term training. However, this was not the case for those evaluation setups that included more considerable time variations. This leads to the question of whether the long-term dataset with an optimal triplet selection could outperform the short-term training protocol. Also, it indicates that the short-term protocol discriminates based on features outside of the bee itself. Another trade-off, not mentioned by Chan et al. 2022 is the effectiveness of the dataset in terms of training time and memory usage. The short-term dataset comprises around 33 times the amount of images, thus increasing the total training time and total memory usage considerably without yielding a significant increase in re-identification. It remains questionable whether tracking the abdomen is feasible for re-identifying bees. The curling of the abdomen greatly impacts the re-identification performance and can not be avoided in real-world re-identification scenarios. Also, the presented results by Chan et al. 2022 were achieved using only 10 distractors, which is infeasible for a real-world scenario, whereas, in an exemplary world where a beehive only has 191 individuals, re-identification would need to be done based on those 191 distractors, i.e. means of the respective embeddings. Increasing the distractors to 50 in the test setup yielded worse results, indicating the infeasibility of a real-world setup. Figure 14 highlights some of the difficulties of reidentifying bees, such as curlings changing the pattern of the abdomen, different levels of blurriness, or orientations of the bee.

6. Conclusion

This seminar presents a reproduction of the paper published by Chan et al. 2022. The baseline results could be reproduced, and a way to improve the results for the long-term dataset is shown. Ultimately questioning the hypothesis of the impact of the short-term dataset and leaving room for further investigations. The results indicate that the bee abdomen is not a valuable feature in re-identifying bees due to biodiversity factors and variations such as curling of the abdomen, which heavily impacts the displayed pattern. This work highlights the importance of labelled datasets that collect drastic variations of the same individual over time. Even though it's invasive, future work could focus on collecting a dataset of bees over a longer time frame than 12 days, capturing more diversity of the same individual. Also, the data quality could be improved by using cameras capable of tracking more frames per second and higher shutter speeds, thus reducing the blurriness of images and potentially relevant features for identification. Future work should also investigate a systematic triplet selection technique yielding the best results. This could be done following the approaches presented by Yu et al. 2018 or Shi et al. 2016.

A. Appendix

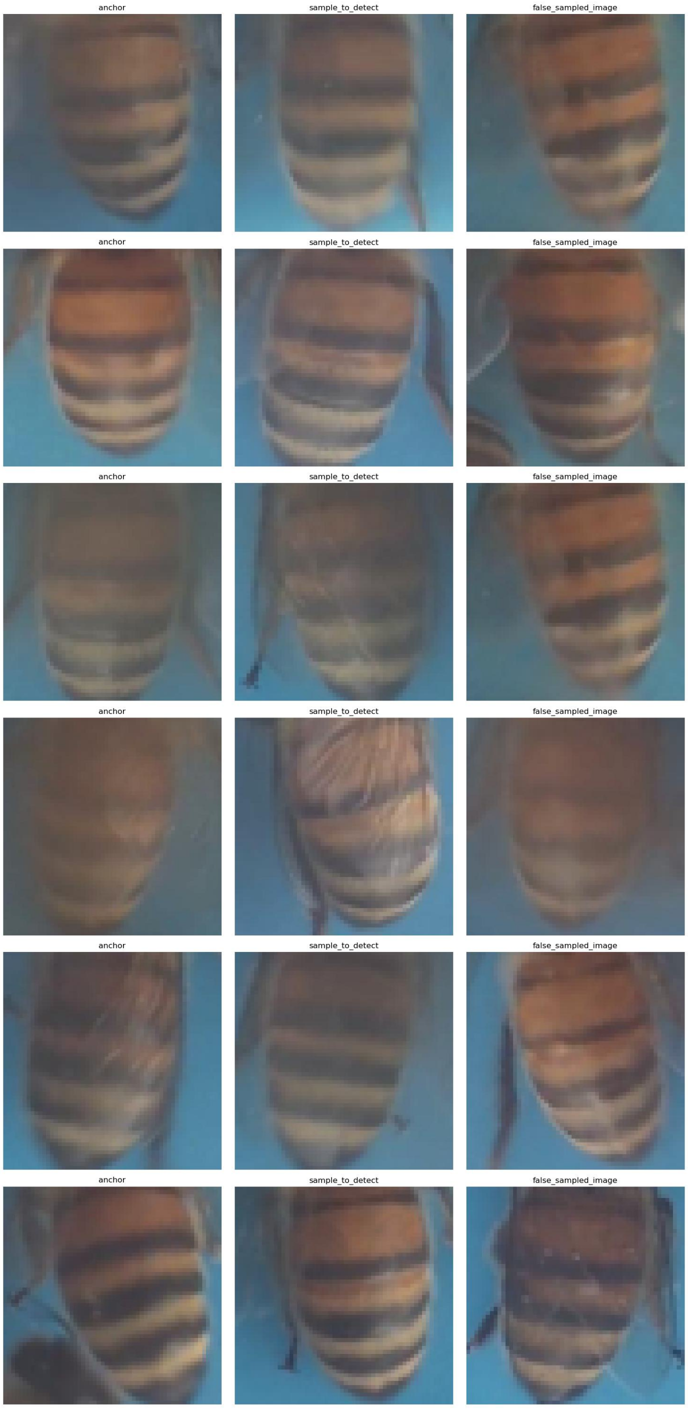

Figure 14: Examples of false detections of the long-term training protocol.

List of Figures

References

Ardovini, A., L. Cinque, and E. Sangineto (2008). "Identifying elephant photos by multi-curve matching". In: *Pattern Recognition* 41.6, pp. 1867–1877. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2007.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0031320307005031.

Borlinghaus, Parzival, Frederic Tausch, and Luca Rettenberger (2023). "A Purely Visual Re-ID Approach for Bumblebees (Bombus terrestris)". In: *Smart Agricultural Technology* 3, p. 100135. ISSN: 2772-3755. DOI: https://doi.org/10.1016/j.atech.2022.100135. URL: https://www.sciencedirect.com/science/article/pii/S2772375522000995.

Chan, Jeffrey et al. (2022). "Honeybee Re-identification in Video: New Datasets and Impact of Self-supervision." In: *VISIGRAPP (5: VISAPP)*, pp. 517–525.

Foster, Rebecca J and Bart J Harmsen (2012). "A critique of density estimation from camera-trap data". In: *The Journal of Wildlife Management* 76.2, pp. 224–236.

Fu, Dengpan et al. (2021). "Unsupervised pre-training for person re-identification". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14750–14759.

Gao, Jing et al. (2021). "Towards self-supervision for video identification of individual holstein-friesian cattle: The cows2021 dataset". In: *arXiv preprint arXiv:2105.01938*.

He, Kaiming et al. (2016a). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

– (2016b). "Identity Mappings in Deep Residual Networks". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, pp. 630–645. ISBN: 978-3-319-46493-0.

Kelly, Marcella J. (May 2001). "Computer-Aided Photograph Matching in Studies Using Individual Identification: An Example from Serengeti Cheetahs". In: *Journal of Mammalogy* 82.2, pp. 440–449. ISSN: 0022-2372. DOI: 10.1644/1545-1542(2001)082<0440:CAPMIS>2.0.CO;2. eprint: https://academic.oup.com/jmammal/article-pdf/82/2/440/7022858/82-2-440.pdf. URL: https://doi.org/10.1644/1545-1542(2001)082%3C0440:CAPMIS%3E2.0.CO;2.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Krebs, Charles J. (1989). *Ecological methodology.* 2nd. Technical report. New York, NY: Harper & Row New York.

Mizroch, Sally, J.A. Beard, and M. Lynde (Jan. 1990). “Computer assisted photo-identification of humpback whales”. In: *Report of the International Whaling Commission*, pp. 63–70.

Rodríguez, IF et al. (2018). “Honeybee detection and pose estimation using convolutional neural networks”. In: *Congres Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP).*

Schneider, Stefan, Graham W Taylor, and Stefan C Kremer (2020). “Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer”. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pp. 44–52.

Schneider, Stefan, Graham W Taylor, Stefan Linquist, et al. (2019). “Past, present and future approaches using computer vision for animal re-identification from camera trap data”. In: *Methods in Ecology and Evolution* 10.4, pp. 461–470.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Sherley, Richard B et al. (2010). “Spotting the difference: towards fully-automated population monitoring of African penguins Spheniscus demersus”. In: *Endangered Species Research* 11.2, pp. 101–111.

Shi, Hailin et al. (2016). “Embedding deep metric for person re-identification: A study against large variations”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14.* Springer, pp. 732–748.

Tuyttens, F.A.M. et al. (2014). “Observer bias in animal behaviour research: can we believe what we score, if we score what we believe?” In: *Animal Behaviour* 90, pp. 273–280. ISSN: 0003-3472. DOI: https://doi.org/10.1016/j.anbehav.2014.02.007. URL: https://www.sciencedirect.com/science/article/pii/S000334721400092X.

Yousra, Taifour et al. (2023). “Self-supervised Animal Detection in Indoor Environment”. In: *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6. DOI: 10.1109/IPTA59101.2023.10320047.

Yu, Baosheng et al. (2018). “Correcting the triplet selection bias for triplet loss”. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 71–87.

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Christopher

Karlsruhe, den July 15, 2024