Unsupervised Canine Emotion Recognition using Momentum Contrast

Aarya Bhave , Alina Hafner , Anushka Bhave , Peter A. Gloor *

Posted Date: 12 October 2024

doi: 10.20944/preprints202410.0927.v1

Keywords: contrastive learning; momentum contrast; Panksepp seven emotions; canine emotions; unsupervised learning

Article

Unsupervised Canine Emotion Recognition using Momentum Contrast

Aarya Bhave [1], Alina Hafner [2], Anushka Bhave [1] and Peter A. Gloor [1,*]

[1] MIT SDM
[2] TUM
* Correspondence: pgloor@mit.edu; MIT SDM, 77 Massachusetts Avenue, Cambridge MA 02142, USA

Abstract: We describe a system for identifying dog emotions based on the dogs' facial expressions and body posture. Towards that goal, we built a dataset with 2184 images of ten popular dog breeds, grouped into seven similarly sized primal mammalian emotion categories defined by neuroscientist and psychobiologist Jaak Panksepp that are 'Exploring', 'Sadness', 'Playing', 'Rage', 'Fear', 'Affectionate' and 'Lust'. We modify the Contrastive Learning framework MoCo (Momentum Contrast for Unsupervised Visual Representation Learning) to train it on our original dataset and achieve an accuracy of 43.2% on a baseline of 14%. We also trained this model on a second publicly available dataset that resulted in an accuracy of 48.46% but had a baseline of 25%. We compared our unsupervised approach with a supervised model based on a ResNet50 architecture. This model when tested on our dataset having seven Panksepp labels resulted in an accuracy of 74.32%

Keywords: contrastive learning; momentum contrast; Panksepp seven emotions; canine emotions; unsupervised learning

1. Introduction

Dogs and humans have been cohabiting as species for over 20,000 years[1,2]. This makes dogs humans' longest and most loyal evolutionary companion, well before horses and bovines were domesticated. However even in the modern era, despite thousands of years of coexistence, it is still hard for humans to accurately anticipate dog emotions. While dog therapists and behavioral researchers demonstrate high accuracy in correctly reading the mood state of a dog, novice owners struggle to correctly understand the behavior of their dog. By creating AI and Deep Learning applications that interpret facial expressions and body posture from images and videos of dogs, expert knowledge can be made generally accessible to permit everybody to interact with their canine companion and correctly read the mood state of the dog.

Jaak Panksepp in his research about affective consciousness [3], specifically on core emotional feelings in animals and humans, puts forth an argument suggesting that the essence of emotional feelings lies within the evolved emotional action systems in all mammalian brains. It presents a view that makes the tough concepts related to the workings of the brain-mind more accessible, and it even proposes that affective feelings might express the neuro-dynamics of brain systems that produce instinctual emotional behaviors [3]. Then, for primary process affective consciousness, it is regarded as instinctive and universal for all mammals, so it is easier to investigate in animals. While some of the secondary processes-for example, consciousness of feelings in behavioral decisions-may be assessed using exceptionally prudent learning procedures, special emphasis must reside in identifying intrinsic emotional action tendencies in the organism.

He argues that core emotional feelings are represented in the neuro-dynamic landscapes of various emotional action systems like SEEKING, FEAR, RAGE, LUST, CARE, PANIC, and PLAY. By studying these systems, researchers can uncover the neural basis of emotional consciousness in both humans and animals. This approach encourages the ethological analysis of emotional action tendencies and accompanying brain changes for effective monitoring of emotional states.

 doi:10.20944/preprints202410.0927.v1

Additionally, it suggests that meaningful progress in this field requires more open discussions among animal brain researchers.

The study of animal emotions, known as affective biology, is becoming increasingly popular across various research fields like evolutionary zoology, affective neuroscience and comparative psychology [4–7]. From 2014 to 2022, scientists used advanced technology to understand and improve the emotional well-being of animals by tracking their movements and recognizing gestures. For instance, Broome et al [8] survey broad research using computer vision to assess animal emotions and recognize signs of pain by closely analyzing facial expressions and body language. Identifying animal emotions is hard because they might hide their internal feelings. Traditionally, researchers would watch or record videos of animals to study their behavior, but now, automatic analysis of facial expressions and body poses allows for a more detailed understanding of their emotional states. Studies on animal emotions have included estimating poses, and using deep learning to identify and track animals. Understanding animal emotions by analyzing their facial expressions and body language is more complex than simply tracking their movements.

Lately, scientists have been using computer vision and deep learning to recognize emotions in dogs. Hussein et. al [9] for example, used different sensors to capture dogs' movements and combined different streams of data to check for wellbeing among dogs. Similarly, Franzoni et al., [10] conducted experiments to evoke emotions in dogs and focused on detecting those emotions through their facial expressions. Ferres and colleagues took a different approach, recognizing emotions from the body poses of dogs by pinpointing key areas on their bodies and faces [5]. Most of these studies cannot attain much accuracy without proper visibility of face and limbs [4]. Research in dog emotion recognition using computer vision and deep learning has mainly centered around clear, high-resolution images of individual dogs. Surveillance cameras have been commonly used, where the emotional states of the animals have mainly been inferred from their physical behavior. In contrast, past research on human emotion recognition has used text, audio, or video data along with various models to achieve high accuracy, often relying on facial expressions or body language analysis as input for supervised or unsupervised learning. However, we are currently not aware of any studies using unsupervised learning for dog emotion recognition.

Semi-supervised learning has emerged as a promising approach for leveraging large amounts of unlabeled data to enhance the performance of learning-based networks. By incorporating both labeled and unlabeled data, semi-supervised learning can improve the generalization capability of models. However, if the semi-supervised learning policy is inefficient, then quality and reliability associated with the learned features would not hold. It basically implies that designing an efficient semi-supervised learning strategy and hence, ensuring the reliability of learned features for every specific application is still an open challenge.

Most facial emotion recognition methods heavily rely on supervision, making it challenging to analyze emotions without considering individuals. Conversely, self-supervised learning offers a way to learn representations without supervision. Kim et al. introduce a new adversarial learning approach [11]. Specifically, they help the face emotion recognition network to better understand complex human emotional elements by learning weak emotion samples from strong ones in an adversarial manner. Their method is able to recognize human emotions independently of individuals, leading to more accurate facial expression understanding by proposing a contrastive loss function to improve the efficiency of adversarial learning. In this paper, we develop a system to recognize emotions in dogs based on their facial expressions and body posture using the Contrastive Learning framework MoCo [12] (Momentum Contrast for Unsupervised Visual Representation Learning) to identify the seven Panskepp emotions.

The chief contributions of this paper are:

1. We construct a novel, high quality, diverse and un-skewed dataset of 2184 images consisting of the ten most popular dog breeds worldwide with varying shapes and sizes. These dog breeds are ‘Siberian Husky’, ‘Rottweiler’, ‘Golden Retriever’, ‘Labrador Retriever’, ‘Pug’, ‘Poodle’, ‘Beagle’, ‘German Shepherd’, ‘Pembroke Welsh Corgi’ and ‘French Bulldog’. We construct equally sized groups of these images with the seven Panskepp emotion labels “Exploring”, “Sadness”, “Playing”, “Rage”, “Fear”, “Affectionate”, and “Lust”.

 doi:10.20944/preprints202410.0927.v1

2. We leverage the Contrastive Learning frameworks SimCLR (A Simple Framework for Contrastive Learning of Visual Representations) and MoCo (Momentum Contrast for Unsupervised Visual Representation Learning) on our dataset to predict the seven Panksepp emotions using unsupervised learning. We significantly modify the MoCo framework to obtain the best possible results on our hardware. We also test the unsupervised learning models on a publicly available dog emotion dataset to compare relative performance on baseline accuracies.
3. We build a supervised learning model based on the ResNet50 architecture and run it on our dataset as well as the publicly available dataset to obtain benchmark results.

Our research makes significant contributions to the field of Animal-Computer Interaction (ACI) by introducing a novel method for recognizing dog emotions, fostering improved human-animal communication, adhering to the highest ethical standards, and utilizing a unique contrastive learning approach. By better recognizing dog emotions, we help bridge the human-animal communication barrier. Additionally, we advance the field of applied machine learning by demonstrating the effectiveness of semi-supervised and self-supervised learning techniques in animal emotion recognition.

2. Method

In this experimental study and subsequent reconfiguration of framework architectures, we have pursued an approach which prioritizes novelty. Consequently, we have prepared an original dataset to carry out this academic research. Following this, we have experimented with contemporary contrastive learning frameworks. A detailed taxonomy for contrastive learning methods was explored by Le-Khac et al. [13], differentiating supervised, semi-supervised and unsupervised contrastive learning techniques. After experimentation with several contrastive learning frameworks, we have implemented a modified version of the MoCo framework [12] developed by He et al.

2.1. Creating the Dataset

To implement unsupervised visual representation learning [14] on image data, we embarked on the task of collecting, analyzing and evaluating images of dog emotional behaviors [15]. Our key focus was capturing the respective physiological displays, for different emotional temperaments among several different dog breeds. The dataset we constructed includes the images of the top ten most popular dog breeds worldwide [16], categorically distributed in accordance to their breed and seven different emotional behaviors. These emotions include ‘Exploring’, ‘Sadness’, ‘Playing and Happy’, ‘Rage’, ‘Fear’, ‘Caring and Affectionate’, and ‘Lust’. The ten dog breeds included in this dataset are ‘Siberian Husky’, ‘Rottweiler’, ‘Golden Retriever’, ‘Labrador Retriever’, ‘Pug’, ‘Poodle’, ‘Beagle’, ‘German Shepherd’, ‘Pembroke Welsh Corgi’ and ‘French Bulldog’. In addition to this we have also added a Miscellaneous section in which images of other dog breeds for reasons discussed further in this section. Interventionary studies involving animals or humans, and other studies that require ethical approval, must list the authority that provided approval and the corresponding ethical approval code.

2.1.1. Methodology for Labelling Images

We have adhered to the comparative cognition research principles, posited by Kujala [17] for analyzing and evaluating our images. The visual identification of dog emotions relies heavily on the detailed observation of facial expressions, body postures, and other non-verbal cues. Facial expressions, such as ear position, eye shape, and mouth movements, are critical indicators of a dog's emotional state. For instance, a dog with relaxed ears, soft eyes, and a slightly open mouth is typically expressing contentment or relaxation, whereas pinned-back ears, a furrowed brow, and a closed mouth might indicate fear or anxiety.

Tail position and movement also provide significant emotional cues; a wagging tail generally suggests excitement or happiness, but the speed and height of the wag can modify the interpretation, such as a slow wag with a low tail often signaling uncertainty or submissiveness. In addition to facial

 doi:10.20944/preprints202410.0927.v1

expressions and tail movements, body posture plays a crucial role in the visual identification of emotions in dogs.

A confident and happy dog will usually have a loose, relaxed posture, often with a playful stance or a wagging tail. In contrast, a dog experiencing fear or aggression might exhibit a tense, rigid body, with weight shifted backward or forward, depending on whether the dog is preparing to flee or confront. Submissive dogs may crouch low to the ground, tuck their tails between their legs, and avoid direct eye contact. By carefully observing these visual cues in various contexts and interactions, owners and researchers can gain valuable insights into the emotional states of dogs, allowing for more effective communication and stronger human-animal bonds. Using the above key factors as described in [17], we were able to successfully classify images into seven emotion labels as discussed in the section that follows.

2.1.2. Dataset Description

We have primarily used Google Images and Wikipedia Images to collect a total of 2184 images of different and distinct individual dogs across the internet, that we collected and manually labelled using the methodology highlighted in the previous section of this paper. It is important to note that 2184 images were collected from the internet independently of each other and this is before applying any augmentation to these images for training. It is also important to signify that the image augmentation techniques discussed in the latter sections of the paper have not increased or decreased the size of the original dataset of 2184 images in any way shape or form. For the sake of preserving copyright, the dataset that we have collected cannot be made public for open-source usage, since it contains several images that have copyrights associated with them. Thus we cannot publicly release them for the scientific community. The search engines were used from January 18th 2024 to April 4th, across a span of three months. The emotional behaviors we have considered are adapted from Panksepp's Seven Core Emotional Systems [3].

Jaak Panksepp was one of the most influential neuroscientists in recent times. Panksepp argued that each of these seven emotions has a dedicated system in the subcortical regions in the brain of all mammals. Dogs were domesticated from now-extinct wolves between 11,000 and 16,000 years ago [2]. This has given evolutionary reasons for dogs and humans alike, to share and adapt each other's emotional behaviors due to our closely bound symbiotic relationship. The images for this dataset show the seven Panksepp emotions. Emotions 'Sadness', 'Exploring', and 'Lust' were found to be universally exhibited by all dog breeds in equal frequencies, while the emotions 'Caring and Affectionate', Playing and Happy', 'Rage', and 'Fear' were found to be exhibited in varying frequencies among different dog breeds. Working dog breeds like German Shepherds and Rottweilers tend not to display fearful behaviors as often as breeds like Pugs and Pembroke Welsh Corgis. Similarly, breeds like the Golden Retriever and Poodle rarely exhibit aggressive behavioral tendencies [15].

These factors influence the image count for each emotion and breed. To tackle this issue, we introduced a 'Miscellaneous' category that contains images of other dog breeds that are not considered in the ten breeds listed above. The dataset that we assembled consists of 2184 images, with 300 images per emotion. Approximately 200 images are included for every considered dog breed, also including the miscellaneous category. We also made certain that each image contained the entire body of the subject, along with the environment in which the subject is present. We ascertained that behavioral displays are largely dependent on the overall stance and body posture of the dog [12]. The temperament of the dog varies with respect to its surrounding terrain, amount of natural lighting and weather conditions [15].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2024 doi:10.20944/preprints202410.0927.v1

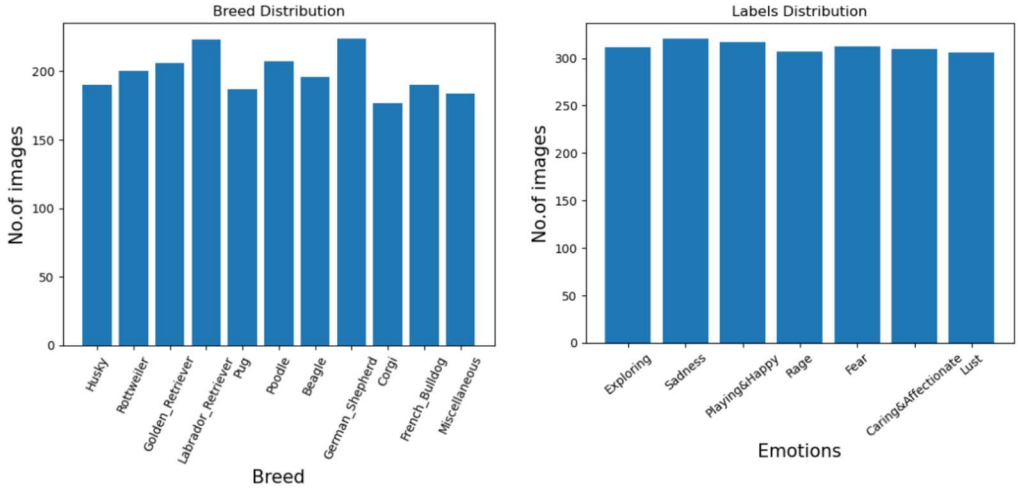

Figure 1. Breed and Emotion distribution of images in dataset.

Figure 2. Examples of seven emotional behaviors in the Golden Retriever breed.

2.2. Supervised Learning Benchmarks

We trained a supervised CNN which builds on the ResNet50 architecture. Overall ResNet50 performs quite well on image classification tasks, although in direct comparison to other image classification models, such as Alex Net and GoogleNet, it gets confused by certain objects such a dogs and deers, classifying them as horses [18]. The model utilizes the PyTorch Lightning framework to streamline the training and validation processes. Image data is preprocessed and augmented through a series of transformations to enhance model generalizability and to mimic a variety of real-world conditions. These transformations include random rotations of ±10 degrees, horizontal flips with a 50% probability, resizing images to 224x224 pixels, center cropping, and normalization using the ImageNet dataset's mean and standard deviation values ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]).

The dataset comprises 2184 images categorized into 7 classes representing the Panksepp's emotions [3] (i.e., 'Exploring', 'Sadness', 'Playing', 'Rage', 'Fear', 'Affectionate' and 'Lust'), with the directory structure facilitating the use of PyTorch's ImageFolder for automatic labeling based on folder names. A custom DataModule class, an abstraction provided by PyTorch Lightning, manages data loading and splitting. The dataset is divided into an 80-20 ratio for training and testing. DataLoaders, a PyTorch sampling library, for each subset shuffles the training data and batch it into sets of size 32, optimizing the loading process and preparing the dataset for efficient training. The core of our model is a CNN based on the ResNet-50 architecture, pre-trained on the ImageNet dataset.

 doi:10.20944/preprints202410.0927.v1

Transfer learning is employed, leveraging the pre-trained convolutive bases to extract features, while the fully connected output layer is adapted to our specific task of classifying emotional states. This output layer is redefined to match the number of classes in our dataset, replacing the original ImageNet classifier. The CNN utilizes the Adam optimizer with a learning rate of 0.001. Training involves the computation of cross-entropy loss, a common choice for multi-class classification problems. Performance metrics, specifically loss and accuracy, are logged and monitored during training to assess model convergence and effectiveness on both training and validation datasets. Using PyTorch Lightning's Trainer, the model undergoes 20 epochs of training, where it learns to minimize the loss function and improve accuracy on the processed images. The framework handles under-the-hood functionalities like GPU acceleration (if available) and model checkpointing, allowing us to focus on model architecture and performance.

2.3. Unsupervised Learning with Contrastive Learning Frameworks

In computer vision research, novel sophisticated methodologies such as contrastive learning and unsupervised visual representation learning have initiated a paradigm shift in representation learning and pretraining strategies. Contrastive learning, while originally developed within the framework of unsupervised learning, has demonstrated remarkable potential for few-shot learning scenarios [19]. Contrasted learning precisely focuses the forcing of the model to push closer representations of connected images, rather pushing apart representatives of the least-related images in an embedding space. The process typically involves constructing pairs or batches of images and applying augmentation techniques to create different views of the same image. These views are then passed through a neural network to generate embeddings [20]. In the context of emotional behavior analysis, this means contrasting instances of similar emotional expressions (positives) while distinguishing them from dissimilar expressions (negatives).

2.3.1. Why Contrsative Learning?

Emotional behaviors often exhibit complex and subtle variations influenced by factors such as cultural background, individual differences, and contextual cues. Contrastive learning can help disentangle these variations by forcing the model to focus on the discriminative features that distinguish between different emotional states, thus enabling more nuanced and robust representations [21]. The emotions chosen for the dataset are such that they produce diametrically opposite physiological displays among dogs. Thus, a contrastive learning framework should be able to produce well-separated plots for all the labels within an embedding space. Additionally, this opens the possibility of finding novel emotional behaviors that may get represented as outgrowths of an existing label cluster. The cluster formation and corresponding clustering methods are motivated by Zhang et al. [22].

In addition, several studies have demonstrated the effectiveness of contrastive learning for learning rich representations from images. For example, Jaiswal, A. et al. [23] conducted a survey of contrastive learning methods and their applications, highlighting the ability of contrastive learning to learn representations that capture semantic information in images with good accuracy metrics. In conclusion, contrastive learning is a powerful technique that can be used for the analysis of emotions using image data. By learning to differentiate between images that represent different emotions, contrastive learning can capture the unique visual features associated with each emotion. This makes it a valuable tool for emotion discovery and analysis in scenarios where labeled data is limited, as it can learn meaningful representations from unlabeled data, as demonstrated by Shen et al. [24].

2.3.2. Experimenting with the SimCLR Framework

A Simple Framework for Contrastive Learning of Visual Representations, developed by Chen T. et al. [25] is a framework which offers effective representation learning on medium to large scale datasets. Its novelty lies in the fact that it is the most straightforward contrastive learning framework which still manages to outperform several other more complex frameworks. In our implementation

 doi:10.20944/preprints202410.0927.v1

of the SimCLR framework, we have utilized a pretrained ResNet-50 as an encoder. We have used a non-linear projection head that includes a fully connected layer with 2048 input features and 2048 output features, a ReLU activation layer, and another fully connected layer with 2048 input features and 128 output features.

In addition to the modification in the architecture of the multi-layer perceptron, we have also made some changes to the pretrained ResNet-50 encoder. We have replaced the parameters of the first convolutional layer, changing the kernel size to (3, 3), stride to (1, 1) and removed the padding. The minimum recommended batch size in the SimCLR [25] paper is 256, with the optimum batch size being 2048. We have used the Layer-wise Adaptive Rate Scaling (LARS) optimizer put forth by You Y. et al. [26]. We have used a learning rate of 0.2 and a weight decay of 10-6.

We have also implemented a linear warmup of 10 epochs, after which we follow a cosine annealing learning rate scheduler for 500 epochs. We have used an input image size of 32x32, in order to be able to train with an NVIDIA RTX 3070 Ti GPU with 8 Gigabytes of video memory. We have used the Normalized Temperature-scaled Cross Entropy Loss (NT Xent) from [27] to train the encoder. The temperature value we have used for the NT Xent loss is 0.5. Because of the resource-intensive nature of the SimCLR framework, we have not implemented a downstream classifier and instead directly generated a t-distributed Stochastic Neighbor Embedding with perplexity value of 50.

2.3.3. Frameworks That Require Less GPU Memory

The contrastive learning frameworks have an intrinsic need for a large batch size, due to the fact that a lot of negative samples are required to properly differentiate data points from each other within the embedding space. The SimCLR [25] framework has a recommended batch size of 2048, which is not implementable with memory sizes of most GPU's.

Momentum Contrast for Unsupervised Visual Representation Learning [12], also known as MoCo, is another popularly used contrastive learning framework that utilizes a technique called 'Cross Batch Memory Accessing' [28] to get around the issue of requiring a large batch size. The MoCo framework involves the implementation of a queue to store the features generated by the past batches; and use them as negative samples during training. The MoCo frameworks uses two encoders unlike SimCLR which uses a single encoder interchangeably. The two encoders are called 'Query Encoder' denoted by 'encQ' and the 'Momentum Encoder' denoted by 'encK'. The query encoder is updated using stochastic gradient descent, while the momentum encoder is updated using exponential moving average.

Another efficient way of resolving the issue with GPU memory is 'Model Parallelism'. Model parallelism involves splitting the model across multiple GPUs, with each GPU responsible for computing a subset of the model [29]. This allows training of larger models that would not fit into a single GPU's memory. However, model parallelism introduces communication overhead between GPUs, which can impact training performance.

In addition to this we also considered Minibatch-Gradient-Checkpointing (MBGC). MBGC is a framework that enables training large models with limited GPU memory by checkpointing activations and recomputing them during backward pass to reduce memory usage [30], [31]. It divides the model into segments and checkpoints intermediate activations, thereby trading off computation for memory. This approach allows training of larger models without increasing the memory footprint significantly.

2.4. Momentum Contrast for Unsupervised Visual Representation Learning

Among the frameworks that use Momentum Contrast are MoCo [12], MoCov2 [32] and MoCov3 [33]. The original MoCo framework is the computationally lightest framework to train. For this project, we have implemented a modified version of the original MoCo framework.

The modifications include alteration of the data augmentations, use of an alternate encoder, and slight changing of the encoder architecture. The most notable change we have made are the data augmentations.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2024 doi:10.20944/preprints202410.0927.v1

The MoCo [12] paper has suggested the use of different data augmentations. After testing performance with different augmentations, we found that the following augmentations produce the best results.

2.4.1. Data Augmentations

The following data augmentations were performed on training data:

1. Random Resizing and Cropping.
2. Random Horizontal Flip (probability = 0.5).
3. Random application of Color Jitter with jitter values of brightness = 0.4, contrast = 0.4, saturation = 0.4, hue = 0.1 and probability = 0.8.

As demonstrated in MoCo [12], we have built our train, test, and validation data loaders. The image data augmentations are applied to the same image twice by the train data loader, converted to a PyTorch tensor, normalized, and fed into the model for training as two positive samples.

These augmentations had the best effect in increasing the diversity of the dataset and encouraging the model to learn embeddings efficiently from our train data loader. The augmentations were initially borrowed from the MoCo Repository [12], but were modified according to their effect on model performance. The size of the image fed into the model was treated as a hyperparameter during training, due to its effect on GPU memory consumption and accuracy.

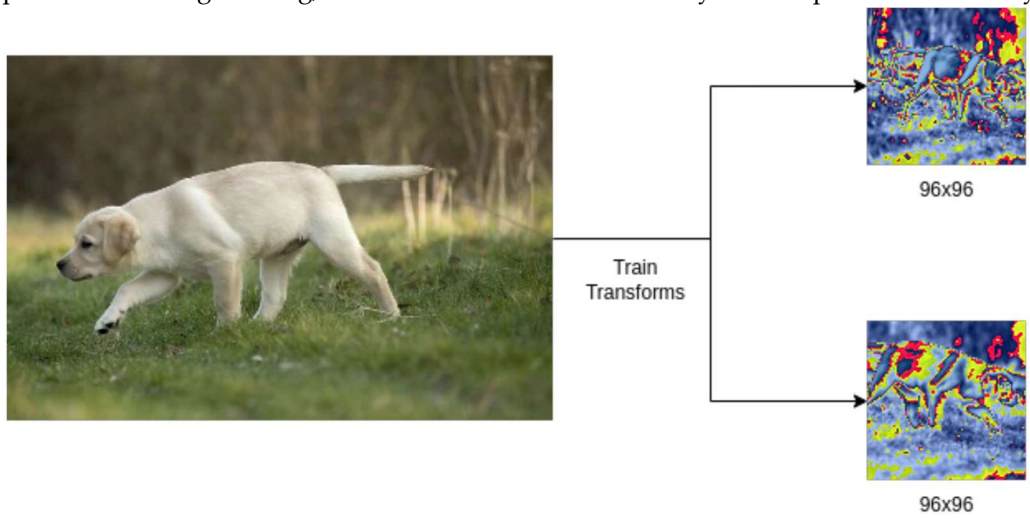

Figure 3. An image of a labrador puppy displaying the 'Exploring' behavior augmented according to above mentioned augmentations (image source Wikipedia).

2.4.2. Modification of ResNet-34 and ResNet-18 Architectures

The MoCo paper discusses experimentation with R50w4x, R50w2x and R50 as encoders. These encoders are resourceintensive and difficult to train with limited hardware. For this project, we had to experiment with ResNet-34 and ResNet18 as encoders. We have implemented the commonly used CIFAR-10 ResNet18/34 recipe, which in comparison to the ImageNet ResNet recipe:

1. Replaces the first convolution layer with kernel size = 3 and stride = 1.
2. Removes the first pooling layer.

In addition to these changes, each batch normalization layer is replaced with Split Batch Normalization Layers. Split Batch Normalization [34] is a technique that allows the simulation of multiple GPU behaviors on a single GPU. Each split batch normalization layer is set to carry out 8 splits across a batch. We have also used Data Parallelism [35], in order to train these split batches in a single GPU.

The query encoder was updated using Stochastic Gradient Descent, and the NT-Xent Loss function was found to be better performing for this project than the Info NCE loss function.

 doi:10.20944/preprints202410.0927.v1

$$\mathcal{L}_q = -\ln\left(\frac{e^{\left(\frac{q.k_+}{\tau}\right)}}{\sum_{i=0}^{k} e^{\left(\frac{q.k_i}{\tau}\right)}}\right)$$

The momentum encoder was updated using Exponential Moving Average as used in the MoCo [12] paper. The ResNet-18 as well as the ResNet-34 encoder were set to produce final feature vectors of length 128.

2.5. KNN Classifier

The feature vectors produced by the encoder were tested for accuracy using a KNN Classifier borrowed from [36]. The value for the 'K' nearest neighbors to be considered in the embedded space was set to 200. The classifier uses cosine similarity as a distance metric for the feature vectors. The implementations for this were taken from the GitHub Repositories http://github.com/zhirongw/lemniscate.pytorch and https://github.com/leftthomas/SimCLR. The learning rate scheduler is a cosine annealing scheduler with the value for temperature set to 0.1, since it performed best with this InstDisc KNN Monitor [36].

It is important to note that a fully connected downstream network that uses the pre-trained encoder features as input will produce an outstanding accuracy for practical tasks. The similarity scores are adjusted by a temperature parameter and converted to probabilities. Next, the classifier creates a one-hot encoding for the labels of the nearest neighbors and calculates a weighted sum of these one-hot vectors to obtain the predicted scores for each class.

The KNN monitor in InstDisc [36] serves as a mechanism to utilize the learned representations for identifying potential negative pairs, which enhances the quality of the learned representations for downstream tasks. The choice of K in the KNN monitor is a key hyperparameter that can be tuned to balance between the diversity and selectivity of the negative pairs.

3. Results

In this section, we explore the performance of our unsupervised learning models with different values of hyperparameters, and different training environments and compare them with two supervised learning models that use a simple CNN based on a ResNet-50 architecture. Among the hyperparameter values, we have experimented with the batch size, the value of K in the KNN Monitor Classifier [36], and the number of epochs for which the model was trained. We have used a modified SimCLR [25] framework and several different modified MoCo [12] frameworks for unsupervised representation learning in this project. All the relevant code and results will available in GitHub [link removed for anonymized review].

3.1. Results Produced by the SimCLR Framework

The SimCLR [25] framework was used in initial exploratory steps for contrastive learning for this project. The maximum resolution that could be used with this framework was 32x32 paired with a batch size of 256. The computational intensity of the SimCLR framework prevented us from training it on our dataset with 7 emotion labels. As a result, we attempted to train the model using a publicly available dataset, put together by Daniel Shan Balico, on Kaggle [37]. Despite being trained on the publicly available dataset with only four labels [37], the model failed to cross the accuracy of 30% where the base level guessing accuracy would be 25%, since only four labels are available in [37]. The implementation for that can be referred to in our repository. This made the model impractical to proceed with and build downstream networks.

3.2. Results Produced by the SimCLR Framework

All the MoCo models were trained on our own proprietary dataset with seven Panksepp emotion labels. The baseline accuracy for this dataset is 14.28%, and our unsupervised learning

 doi:10.20944/preprints202410.0927.v1

models have managed to cross the accuracy of 40%. Table 1 shows the hyperparameter values and the performance metrics for our best performing models.

Table 1. Best four results produced by MoCo frameworks.

Run	Image Resolution	Encoder	GPU	Batch Size	LR	K Value KNN	Momentum	Epochs	NT-Xent Temperature	Accuracy
1	96	ResNet-18	NVIDIA RTX-3070	256	0.3	150	0.99	1200	0.1	40.24%
2	96	ResNet-18	NVIDIA RTX-3070	128	0.3	200	0.99	1800	0.1	42.19%
3	96	ResNet-34	NVIDIA Tesla P100	128	0.25	150	0.95	1200	0.1	39.71%
4	96	ResNet-34	NVIDIA Tesla P100	128	0.3	200	0.99	1800	0.1	43.42%

3.2.1. Testing Accuracy and Training Loss for Our Unsupervised Learning Models

In the following section, we display our test accuracy and training loss graphs. In all of these training runs, we have seen a plateau appear after 700 epochs. We have observed more variance in the accuracy values, when training with the more resource-intensive ResNet-34 encoder.

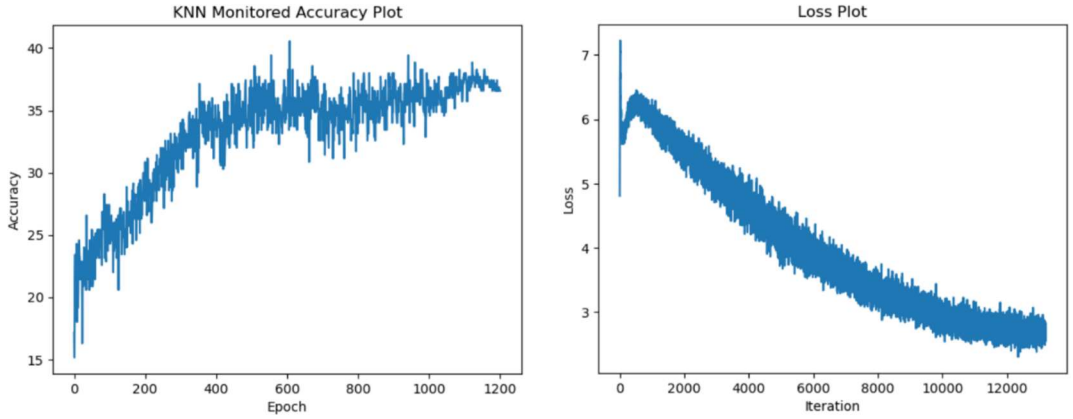

Figure 4. Test accuracy and training loss results for ResNet-18 encoder with 1200 epochs.

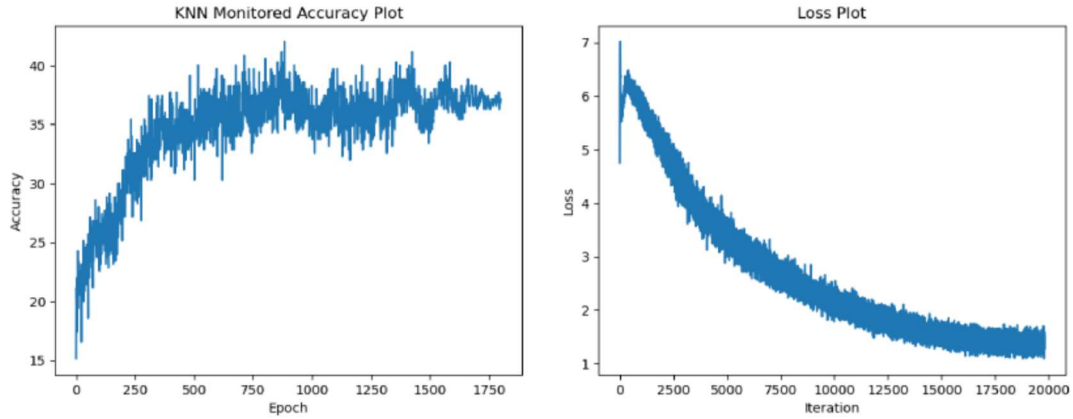

Figure 5. Test accuracy and training loss results for ResNet-18 encoder with 1800 epochs.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2024 doi:10.20944/preprints202410.0927.v1

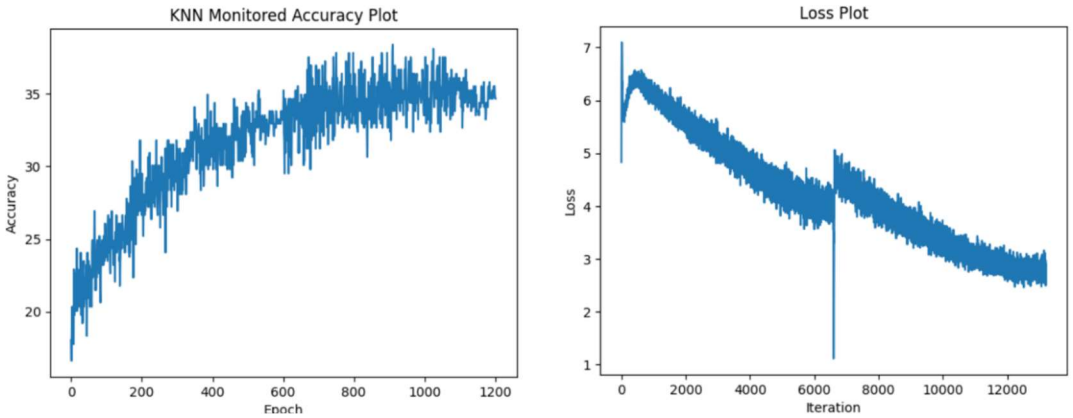

Figure 6. Test accuracy and training loss results for ResNet-34 encoder with 1200 epochs.

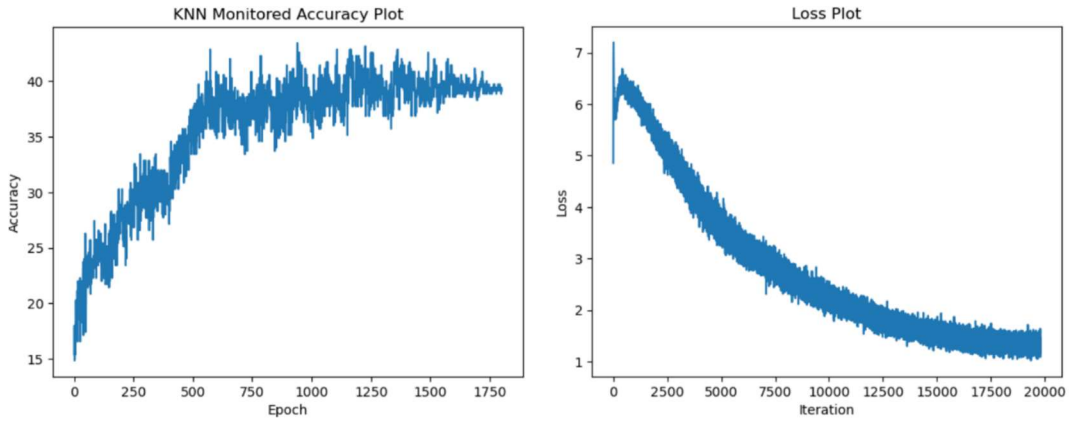

Figure 7. Test accuracy and training loss results for ResNet-34 encoder with 1800 epochs.

Figure 8 compares the performance shown by the different unsupervised models and the supervised control model on our dataset.

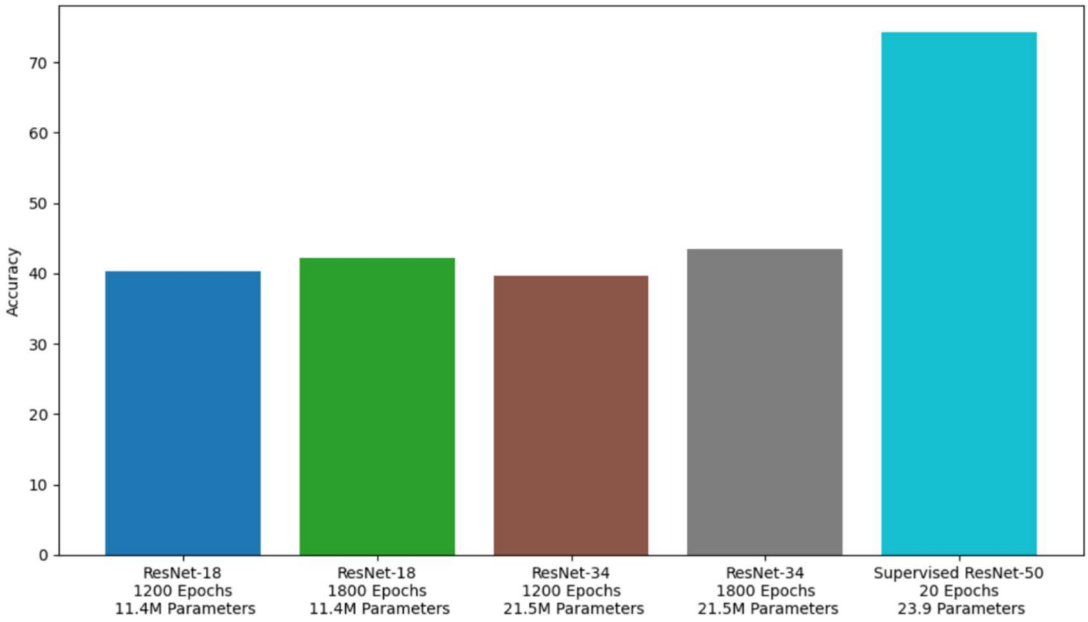

Figure 8. Unsupervised and Supervised model comparison.

 doi:10.20944/preprints202410.0927.v1

3.3. Comparison of Supervised and Unsupervised Results

Post-training, the supervised model described in section 3.2 is evaluated on two datasets: a separate test dataset to measure its generalization capabilities, and our Panksepp 7-emotions dataset. A classification report as shown in Table 3 provides detailed insights into the model's performance across all 7 Panksepp emotional categories. The model achieved high precision with the 'Rage' category (100%), indicating that all predictions for this class were correct, though the recall suggests that it missed a significant portion of actual 'Rage' cases. Conversely, the 'Fear' class showed an impressive recall of 94.29%, suggesting the model is highly sensitive in identifying this emotion but less precise, as indicated by a lower precision of 51.56%. This implies a higher number of false positives for 'Fear'.

The overall accuracy of the model stands at 74.32%, which reflects a solid ability to generalize across unseen data. The macro average F1-score of 74.90% suggests a balanced mean performance across classes, accounting for the imbalance in class distribution as indicated by the 'support' values. However, the disparity in performance among different emotions suggests that while the model is adept at recognizing certain emotional states (like 'Playing' and 'Rage'), it struggles with others such as 'Sadness' and 'Exploring'.

Comparing the supervised learning (see Table 2) scores to the results achieved by our approach based on the MoCo framework, we conclude that the contrastive learning results are promising especially for predicting 'Lust' followed by 'Rage' and 'Sadness'. 'Caring' however, compared to the supervised learning approach, seems to be harder to predict.

Table 2. Comparison of accuracy of ResNet50 to the accuracy of MoCo results on our own dataset.

Emotion	Accuracy ResNet50	Accuracy MoCo
Caring	94.74%	34.61%
Exploring	83.75%	40.40%
Fear	28.95%	35.51%
Lust	47.05%	62.79%
Playing	46.34%	38.88%
Rage	87.09%	45.91%
Sadness	78.57%	44.37%

Table 3. Further classification results of ResNet50 model on our own dataset.

Emotion	Precision	Recall	F1-Score
Caring	0.9796	0.7059	0.8205
Exploring	0.7609	0.6364	0.6931
Fear	0.5156	0.9429	0.6667
Lust	0.8125	0.7091	0.7573
Playing	0.7778	0.9403	0.8514
Rage	1.0000	0.7037	0.8216
Sadness	0.7600	0.5352	0.6281

Furthermore, we tested our CNN architecture with a publicly available dataset [37] containing 800 images representing each of the 4 emotional states 'Angry', 'Happy', 'Relaxed' and 'Sad'. Again, a classification report as shown in Table 5 provides detailed insights into the model's performance across all emotional categories.

The CNN demonstrated strong performance across all classifications, as evidenced by the precision, recall, and F1-score metrics for each category. Specifically, it achieved the highest precision with the 'Angry' emotion at 93.79%, indicating a high degree of accuracy in identifying this specific state. However, its recall for 'Angry' was lower at 77.04%, suggesting some instances were missed.

 doi:10.20944/preprints202410.0927.v1

The 'Happy emotion scored impressively in both precision (87.85%) and recall (94%), leading to the highest F1-score of 90.82% among all categories. The 'Relaxed and 'Sad emotions also showed robust results, with 'Sad exhibiting a notably high recall rate of 91.30%. Overall, the model achieved an accuracy of 85.12% on the dataset of 800 images, with the macro and weighted averages of precision and F1-score at around 85%. This indicates a consistent and balanced performance across different emotional expressions.

Again, comparing the supervised learning results (see Table 4) to the results achieved by our approach based on the MoCo framework, we can again conclude that the contrastive learning results are promising especially for predicting 'Angry' and 'Sad' dogs but still significantly less accurate than the supervised learning results.

Table 4. Comparison of accuracy of ResNet50 model to the accuracy of MoCo results on publicly available dataset [37].

Emotion	Accuracy ResNet50	Accuracy MoCo
Angry	81.08%	55.50 %
Happy	96.47%	45.65%
Relaxed	91.13%	35.90%
Sad	80.00%	54.55%

Table 5. Further classification results of ResNet50 model on publicly available dataset [37].

Emotion	Precision	Recall	F1-Score
Caring	0.9796	0.7059	0.8205
Exploring	0.7609	0.6364	0.6931
Fear	0.5156	0.9429	0.6667
Lust	0.8125	0.7091	0.7573

Comparing the different variants of the MoCo [12] framework, we found that contrastive learning show promising performance in the analysis of emotional behaviors in dogs. In comparison to the ResNet50 architecture which uses 23.9 million parameters, the ResNet18 and ResNet34 encoders use only 11.4 million and 21.5 million parameters. The primary requirements of any contrastive learning framework are a large batch size and a heavy encoder. In spite of this, the models that we built have shown promising results with their lighter encoders and computationally limited training environments [38].

3.4. Generalizability of Our Results

Our results demonstrate the strong potential of contrastive learning frameworks, particularly the MoCo variants, in analyzing emotional behaviors in dogs. The effectiveness of our models, despite using lighter encoders and operating within computational constraints, underscores the adaptability and robustness of these approaches. However, it's essential to recognize the limitations posed by the specific dataset and the potential variability in real-world scenarios. Future research should focus on evaluating these models across diverse datasets, including those representing different breeds and environments, to ensure broader applicability. Additionally, exploring transfer learning techniques could enhance the generalizability, allowing models trained on dog emotion datasets to be adapted for other species or even human emotional recognition tasks. This cross-species application could open new avenues in understanding and interpreting emotional behaviors using machine learning, fostering advancements in both animal behavior research and human-animal interaction studies.

The potential implications of the generalizability of our results are significant. If these models can be effectively applied across diverse datasets, it could lead to more accurate and nuanced understanding of dog emotions in various contexts, such as different breeds and environments. This

 doi:10.20944/preprints202410.0927.v1

could enhance the development of applications in areas such as animal welfare, training, and therapy. Moreover, exploring transfer learning techniques could enable the adaptation of these models for other species or human emotional recognition tasks. Such cross-species applications could advance our understanding of emotional behaviors more broadly, facilitating improvements in human-animal interaction studies and potentially contributing to the development of better tools for emotional recognition and response in both animals and humans.

4. Discussion

Our results have shown that unsupervised learning can achieve promising results comparable to supervised learning. Additionally, using unsupervised learning has the potential to identify new mood categories of dog emotions, beyond the seven emotional feelings of Panksepp [3]. By initially comparing prediction accuracy of five simple emotional categories first identified in [39] which are publicly available on Kaggle [37], we demonstrate that contrastive learning is a viable way to differentiate between emotional behaviours of mammals. Contrastive Learning is well-known for its few-shot learning capabilities [40].

The encoders used in this project are standard two-dimensional ResNet CNN's. The encoders are put through pretraining, and are able to produce good accuracies with a trivial downstream classifier such as KNN. In the future for practical applications for this technique, it will be feasible to use a dedicated fully connected downstream classifier neural network to boost accuracy scores. However, this is beyond the scope of this preliminary experimental study. With the advent of advanced vision transformer architectures, originally put forth by Dosovitskiy et al. [41], the encoders used for contrastive learning can produce much better results.

4.1. Broader Impact to the Field

The research paper on dog emotion recognition using contrastive learning is poised to significantly impact the broader field of animal behavior analysis and artificial intelligence. By introducing a novel dataset comprising images of ten popular dog breeds, each categorized into seven core emotional states, the study provides a robust foundation for future research. This comprehensive dataset, combined with the innovative application of the Momentum Contrast (MoCo) framework for unsupervised visual representation learning, contributes to the automatic recognition of canine emotions. The paper's approach not only bridges the gap between supervised and unsupervised learning methodologies but also demonstrates the feasibility and efficacy of using contrastive learning for emotion recognition in non-human subjects. This has the potential to enhance our understanding of canine emotions, fostering better human-dog interactions and contributing to the well-being of dogs by enabling more precise and empathetic responses to their emotional states.

Furthermore, the broader implications of this research extend to various practical applications, including the development of intelligent systems for dog training, therapy, and assistance, where accurate emotion recognition is crucial. By making expert knowledge accessible through AI-driven tools, the study democratizes the ability to understand and interpret dog emotions, benefiting novice pet owners, veterinarians, and animal behaviorists alike. The methodological advancements presented in the paper can also inspire similar approaches in the study of other animal species, thus broadening the scope of affective computing and ethology. As the paper integrates cutting-edge AI techniques with ethological insights, it paves the way for interdisciplinary collaborations that could change the way we study and interact with animals, ultimately promoting a deeper and more nuanced understanding of animal cognition and emotions.

4.2. Further Scope of This Research

This is also indicative of the prospects of fine-tuning these encoders. They can easily be imported with their current weights, and be fine-tuned to support identification of the seven Panksepp emotional behaviours [3] in various different domesticated mammalian species. Obvious species which could be supported with minimal fine-tuning and less training data include other four-legged

 doi:10.20944/preprints202410.0927.v1

furry animals such as horses, cats, cows and goats. 19 Another promising future avenue of research would be to explore transfer learning of this approach to humans. Dogs and humans have co-existed for several thousand years, which have introduced similarities within our emotional behaviours [42]. This could be demonstrated using a novel perspective of machine learning by comparing the encoders used for dogs and humans, and analysing their similarities and differences.

Researchers at University of California, Berkley identified 27 categories of emotions in humans [43]. Currently, canine emotions have not been explored in extensive depth, and our research aims to lay down a foundation on which extensive emotional research can be carried out on dogs with a new perspective and assistance from the cutting-edge methodologies developed in machine learning. The global population of dogs is steadily increasing, with estimates ranging from 900 million to one billion dogs in 2024 [44]. This calls for an increase in overall academic research conducted on canines. Our research aims to add an aspect of animal and computer interactions, to aid this research carried out on dogs.

4.3. Ethical Considerations

In our pursuit of advancing the understanding of canine emotions through contrastive learning, we must also consider the ethical implications of this research. Consequently, we adhere to the welfare-centered ethics framework introduced by [45] and guidelines from [46] that address ethical research in the ACI discipline.

Our research primarily focuses on improving animals' life quality [45] by enhancing the understanding of dogs' feelings, thus enabling more accurate handling of them. Additionally, we aim to foster interspecies relationships [45] by improving the mutual understanding and communication between humans and animals. Throughout this process, we ensure animals' welfare [47] at all times. Here, we outline the ethical considerations made before engaging in this research.

One of the primary ethical considerations revolves around the welfare of the animals involved. To ensure this, we designed our data collection process [47] to minimize any potential distress or harm to the dogs [47]. We conducted photographing and filming sessions in the dogs' natural or familiar environments, preferably done by their owner, avoiding any intrusive or stressful interventions [47].

Another critical ethical concern is the potential misuse of the technology developed from this research. To address this, we established clear guidelines and regulations to govern the application of this technology. These measures are designed to prevent the exploitation of the technology in ways that could harm the animals, such as inappropriate manipulation or commercialization of their emotions for entertainment purposes [46]. By implementing these guidelines, we ensure that the technology is used responsibly and ethically, enhancing the welfare and understanding of dogs rather than serving exploitative interests.

Lastly, we addressed ethical considerations related to the privacy and data protection of pet owners and their dogs. We implemented strict measures to handle potentially sensitive data, such as images and videos of the dogs [48]. Before collecting any data, we obtained informed consent from pet owners, clearly explaining the purpose, methods, and potential uses of the research. We anonymized the data to prevent the identification of individual dogs or owners and implemented secure storage protocols to safeguard the information from unauthorized access. As a result, we are unable to make our dataset publicly available.

By addressing these ethical concerns, our research contributes positively to the field of canine emotion recognition and the research field of ACI while maintaining the highest standards of ethical integrity.

5. Conclusions

We presented a Contrastive Learning approach for detecting the seven Panksepp emotions in dogs based on their body posture and facial emotions using Momentum Contrast for Unsupervised Visual Representation Learning. We achieved an accuracy of 43.2%, surpassing the 14% baseline on our self-curated dataset. We were able to draw conclusions about our model by comparing it with

supervised approaches applied to our dataset and the publicly available Kaggle dataset. Our approach offers valuable insights into dog emotion recognition without having to rely on labeled data. It also uses less parameters compared to supervised learning. Overall, our study demonstrates the feasibility of using unsupervised learning techniques for dog emotion recognition, providing a promising avenue for increasing our understanding of the emotional world of human's best friend.

References

1. Perri, A.R.; Feuerborn, T.R.; Frantz, L.A.F.; Larson, G.; Malhi, R.S.; Meltzer, D.J.; Witt, K.E. Dog Domestication and the Dual Dispersal of People and Dogs into the Americas. *Proc. Natl. Acad. Sci.* **2021**, *118*, e2010083118, doi:10.1073/pnas.2010083118.
2. Reed, C.A. Animal Domestication in the Prehistoric Near East: The Origins and History of Domestication Are Beginning to Emerge from Archeological Excavations. *Science* **1959**, *130*, 1629–1639, doi:10.1126/science.130.3389.1629.
3. Panksepp, J. Affective Consciousness: Core Emotional Feelings in Animals and Humans. *Conscious. Cogn.* **2005**, *14*, 30–80, doi:10.1016/j.concog.2004.10.004.
4. Chen, H.-Y.; Lin, C.-H.; Lai, J.-W.; Chan, Y.-K. Convolutional Neural Network-Based Automated System for Dog Tracking and Emotion Recognition in Video Surveillance. *Appl. Sci.* **2023**, *13*, 4596, doi:10.3390/app13074596.
5. Ferres, K.; Schloesser, T.; Gloor, P.A. Predicting Dog Emotions Based on Posture Analysis Using DeepLabCut. *Future Internet* **2022**, *14*, 97, doi:10.3390/fi14040097.
6. Chavez-Guerrero, V.O.; Perez-Espinosa, H.; Puga-Nathal, M.E.; Reyes-Meza, V. Classification of Domestic Dogs Emotional Behavior Using Computer Vision. *Comput. Sist.* **2022**, *26*, doi:10.13053/cys-26-1-4165.
7. Hernández-Luquin, F.; Escalante, H.J.; Villaseñor-Pineda, L.; Reyes-Meza, V.; Villaseñor-Pineda, L.; Pérez-Espinosa, H.; Reyes-Meza, V.; Escalante, H.J.; Gutierrez-Serafín, B. Dog Emotion Recognition from Images in the Wild: DEBIw Dataset and First Results. In Proceedings of the Proceedings of the Ninth International Conference on Animal-Computer Interaction; ACM: Newcastle-upon-Tyne United Kingdom, December 5 2022; pp. 1–13.
8. Broomé, S.; Feighelstein, M.; Zamansky, A.; Carreira Lencioni, G.; Haubro Andersen, P.; Pessanha, F.; Mahmoud, M.; Kjellström, H.; Salah, A.A. Going Deeper than Tracking: A Survey of Computer-Vision Based Recognition of Animal Pain and Emotions. *Int. J. Comput. Vis.* **2023**, *131*, 572–590, doi:10.1007/s11263-022-01716-3.
9. Hussain, A.; Ali, S.; Abdullah; Kim, H.-C. Activity Detection for the Wellbeing of Dogs Using Wearable Sensors Based on Deep Learning. *IEEE Access* **2022**, *10*, 53153–53163, doi:10.1109/ACCESS.2022.3174813.
10. Franzoni, V.; Milani, A.; Biondi, G.; Micheli, F. A Preliminary Work on Dog Emotion Recognition. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence - Companion Volume; ACM: Thessaloniki Greece, October 14 2019; pp. 91–96.
11. Kim, D.; Song, B.C. Contrastive Adversarial Learning for Person Independent Facial Emotion Recognition. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 5948–5956, doi:10.1609/aaai.v35i7.16743.
12. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning 2019.
13. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive Representation Learning: A Framework and Review. **2020**, doi:10.48550/ARXIV.2010.05113.
14. Shen, Z.; Liu, Z.; Liu, Z.; Savvides, M.; Darrell, T.; Xing, E. Un-Mix: Rethinking Image Mixtures for Unsupervised Visual Representation Learning. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 2216–2224, doi:10.1609/aaai.v36i2.20119.
15. Konok, V.; Nagy, K.; Miklósi, Á. How Do Humans Represent the Emotions of Dogs? The Resemblance between the Human Representation of the Canine and the Human Affective Space. *Appl. Anim. Behav. Sci.* **2015**, *162*, 37–46, doi:10.1016/j.applanim.2014.11.003.
16. Pasols, A. 20 Most Popular Dog Breeds (2024). *Forbes Advis.* 2024.
17. Kujala, M.V. Canine Emotions: Guidelines for Research. *Anim. Sentience* **2018**, *2*, doi:10.51291/2377-7478.1350.
18. Sharma, N.; Jain, V.; Mishra, A. An Analysis Of Convolutional Neural Networks For Image Classification. *Procedia Comput. Sci.* **2018**, *132*, 377–384, doi:10.1016/j.procs.2018.05.198.
19. El-Nouby, A.; Izacard, G.; Touvron, H.; Laptev, I.; Jegou, H.; Grave, E. Are Large-Scale Datasets Necessary for Self-Supervised Pre-Training? 2021.
20. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Multiview Coding 2019.
21. Yang, K.; Zhang, T.; Alhuzali, H.; Ananiadou, S. Cluster-Level Contrastive Learning for Emotion Recognition in Conversations. *IEEE Trans. Affect. Comput.* **2023**, *14*, 3269–3280, doi:10.1109/TAFFC.2023.3243463.

22. Zhang, D.; Nan, F.; Wei, X.; Li, S.; Zhu, H.; McKeown, K.; Nallapati, R.; Arnold, A.; Xiang, B. Supporting Clustering with Contrastive Learning 2021.
23. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies* **2020**, *9*, 2, doi:10.3390/technologies9010002.
24. Shen, X.; Liu, X.; Hu, X.; Zhang, D.; Song, S. Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition. *IEEE Trans. Affect. Comput.* **2023**, *14*, 2496–2511, doi:10.1109/TAFFC.2022.3164516.
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations 2020.
26. You, Y.; Gitman, I.; Ginsburg, B. Large Batch Training of Convolutional Networks 2017.
27. *Advances in Neural Information Processing Systems 29: 30th Annual Conference on Neural Information Processing Systems 2016: Barcelona, Spain, 5-10 December 2016*; Lee, D.D., Luxburg, U. von, Garnett, R., Sugiyama, M., Guyon, I., Neural Information Processing Systems Foundation, Eds.; Curran Associates, Inc: Red Hook, NY, 2017; ISBN 978-1-5108-3881-9.
28. Wang, X.; Zhang, H.; Huang, W.; Scott, M.R. Cross-Batch Memory for Embedding Learning 2019.
29. Choi, H.; Lee, B.H.; Chun, S.Y.; Lee, J. Towards Accelerating Model Parallelism in Distributed Deep Learning Systems. *PLOS ONE* **2023**, *18*, e0293338, doi:10.1371/journal.pone.0293338.
30. Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; Swersky, K. Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One 2019.
31. Sohoni, N.S.; Aberger, C.R.; Leszczynski, M.; Zhang, J.; Ré, C. Low-Memory Neural Network Training: A Technical Report 2019.
32. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved Baselines with Momentum Contrastive Learning 2020.
33. Chen, X.; Xie, S.; He, K. An Empirical Study of Training Self-Supervised Vision Transformers 2021.
34. Zając, M.; Zolna, K.; Jastrzębski, S. Split Batch Normalization: Improving Semi-Supervised Learning under Domain Shift 2019.
35. Li, S.; Zhao, Y.; Varma, R.; Salpekar, O.; Noordhuis, P.; Li, T.; Paszke, A.; Smith, J.; Vaughan, B.; Damania, P.; et al. PyTorch Distributed: Experiences on Accelerating Data Parallel Training 2020.
36. Wu, Z.; Xiong, Y.; Yu, S.; Lin, D. Unsupervised Feature Learning via Non-Parametric Instance-Level Discrimination 2018.
37. Balico, D. Dog Emotions.
38. Keshtmand, N.; Santos-Rodriguez, R.; Lawry, J. Understanding the Properties and Limitations of Contrastive Learning for Out-of-Distribution Detection 2022.
39. Ferres, K.; Schloesser, T.; Gloor, P.A. Predicting Dog Emotions Based on Posture Analysis Using DeepLabCut. *Future Internet* **2022**, *14*, 97, doi:10.3390/fi14040097.
40. Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; Zhang, L. Learning a Few-Shot Embedding Model with Contrastive Learning. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 8635–8643, doi:10.1609/aaai.v35i10.17047.
41. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale 2020.
42. Thompkins, A.M.; Lazarowski, L.; Ramaiahgari, B.; Gotoor, S.S.R.; Waggoner, P.; Denney, T.S.; Deshpande, G.; Katz, J.S. Dog–Human Social Relationship: Representation of Human Face Familiarity and Emotions in the Dog Brain. *Anim. Cogn.* **2021**, *24*, 251–266, doi:10.1007/s10071-021-01475-7.
43. Cowen, A.S.; Keltner, D. Self-Report Captures 27 Distinct Categories of Emotion Bridged by Continuous Gradients. *Proc. Natl. Acad. Sci.* **2017**, *114*, doi:10.1073/pnas.1702247114.
44. Cosgrove, N. *How Many Dogs Are There? US & Worldwide Statistics 2024*; 2024;
45. Mancini, C. Towards an Animal-Centred Ethics for Animal–Computer Interaction. *Int. J. Hum.-Comput. Stud.* **2017**, *98*, 221–233, doi:10.1016/j.ijhcs.2016.04.008.
46. Coghlan, S.; Parker, C. Harm to Nonhuman Animals from AI: A Systematic Account and Framework. *Philos. Technol.* **2023**, *36*, 25, doi:10.1007/s13347-023-00627-6.
47. Ilyena Hirskyj-Douglas; Read, J. The Ethics of How to Work with Dogs in Animal Computer Interaction. *Proc. Meas. Behav. 2016 Anim. Comput. Interact. Workshop* **2026**.
48. Paci, P.; Mancini, C.; Nuseibeh, B. The Case for Animal Privacy in the Design of Technologically Supported Environments. *Front. Vet. Sci.* **2022**, *8*, 784794, doi:10.3389/fvets.2021.784794.