Self-supervised Pose Representation Learning For Nonhuman Primates

Kenza QITOUT[1], Mathilde ADET[2], Xavier DESQUESNES[1], Audrey MAILLE[2], and Bruno EMILE[1]

[1] Laboratoire PRISME, Université d'Orléans, INSA CVL, Orléans, 45072, France
[2] Laboratoire Eco-anthropologie, UMR 7206, MNHN, CNRS, Université Paris Cité, Paris, 75016, France

Abstract. Self-supervised image representation learning through contrastive learning is highly effective for extracting robust representations in various contexts, and leads to efficient applications in object classification from unlabeled or poorly labeled data. In this article, we investigate the potential of such self-supervised learning strategy for global body pose representation with unannotated data. Using contrastive learning models, we tested several strategies for data comparison, data augmentation, and loss functions in order to improve both representation quality and temporal coherence across consecutive images from videos, aiming to construct a robust embedding space. We assess a new self-supervised approach, which exploits the temporal similarity that may exist between two successive video frames to extract pose information weighted by the motion differences between the two frames. We also present a comparative study on a new annotated dataset of poses of Barbary macaque (*Macaca sylvanus*) on video frames collected in a zoo. We obtained an accuracy of 0.97 with BYOL model for image classification task.

Keywords: Contrastive Learning · Pose Representation · Motion-Based Loss · Temporal Similarity.

1 Introduction

Pose and action recognition is a broad research field with applications in neuroscience, engineering, security, sports, and medicine [22]. It enables images and videos to be analysed to identify an individual's body position and movements in space. While pose recognition in humans has been widely studied [13], recognition of animal poses has received less attention. Studying animals is indeed a significant challenge due to the difficulty of observing them continuously and collecting usable information. In this context, the study of nonhuman primates that share morphological characteristics with humans is particularly relevant, especially the analysis of poses, as they can provide important information for action recognition and analyses of complex behaviors, including social interactions [1].

Supervised methods such as pose-estimation models are highly effective for pose recognition. They eliminate the need for heavy material or human resources,

as well as physical markers that could modify behaviours. Indeed, these models enable automatic, non-invasive, and low-cost analysis of visual data in different environments, as well as the extraction of information related to global body pose. However, supervised models have limitations: they require large amounts of annotated data, which is time-consuming, prone to errors, subjective, and they may lack precision. Moreover, most of the pose-estimation models are designed to reconstruct the skeleton of the observed individual [13], which forces the representation into that specific structure even though it is not necessarily optimal. Our goal is to create a system that provides an immediate representation of global body poses using a model that learns its own relevant representations, avoiding intermediate steps or imposed structures. We focus on representing the diversity of global body poses from images in a latent space that groups similar poses together and pushes apart dissimilar ones, while preserving a gradual progression between poses to capture the full continuum from extreme to intermediate poses, despite limited annotated data.

Self-supervised learning is a promising alternative to the limitations of supervised approaches. This method requires no prior data annotation, as the model generates its own supervisory signals from raw data. It allows pose representations to be learned without large annotated dataset, reducing the time and cost of annotation while remaining robust to variation. Among self-supervised methods, the contrastive learning approach seems particularly suitable for learning visual representations [12],[23]. Recent studies used ordinal contrastive learning to represent samples while taking into account an ordinal ranking. Ranks preserve the order relationship between samples and guide embeddings to reflect this order in the latent space: the smaller the rank difference between two samples, the more similar they are. These ordinal levels can come from various sources, such as spectral ordering [5] or spatio-temporal data [6], and are used to weight the loss function and smooth transitions between classes. Contrastive learning models generally compare a pair of views transformed from an anchor image. However, since these views are artificially generated, performance can vary depending on the transformations applied [3]. Moreover, contrastive learning models typically use positive (and negative) samples to distinguish similar elements from different ones. However, as body poses can vary greatly between and within individuals, representing poses through binary or discrete contrast tends to minimize variation.

One approach to pose representation is to adopt a continuous context rather than a binary or discrete context in order to capture the diversity and spectrum of poses. For this, we adapted self-supervised methods of contrastive learning in order to compare multiple real-world views without depending on the transformations applied to an anchor image to generate images considered as similar, while accounting for natural variation between these images. Using images from the same video sequence has the advantage of naturally generating image variations comparable to data augmentation. Moreover, videos are a rich source of unannotated visual data with temporal continuity. Poses in a video are linked in a particular sequence to form an action, which can be represented as a succession

of poses that differ progressively as the temporal distance increases. We therefore based our work on the temporal similarity among the poses of an individual within a video sequence.

In this article, we (1) test existing contrastive learning methods for extracting representations of global body poses from unannotated images, (2) modify the data comparison strategy by using real, temporally adjacent images, (3) test the effect of a contrastive loss based on motion, and (4) evaluate these models on an annotated global body pose dataset, describing how we constructed it.

2 Related Work

2.1 Contastive Learning

The principle of contrastive learning is to bring similar images closer together while increasing the distance between different images through their representations in a latent space. There is a diversity of contrastive learning models with different architectures and data comparison strategies [12] [20]. In this study, we focused on the two contrastive learning models most widely used in the literature, SimCLR [3] and BYOL [10], which show strong performance across a wide range of applications.

SimCLR model [3] is a classic contrastive learning model [20] with a two-part architecture: an encoder (ResNet) and a projection head (MLP). This model generates n augmented views (generally 2 views) in the training dataset through data augmentation, treating them as positive pairs. The model maximizes similarity between these views in the latent space while pushing apart representations of other images considered negatives. For a batch of N images, there are 2N views: each view is compared to its positive sample and 2N-2 negatives. This contrastive learning strategy prevents representation collapse and enables effective learning of discriminative features. This model achieves very strong performance with 76.5% top-1 accuracy on ImageNet, but it requires large batch sizes to generate enough negative samples. Alternatives models are being developed such as MoCo model [11] with a queue and a momentum encoder and NNCLR model [9] matching views to nearest neighbors in a memory bank. Other models perform training without negative samples [10], [4], [2].

BYOL model [10] has a specific architecture with an online network (encoder, projector, predictor) and a target network (encoder, projector). Similar to SimCLR, it generates two augmented views of an image with data augmentation. Each network produces two slightly different representation of the same view. The model encourages predictions and projections to be closer in the latent space by updating the target network with an exponential moving average (EMA). This dual-network design stabilizes training and prevents model collapse without relying on negative sample comparisons. Among other developed models, SimSiam model [4] uses a stop-gradient instead of EMA and SwAV model [2] creates prototypes and maximizes the agreement between augmented views with the same prototype.

2.2 Temporal Contastive Learning

Contrastive learning models for video representation apply same strategies as image representation models [16], [8]. Most of these models maximize the similarity between two nearby clips from the same video while minimizing the similarity with clips from other videos [17], [23], [21], [7]. The CVRL model [17] and SimCLR model follow this strategy by bringing temporally close video clips from the same video closer together. The VideoMoCo model [16] additionally uses a memory queue to include negative clips from other videos. Temporal similarity has been primarily used for video representation [7] (but see applications to image representation: [23], [8], [14]).

The TempCLR approach [23] includes a temporal contrastive learning model based on temporal consistency [14]. The principle of this model is to extract images from a video: for an anchor image x_i, a temporal window T_i is defined around x_i. Images within the temporal window are considered positive samples, while images outside the window are considered negative samples.

2.3 Loss Functions

Several loss functions are used in contrastive learning models to train models without supervision [20]. The most common are the contrastive loss, the InfoNCE loss (Information Noise-Contrastive Estimation) or the NT-Xent loss (Normalized Temperature-Scaled Cross-Entropy).

The NT-Xent loss, which is widely used notably in the SIMCLR model and inspired other models [3], [11], [16], [17], [23], [9], [8], aims to bring positive pairs closer together than negative pairs.

$$\ell_{i,j} = -\log \frac{\exp\big(\mathrm{sim}(z_i, z_j)/\tau\big)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\big(\mathrm{sim}(z_i, z_k)/\tau\big)} \tag{1}$$

where samples i and j are a positive pair, $\mathrm{sim}(z_i,\ z_j)$ is the cosine similarity between i and j representations and τ is the temperature parameter. Minimizing this loss increases similarity for positive samples while reducing similarity for negative ones, preventing collapse and promoting discriminative representations.

The mean squared error (MSE) is used in the BYOL model to minimize the euclidean distance between the prediction p_i and the projection z_j:

$$\mathcal{L}_{i,j} = \|p_i - z_j\|_2^2 = 2 - 2 \cdot \frac{\langle p_i, z_j \rangle}{\|p_i\|_2 \cdot \|z_j\|_2} \tag{2}$$

where ||u|| is the norm of u and <u,v> is the dot product between u and v.

The triplet loss function [19] minimizes the distance between an anchor and its positive sample while maximizing the distance to a negative sample using a margin. This encourages positive samples to be brought closer in the representation space while pushing dissimilar ones further apart. Other approaches include geometric loss [15], temporal-discriminative loss [21], combination of invariance, variance, and covariance in the loss, as well as global and local temporal contrastive losses [7].

3 Method

This study aims at representing images continuously by considering both temporality and motion quantity, in order to capture the inherent continuity between poses and thus provides a more coherent and realistic modeling of motion sequences. This work contributes to (i) extend the strategies of existing contrastive learning approaches by incorporating temporal similarity into the data comparison strategy and (ii) introduce a motion-based loss function (Figure 1).

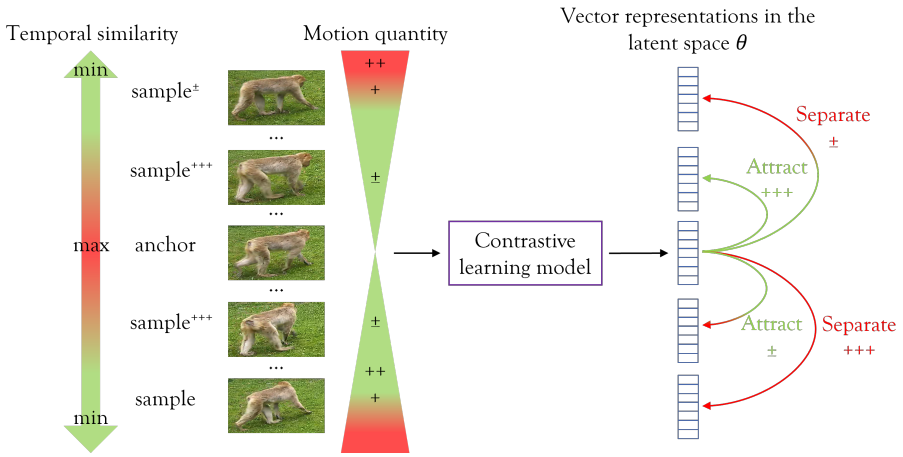

Fig. 1. Overview of the proposed framework based on temporal and motion similarity

3.1 Temporal similarity

The objective is to bring closer images of individuals with similar poses and to push apart those with dissimilar poses. We estimate that there is a temporal similarity among the poses of an individual within a video sequence, where the pose at time t is probably similar to those at $t+1$ and $t-1$, and different from those at $t+n$ and $t-n$ (for larger values of n).

As the model TempCLR [23], we used a data comparison strategy based on the temporal dimension between images in a video. Images from the same video are considered positive samples, with k positive pairs forming the mini-batches and compared to pairs from other videos treated as negatives. Unlike standard contrastive approaches that artificially generate positive views through data augmentation, our approach uses naturally occurring more or less positive samples within video sequences to capture natural variations between consecutive images. Images in each positive pair can be transformed with the same random geometric transformations to preserve temporal consistency in motion between images. Moreover, different random color transformations were applied to each image in the pair.

In addition to temporal proximity, incorporating motion information is useful for increasing robustness and improving the grouping of images showing different

poses but considered similar. This approach allows the model to better capture the intrinsic variability present in real image sequences. Positive samples are associated with a term representing motion between images within a sequence, allowing us to move from a binary to a continuous system and for the model to take into account varying degrees of similarity from natural differences between images and not just whether images are similar or not.

3.2 Motion-Based Loss function

We introduce a novel loss function based on motion values to better preserve continuity in image representation. Our approach preserves the architecture of existing contrastive learning frameworks while introducing an additional regularization term to model pose continuity. The objective is to maintain the proximity of representations corresponding to images from the same video in the latent space, proportionally to the observed motion quantity between these images.

The proposed method reduces the distance between positive samples from the same video (without overlap) and between groups of similar images across different videos, ensuring that similar poses are mapped to consistent regions of the latent space. Importantly, the distance between positive samples does not converge to zero but is modulated by a motion-dependent value, penalizing excessive similarity and preserving ordinal temporal relationships among representations. This constraint ensures temporal coherence and continuity by maintaining the sequential structure of motion stages within the latent space.

Our approach combines the baseline contrastive loss used to discriminate positive and negative samples with a motion-based term acting on positive pairs. This term constrains the distances between consecutive image representations proportionally to the motion quantity, thereby promoting a temporally coherent, structured embedding space. The proposed regularization can be seamlessly integrated into various contrastive learning loss functions.

For loss computation, we employed the euclidean distance, normalized to the range $\in [0, 1]$, where smaller values indicate higher similarity. For two positive samples (x_i, x_j), the euclidean distance is defined as $d(x_i, x_j) = \|x_i - x_j\|_2$. We enforce proportionality between this distance and the motion quantity φ_{ij}, such that for any positive pair (x_i, x_j) and (x'_i, x'_j) with motion values φ_{ij} and $\varphi_{i'j'}$, the ratio $\frac{d_{ij}}{\varphi_{ij}}$ remains approximately constant across pairs: $\frac{d_{ij}}{\varphi_{ij}} \approx \frac{d_{i'j'}}{\varphi_{i'j'}}$. We minimize the proportionality ratio variance across all positive pairs: $\mathrm{Var}\left(\frac{d_{ij}}{\varphi_{ij}}\right)$. We define the Motion-Based loss like:

$$\mathcal{L}_{\text{Motion-Based}} = \mathrm{Var}\left(\frac{d_{ij}}{\varphi_{ij} + \delta}\right) \tag{3}$$

where $\delta > 0$ is constant to ensure numerical stability and maintain a minimal separation between positive samples. For any given model, the new loss function is defined by the following formula with a weighting parameter $\alpha \in [0, 1]$ that controls the relative contribution of the Motion-Based loss:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{Model}} + (1 - \alpha) \cdot \mathcal{L}_{\text{Motion-Based}} \tag{4}$$

4 Experiments

The numerical calculations were carried out in part on the machines of the Centre de Calcul Scientifique en région Centre-Val de Loire using two GPU NVidia Tesla V100 with 32 GB and on local server using two GPU NVIDIA TITAN RTX with 24 GB. For the training, we chose optical flow to represent the motion quantity, which is inversely proportional to temporal similarity as it represents the amount of motion between two image pixels by calculating the apparent pixel displacement between two images.

4.1 Dataset Construction

We created large, diverse and non-redundant datasets containing a wide variety of poses to train each model. As video sequences present a variety of individual movements without local displacement, avoiding background bias, we assumed that two similar poses should be recognized as such, even if the background differs.

We used a dataset of 1,344 videos, each lasting between 30 minutes and 1 hour (for a total of 794 hours), recorded in the Barbary macaques' enclosure at ZooParc de Beauval, covering approximately 600 $\mathbf{m^2}$ and hosting between 33 and 37 individuals depending on natural births ans deaths. For each video, we extracted one image per minute and a 5-second video sequence, starting from this image only if it differed from the last extracted image (the videos have an average FPS of 20); images were considered different when their optical flow was above 1.0. We considered a 1-minute interval as the minimum time to capture individuals in motion without introducing excessive variation.

We detected individuals in the sequences with a YOLO object detection model pre-trained on similar data with confidence score of 0.7. We tracked individuals with SORT and extracted the video sequences centered on each individual to avoid background effects. Each selected video was longer than 30 images and without local spatial displacement (Figure 2). We smoothed tracking by taking the maximum coordinates enclosing all bounding boxes.

Images were extracted from the bounding boxes by modifying their coordinates, taking the smallest square enclosing the bounding box with a 10% buffer relative to the total image size. All images were resized to 128 $\times$ 128 pixels to standardize image size. To ensure that the individual shows motion even without spatial displacement, we selected video sequence with a maximum optical flow above 0.07. This threshold was chosen through visual inspection and by considering the minimal motion to preserve. From all images in each individual-centered videos, we created three datasets by keeping images with an optical flow above a threshold (0.5, 0.75, 1.0) compared to the last kept image. Only videos with at least two images were selected to ensure that at least one valid image pair

could be generated per video. Datasets contain a total of 21.509, 11.011 and 6.461 images for optical flow thresholds of 0.5, 0.75 and 1.0, respectively.

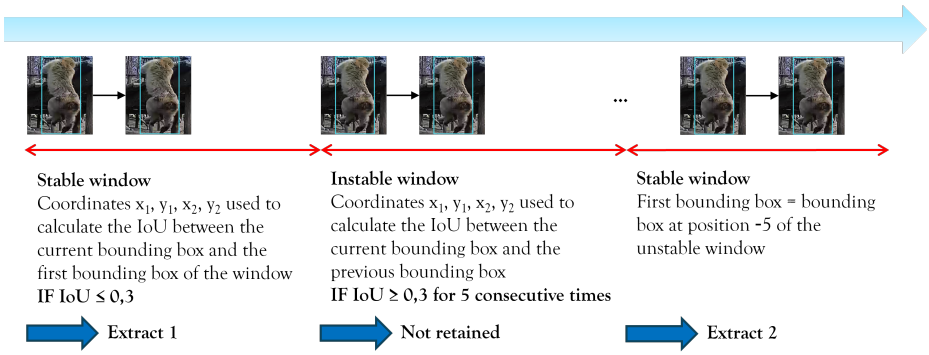

Fig. 2. Stable and moving windows determined by IoU metric. From the first bounding box, we calculated the IoU with the box of the next image. If IoU ≤ 0.3, the box belong to a stable window without local displacement of the individual. If IoU value > 0.3 for five consecutive boxes, local displacement was detected. Bounding boxes were ignored until IoU $\leq$ 0.3 for five successive comparisons, after which a new video sequence centered on the individual was created from the first stable box.

4.2 Training of models

Images were separated into training (90%) and validation (10%) datasets based on their source videos. The data was split into ten folds (cross-validation) to obtain 5 distributions with all different validation datasets, which were defined according to a given optical flow threshold value. The dataset was split by video and not by images to avoid having similar images from the same video in training and validation datasets. We conducted several training sessions, varying the data comparison strategy and the loss function (Table 1).

Table 1. Table of configurations of data comparison strategies and losses for trainings. -I: selected images. -P: selected image pairs.

Dataset	**Data comparison**	**Data Augmentation**	**Data Transformation**	**Loss function**
D1-I	Augmented views	Yes	T_1	Baseline
D'-P	Temporal views	No	T_1 or T_2	Baseline
D'-P	Temporal views	No	T_1 or T_2	Motion-based

According to data comparison strategy, each sample is defined either as a single image or as an image pair, from which two views are constructed and compared. Augmented views are two images generated via data augmentation from the same anchor image (baseline comparison strategy). Temporal views are

two different images extracted from the same video. We randomly selected one sample per video from the dataset with an optical flow threshold of 0.5 (D1): the training dataset size was 1,813 and the validation dataset size was 202. We also selected all samples per video from datasets with the optical flow threshold of 0.5 (D1'), 0.75 (D2') and 1.0 (D3'): training datasets contained an average of 19,396, 9,901 and 5,836 images or image pairs and validation datasets contained an average of 2,113, 1,110 and 625 images or image pairs, respectively.

Optical flow was computed between all pairs of images from the same video and normalized by the mean across all pairs from all videos:

$$\text{OpticalFlow}_{\text{norm}} = \frac{\text{OpticalFlow} - \text{OpticalFlow}_{\text{min}}}{\text{OpticalFlow}_{\text{max}} - \text{OpticalFlow}_{\text{min}}} \tag{5}$$

with optical flow quantity values $\in [0, 1]$. Same datasets of image pairs were used for the temporal comparison strategy with the two loss functions. Optical flow values were extracted for each pair. The output of the projection head returns the encoded images, which are normalized to calculate the loss.

Training batches were built only from image pairs from the same video, across all videos in the dataset; comparisons with images from other videos and with themselves are ignored. Parameters of the Motion-based loss were fixed at $\alpha = 0.5$ and $\delta = 1e - 8$. Image transformations T_1 are inspired by those used in SimCLR and BYOL models (Figure 3). Image transformations T_2 are the same as T_1, except that geometric transformations (rotation) were applied in the same way to both images, while color transformations remained independent. Same datasets were used to train SimCLR and BYOL models.

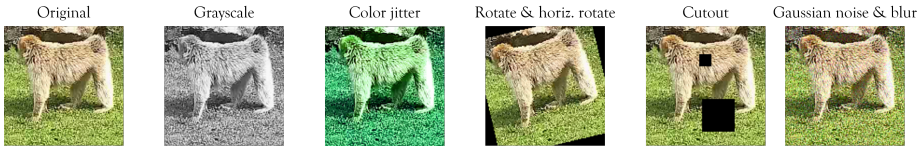

Fig. 3. Data augmentation techniques used

SimCLR model is trained with the LARS optimizer on very large batch sizes (i.e. 4096; more details in [3]). Since datasets used here are smaller, we fine-tuned models by varying hyperparameters such as the learning rate (10^{-4}, 10^{-6}), the batch size (64, 128), the optimizer (Adam, SGD) with a cosine decay learning rate schedule after 10 epochs. Hyperparameters were adjusted based on the validation dataset. The image size was fixed at 128, the patience at 50, the projection size at 128, the temperature at 0.07, the weight decay at 10^{-4} and the momentum at 0.9, with the encoder ResNet50.

BYOL was trained in the same way as SimCLR, but with the projection size fixed at 256, the projection hidden size at 4096 and the moving average decay at 0.99.

The training was stopped when the loss stabilized (Early Stopping), by taking a high number of epochs of 1,000 to reach the model convergence. The

model showing the lowest validation loss was considered the best model. Loss, top-1 accuracy, and top-5 accuracy were calculated on the training and validation datasets. The models were evaluated on the same test dataset described in Section 4.3.

4.3 Evaluation of models

To evaluate models, we used a SVM classification model on an annotated pose dataset. This dataset was created following the same protocol as the dataset used for training. Three pose classes were recognized and annotated on images across a total of 422 images: 230 images of *Sitting*, 96 images of *Lying down*, 96 images of *Standing quadrupedally*. A classification into seven classes was also conducted.

Image embeddings were separated into training (80%) and test (20%) datasets, with stratification based on the pose class. The data was split into ten folds with the training in eight folds and the test in two folds (cross-validation). We obtained 5 distributions with all different test datasets. We trained the SVM model by varying hyperparameters (C, kernel, gamma) using a hyperparameter grid search. Accuracy, precision, recall, and F1-score metrics were calculated for each training.

We evaluated whether image embeddings preserved temporal order. We selected videos with a maximum length of 20 images containing at least 10 images from the dataset filtered by an optical flow threshold of 0.5, resulting in a total of 234 videos. For each video, we selected ten images randomly. Like in [18], the image ordering was estimated using a greedy algorithm, which sequentially constructs the sequence by adding at each step the image most similar to the previous one according to the cosine similarity. The predicted order was compared to the ground-truth order using Kendall's Tau [18]. We computed similarity scores for each image embedding and retrieved the two closest neighbors. Order coherence was measured as the percentage of cases in which these neighbors corresponded to the preceding and following image embeddings in the true order.

5 Results

We trained SimCLR and BYOL models for 2 data comparison strategies and 2 loss functions. Table 2 reports only the best result obtained for each configuration (no significant performance improvement was observed when training on other datasets or using different parameter settings). The best results were obtained by BYOL model, which achieved an average precision score of 0.98 when trained with augmented views and temporal views from the dataset *D3'*, reaching its best performance for a learning rate of 0.0001, Adam optimizer and a batch size of 64 and 128. However, incorporating temporal information into the motion-based loss function does not lead to performance improvements and, in some cases, even results in degraded performance, which is contrary to the initial expectations.

Figure 4 illustrates the structure of the embedding space on one video of 55 image embeddings. The visualization aligns with the expected temporal ordering of image embeddings, with temporally adjacent images appearing closer in the embedding space. This observation supports the coherence of the learned representations of images with respect to temporal coherence.

Table 2. Table of mean scores for all configurations presented in Table 1 for 3 classes. Precision, Recall and F1-score results are expressed as percentages. S: SimCLR model. B: BYOL model. T': Temporal views with Transformation T_1. T: Temporal views with Transformation T_2. M: Motion-based loss function. Order: Order coherence

Method	Precision	Recall	F1-score	K-τ	Order
RGB	79 ± 2	74 ± 5	75 ± 4	**98**	**90**
S(D1)	88 ± 3	79 ± 6	81 ± 5	95 ± 0.5	89 ± 0.5
S(D1')	86 ± 4	78 ± 6	80 ± 6	96 ± 0.3	89 ± 0.2
S(D1)+T	76 ± 5	70 ± 6	71 ± 6	94 ± 3.1	87 ± 2.4
S(D1')+T	82 ± 4	74 ± 6	76 ± 5	$\mathbf{98 \pm 0.6}$	89 ± 0.3
S(D1)+T+M	77 ± 4	70 ± 5	72 ± 5	96 ± 0.6	88 ± 0.4
S(D3')+T+M	77 ± 5	72 ± 5	73 ± 4	94 ± 0.6	86 ± 0.8
S(D1)+T'	85 ± 6	78 ± 6	81 ± 6	96 ± 1.2	89 ± 0.6
S(D1')+T'	85 ± 3	81 ± 4	83 ± 4	$\mathbf{98 \pm 0.6}$	$\mathbf{90 \pm 0.6}$
S(D1)+T'+M	77 ± 4	73 ± 6	75 ± 5	97 ± 0.4	89 ± 1.2
S(D3')+T'+M	81 ± 4	74 ± 5	76 ± 4	97 ± 0.4	$\mathbf{90 \pm 0.2}$
B(D1)	97 ± 2	$\mathbf{97 \pm 3}$	$\mathbf{97 \pm 2}$	89 ± 1.2	80 ± 0.6
B(D3')	$\mathbf{98 \pm 1}$	$\mathbf{97 \pm 2}$	$\mathbf{97 \pm 1}$	93 ± 2.3	84 ± 2.0
B(D1)+T	95 ± 2	92 ± 3	93 ± 2	93 ± 1.5	84 ± 1.1
B(D3')+T	$\mathbf{98 \pm 2}$	$\mathbf{97 \pm 2}$	$\mathbf{97 \pm 2}$	93 ± 1.1	83 ± 1.3
B(D1)+T+M	94 ± 2	90 ± 2	92 ± 2	73 ± 0.5	65 ± 0.3
B(D1')+T+M	81 ± 5	76 ± 5	78 ± 5	70 ± 1.2	66 ± 1.1
B(D1)+T'	96 ± 2	95 ± 2	95 ± 2	91 ± 2.2	83 ± 2.0
B(D1')+T'	95 ± 3	93 ± 4	94 ± 3	92 ± 0.4	83 ± 0.3
B(D1)+T'+M	95 ± 2	91 ± 3	93 ± 2	73 ± 1.2	65 ± 0.3
B(D3')+T'+M	81 ± 4	76 ± 5	78 ± 4	71 ± 1.2	66 ± 0.4

6 Discussion

Our objective was to represent global body poses in a continuous manner, based on an intrinsic image property capturing motion quantity and temporal dynamics. To this end, we leveraged image embeddings from contrastive learning with a novel loss term and a dedicated data comparison strategy. Results show improved classification performance with SimCLR and BYOL models, with BYOL

model embeddings outperforming RGB images by 20% in precision, while no significant improvement is observed in temporal ordering. This suggests that pose representations can be effectively learned in a compact embedding space using contrastive frameworks. BYOL model performs poorly on temporal ordering, with temporal information and the motion-based loss function further degrading performance. While SimCLR model shows lower classification performance, it achieves better temporal ordering performance, highlighting its relevance for temporal ordering analysis in video sequences.

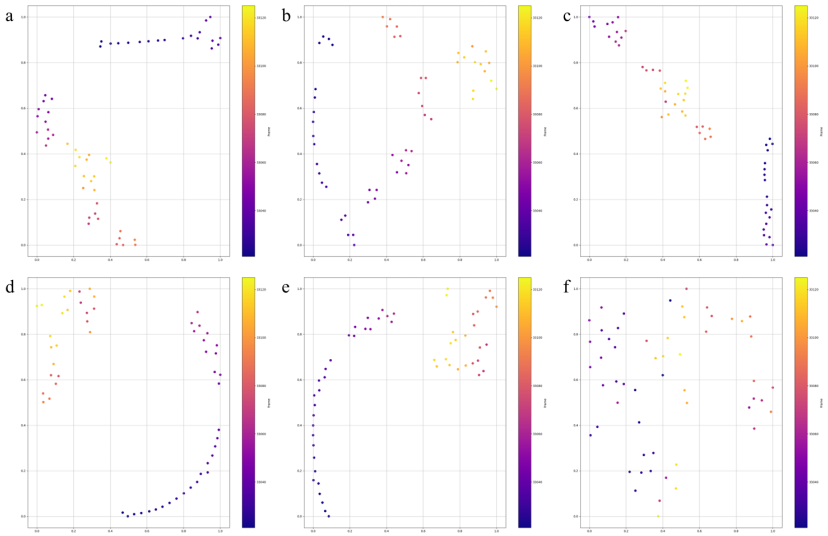

Fig. 4. t-SNE visualizations of embeddings from one video. Panels (a–c) correspond to BYOL model and panels (d–f) to SimCLR model. Columns represent configurations used: Augmented views (a, d), Temporal views T' (b, e) and Temporal views T' + Motion-based loss function (c, f)

The best performance for SimCLR model on temporal views is obtained with data transformations T_1, similar to baseline strategy used in SimCLR and BYOL models. In contrast, the data comparison strategy T_2, based on temporal coherence and used in video representation models [23], does not improve image classification or temporal ordering. This may be due to the limited variability introduced by color transformations alone, which may be insufficient to generate the diversity required for effective pose representation learning. Although temporal variability exists between consecutive frames, it appears insufficient on its own to fully support robust representation learning [3]. Relying on a single transformation strategy may limit representations diversity, while multiple complementary augmentations remain essential for learning discriminative features.

No performance improvement is observed with temporal views, despite the use of temporal image pairs expected to capture temporal ordering and relations between images. We tested a motion-based loss function to reflect action

dynamics and net changes across successive images, aiming for a continuous similarity representation and to mitigate bias from repetitive or cyclic movements. However, this loss function did not yield noticeable gains nor effectively capture temporal relationships, and substantially degrades temporal ordering performance for BYOL model relative to RGB images. For both SimCLR and BYOL models, representations show reduced temporal ordering quality, reflecting a loss of ordering information. While motion does not improve performance, incorporating temporal information leads to more favorable results. In particular, temporal views alone improve ordering for BYOL model, highlighting their importance. The degradation in performance introduced by the additional loss term likely due to conflicting objectives: contrastive methods enforce a global structure, while the proposed loss imposes a local constraint by enforcing distances proportional to a motion-based metric, bringing similar postures closer with pairwise distances reflecting motion differences. These objectives may interfere (e.g., strong data augmentations weakening temporal consistency [23], limited variability between temporally adjacent frames, or small differences in motion values), ultimately impairing the formation of a well-structured latent space and degrading downstream performance. As a direction for future work, we recommend exploring alternative variables for motion-based loss function to better capture temporal coherence, as well as sampling frames further apart in time from longer video sequences and leveraging more discriminative optical flow. Although a linear temporal structure is apparent in the visualizations, the order metric fails to reflect this behavior, resulting in relatively weak quantitative performance. Although Kendall's Tau [18] is used to evaluate the relative ordering of elements within a sequence, it may not fully capture the temporal structure observed in the embedding space. Exploring alternative metrics that better reflect temporal ordering may lead to a better evaluation of temporal structure in the learned representations.

We trained our methods on new unlabeled datasets and evaluated them on a dataset with subtle inter-class variations and highly similar individuals. Best performance is achieved with dataset D3', despite its smaller size compared to D1'. Performance with dataset D1 remains comparable, suggesting redundancy in larger datasets with highly similar images, which may reduce diversity and lead to overfitting. Smaller but more diverse datasets appear sufficient. Representations are evaluated using an SVM on a limited three-class dataset, providing a representation quality assessment but remaining insufficient to fully capture their discriminative power; finer-grained posture categories or larger, more diverse datasets would better evaluate robustness and generalization.

7 Conclusion

In this article, we investigated unsupervised approaches for learning pose representations from images and introduced a new primate pose dataset from zoo videos. Results indicate that contrastive learning models are well-suited for capturing subtle pose differences, with BYOL model combined with the baseline

data comparison strategy and loss function achieving particularly strong performance. However, no comparable improvements were observed when incorporating motion-based information. These findings highlight the potential of self-supervised contrastive methods for pose representation learning and action recognition from video data.

Acknowledgements This work is part of the APR IR 2021 MICMAC project supported by the region Centre-Val de Loire (France) and coordinated by Yann Locatelli at the MNHN. Authors are thankful to the Regional Council of Centre-Val de Loire for its support and to the partners of this project: MNHN, Université d'Orléans, INRAE (EA, PRISME and PRC laboratories), companies Tekin and ACTI'COM, ZooParc de Beauval and association Beauval Nature. Authors are also grateful to ZooParc de Beauval for collecting and transmitting the data.

References

1. Abassi, E., Bognár, A., De Gelder, B., Giese, M., Isik, L., Lappe, A., Mukovskiy, A., Solanas, M.P., Taubert, J., Vogels, R.: Neural encoding of bodies for primate social perception. Journal of Neuroscience **44**(40) (2024). https://doi.org/10.1523/JNEUROSCI.1221-24.2024
2. Caron, M., Goyal, P., Misra, I., Mairal, J., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607 (2020)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15745–15753 (2021). https://doi.org/10.1109/CVPR46437.2021.01549
5. Dai, W., Du, Y., Bai, H., Cheng, K.T., Li, X.: Semi-supervised contrastive learning for deep regression with ordinal rankings from spectral seriation. Advances in Neural Information Processing Systems **36**, 57087–57098 (2023)
6. Dai, Y., Shen, J., Zhai, Z., Liu, D., Chen, J., Sun, Y., Li, P., Zhang, J., Zhang, K.: High-order contrastive learning with fine-grained comparative levels for sparse ordinal tensor completion. In: Forty-first International Conference on Machine Learning (2024)
7. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: Tclr: Temporal contrastive learning for video representation. Computer Vision and Image Understanding **219**, 103406 (2022). https://doi.org/10.1016/j.cviu.2022.103406
8. Diba, A., Sharma, V., Safdari, R., Lotfi, D., Sarfraz, M.S., Stiefelhagen, R., Van Gool, L.: Vi2clr: Video and image for visual contrastive learning of representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1502–1512 (2021). https://doi.org/10.1109/ICCV48922.2021.00153
9. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9588–9597 (2021)
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning pp. 9729–9738 (2020)
12. Hu, H., Wang, X., Zhang, Y., Chen, Q., Guan, Q.: A comprehensive survey on contrastive learning. Neurocomputing **610**, 128645 (2024). https://doi.org/10.1016/j.neucom.2024.128645
13. Jenni, S., Favaro, P.: Self-supervised multi-view synchronization learning for 3d pose estimation. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020). https://doi.org/10.1007/978-3-030-69541-5_11
14. Knights, J., Harwood, B., Ward, D., Vanderkop, A., Mackenzie-Ross, O., Moghadam, P.: Temporally coherent embeddings for self-supervised video representation learning. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8914–8921 (2021). https://doi.org/10.1109/ICPR48806.2021.9412071
15. Koishekenov, Y., Vadgama, S., Valperga, R., Bekkers, E.J.: Geometric contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 206–215 (2023). https://doi.org/10.1109/ICCVW60793.2023.00028
16. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11205–11214 (2021). https://doi.org/10.1109/CVPR46437.2021.01105
17. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6964–6974 (2021). https://doi.org/10.1109/CVPR46437.2021.00689
18. Ramanathan, V., Tang, K., Mori, G., Fei-Fei, L.: Learning temporal embeddings for complex video analysis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4471–4479 (2015). https://doi.org/10.1109/ICCV.2015.508
19. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (June 2015). https://doi.org/10.1109/CVPR.2015.7298682
20. Si, L., Wang, J., Qiang, W.: A generalized learning framework for self-supervised contrastive learning. arXiv preprint arXiv:2508.13596 (2025). https://doi.org/10.48550/arXiv.2508.13596
21. Wang, J., Lin, Y., Ma, A.J., Yuen, P.C.: Self-supervised temporal discriminative learning for video representation learning. arXiv preprint arXiv:2008.02129 (2020). https://doi.org/10.48550/arXiv.2008.02129
22. Zhao, L., Lin, Z., Sun, R., Wang, A.: A review of state-of-the-art methodologies and applications in action recognition. Electronics **13**(23), 4733 (2024). https://doi.org/10.3390/electronics13234733
23. Ziani, A., Fan, Z., Kocabas, M., Christen, S., Hilliges, O.: Tempclr: Reconstructing hands via time-coherent contrastive learning. In: 2022 International Conference on 3D Vision (3DV). pp. 627–636 (2022). https://doi.org/10.1109/3DV57658.2022.00073