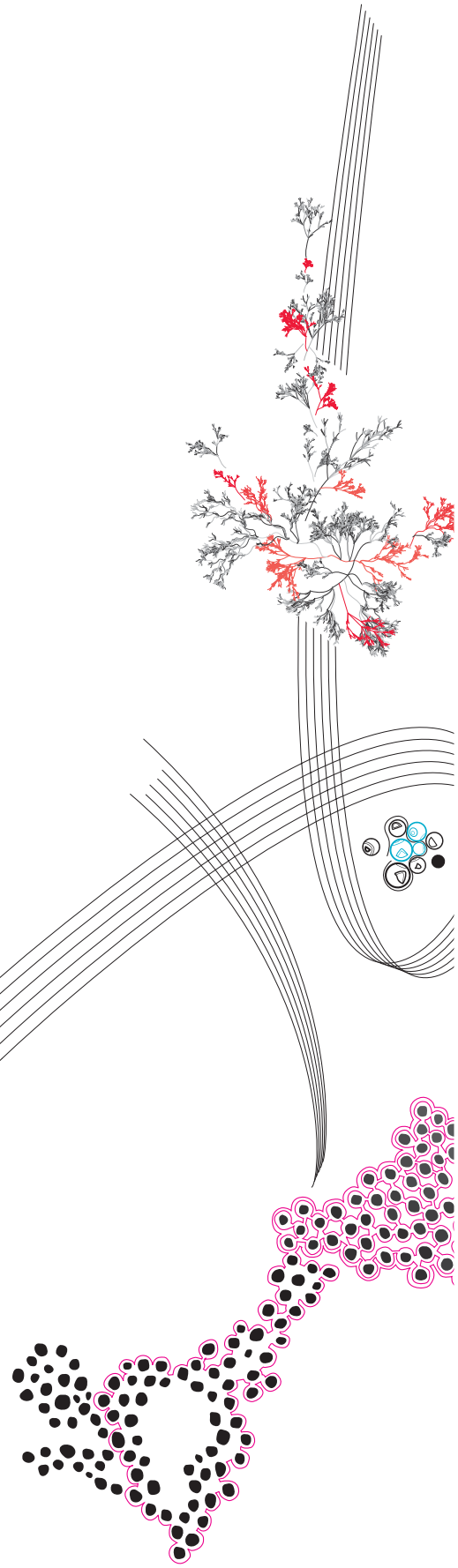


MSc Computer Science
Master thesis

Learning on a budget: tackling the cold-start problem in active learning for livestock behavior monitoring

Wannes Vanwinsen

Committee:
Dr. D. V. Le,
Dr. N. Strisciuglio,
Ir. G.P. Krabbenborg

December, 2024

Department of Computer Science
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Methods for recognizing cattle behavior via video often neglect the time-consuming and costly data annotation process, hindering practical and scalable implementation. Active learning can reduce this burden but is typically not as effective in early, low-data stages. Initial data selection is important yet challenging in these stages due to the lack of *a priori* information to select informative samples. This lack of initial knowledge is referred to as the *cold start* problem in active learning. This challenge becomes even more difficult in imbalanced classification tasks, such as livestock behavior monitoring.

This study explores cold start effects in active learning for livestock behavior monitoring and introduces a novel framework that combines self-supervised contrastive learning with active learning. The framework offers a structured way to select a representative and diverse initial training set for labeling by pre-training a model on a contrastive learning objective. Experiments were conducted on the CVB dataset, a benchmark for cattle behavior classification.

Results reveal that active learning strategies with and without self-supervised pre-training are sensitive to the size and diversity of the initial training set. The proposed framework failed to consistently outperform baseline models without pre-training and pre-training was found to have detrimental effects on model performance in most cases. The findings highlight that active learning's effectiveness depends on careful initial data selection and the selection of different sampling strategies across learning stages, emphasizing that the cold start problem remains a challenging problem in active learning.

Keywords: livestock monitoring, behavior recognition, active learning, cold start, video representation learning, contrastive learning

Chapter 1

Notations

Active learning: functions and variables

- D: dataset
- L_t: the subset of labeled data used for model training at the t'th active learning cycle. Where L_0 is the initial set of labeled training data
- U_t: the subset of unlabeled data used that is queried from at the t'th active learning cycle. Where $L_t \cap U_t = \emptyset$ due to the transfer of data between sets
- B_t: the budget size, or number of samples to be selected from U_t at the t'th active learning cycle
- q: query/acquisition function used to evaluate and propose the subset of label candidates from the unlabeled data set
- Q_t: the subset of U_t selected by the query function.
- $\sigma_{softmax}$: the softmax activation function
- $p_{softmax}$: the probability distribution derived from the softmax activation function
- ϕ_{enc}: encoder function
- ϕ_{clf}: classification head function.
- m: model function, defined as the composition $(\phi_{\text{enc}} \circ \phi_{\text{clf}})$ of the encoder- and classification head function
- $\mathcal{Z}$: a set of representations/embeddings learned by the encoder function
- θ: model parameters
- $\mathcal{L}$: loss function

Chapter 2

Introduction

As early as 1997, the need to transition from traditional livestock farming to precision livestock farming (PLF) was recognized and investigated [1]. Due to a rising global demand for dairy products, coupled with urbanization and a declining farming workforce, the average herd size in the Netherlands has increased from 57 to 114 cows per farm since the year 2000 [2]. Consequently, automated PLF systems are becoming increasingly important to uphold productivity, sustainability, and animal welfare on the dairy farm [3].

Video monitoring systems can provide continuous surveillance and analysis of cattle behavior. In current scientific literature, different automated methods for video behavior recognition of cattle have been proposed [4]. However, such studies often forego expensive and time-consuming data collection and annotation. Annotating behavioral data for livestock is not a trivial task, as it requires a lot of time and domain-specific knowledge. As with many vision-based learning tasks, the required data is large with diminishing returns and bridging the gap between the experimental setting and scalable, real-world solutions is not sustainable using typical supervised learning practices [5].

Active learning strategies show promise to ease this annotation burden. However, most active learning studies rely on pre-trained models that have already learned the underlying task sufficiently to be effective [6, 7, 8, 9]. Empirical studies show that active learning strategies often fail to identify informative samples during the early stages of model training due to a lack of a sufficiently learned data distribution [10, 11, 12].

The lack of *a priori* information in active learning is called the *cold-start problem*. This problem poses a challenge for use cases like video-based livestock monitoring due to multiple reasons: (1) the data is expensive to annotate due to the need for expert knowledge (2) the high dimensionality of video data, which often necessitates large datasets to achieve desirable performance, and (3) the imbalanced nature of livestock behaviors, where many behaviors are rarely observed. As a result, random selection of the initial training set often leads to uninformative or insufficiently diverse samples, which can negatively impact the effectiveness of active learning in later stages.

Prior works on the cold-start problem have achieved promising results on image classification tasks by using self-supervised contrastive learning (CL) to overcome the initial data gap [11, 13, 10]. However, these methods do not investigate or address the effectivity of such approaches in imbalanced classification problems. In addition, contrary to image classification, accurate monitoring of livestock behavior benefits from learning spatiotemporal relations [14, 15, 16, 17]. Representation learning methods for video sequences differ from those used for static images as they need to accommodate for the additional temporal dimension [18, 19].

To address these gaps, this research proposes and evaluates a framework that addresses

the cold-start problem for video-based livestock monitoring. It provides an analysis of cold start effects and the conditions in which they occur when applying active learning to highly imbalanced video data. Furthermore, the research explores the potential of self-supervised pre-training to reduce cold-start effects. By combining representation learning with active learning, this work aims to improve early-stage model performance and contribute to more sustainable and scalable deep-learning solutions for video-based livestock monitoring.

2.1 Research outline

This research builds upon previous work in active learning, in which the cold-start problem is addressed through self-supervised contrastive learning methods. The research investigates the conditions under which cold-start effects occur for different active learning strategies when applied to the behavior classification task. Furthermore, a framework for early-stage active learning is proposed and evaluated. In this framework, self-supervised contrastive learning is applied before active learning, after which the learned representations are leveraged to guide a more informed selection of the initial training set.

To this end, we formulate the following research question:

RQ **How can we design an active learning framework that addresses the cold-start problem using limited labeled data in livestock monitoring scenarios?**

The following sub-questions are complementary to answering said research question:

SQ1 *What is the impact of the selection and size of the initial training seed on the effectiveness of different active learning strategies?*

SQ2 *How does integrating self-supervised contrastive learning influence model generalization and active learning outcomes in video-based livestock monitoring?*

SQ3 *To what extent can features derived from contrastive learning improve the initial seed selection process compared to random sampling?*

In chapter 3 relevant background information is given about precision livestock farming and some of the techniques and algorithms used in this research. Relevant prior work is discussed in chapter 4, followed by formalizing a framework for early-stage active learning in chapter 5. Results of the experimental evaluation are reported in chapter 6, with supplementary analysis and discussion in chapter 7. Finally, the limitations and potential future directions of the research are addressed in chapter 8, followed by some concluding remarks in chapter 9.

Chapter 3

Research background

This chapter is divided into three sections. In the first two sections, we briefly discuss the importance of monitoring livestock behavior as an indicator of animal welfare. In the second section, some background information on active learning is given.

3.1 The dairy industry: the Netherlands as leading dairy country

The dairy industry is a major part of the agri-food industry. It is a global industry where fresh milk is mostly consumed locally and other products such as cheese or milk powder are traded on a global export market. Within Europe, the Netherlands are a very important milk-producing country with an annual production value of roughly 6.4 billion euro and a turnover of 10 billion euro. Due to the limited space in the Netherlands, farmers are forced to practice a specialized way of dairy production with a strong focus on a high productivity per cow.

In recent decades, there has been an increasing emphasis on sustainable practices within the Dutch dairy sector. This shift is largely in response to the environmental challenges posed by intensive farming methods. The concentration of livestock has led to concerns regarding manure management. The sheer volume of manure generated can exceed the natural absorption capacity of the land, resulting in nutrient runoff that contaminates water bodies, contributing to eutrophication and other ecological disturbances [20]. In addition, dairy production contributes heavily to greenhouse gas emissions. The industry has been striving to implement practices that mitigate these effects, such as improving feed efficiency, optimizing manure management, adopting circular farming practices and a strong emphasis on animal health. These changes translate to higher milk yield and reduced costs for the farmer according to latest research [21].

Despite these efforts, the Dutch dairy industry often approaches and occasionally exceeds environmental limits. The balance between maximizing production and ensuring environmental sustainability is delicate and requires continuous innovation.

3.2 Behavior: a welfare and productivity indicator for livestock

Most dairy farmers in the Netherlands try to optimize their production with a strong focus on animal health, the extension of the cow's productive life, and animal well-being. For the productivity and well-being of dairy cows, the following parameters are of importance:

- **Fertility**: one of the main goals for dairy farmers is to achieve one calf per cow per year. This is important for maintaining milk production levels, as cows need to calve regularly to produce milk. Heath detection plays a large role in this. Detecting when a cow is in heat (oestrus) is important for successful breeding. Missed heat cycles can lead to extended calving intervals, reducing overall milk production. During estrus, cows often exhibit increased physical activity and may show changes in other behaviors, such as increased restlessness or mounting behavior [22, 23].
- **Hoof health (lameness)**: lameness, often caused by lesions in the hoof of the cow, most often results in decreased productivity and well-being in dairy cows [24]. Lame cows often have difficulty walking, which affects their ability to feed and reproduce. Proper hoof care and management can significantly extend a cow's productive life.
- **Feeding:** tailored feeding can maximize milk yield and quality. Adjusting feed based on the cow's lactation stage, activity level, or milk production rate ensures that cows have the energy and nutrients required for optimal milk production. In addition, abnormal feeding patterns, such as reduced feed intake, have been linked to certain health conditions [25].
- **Digestion:** cows with healthy rumination patterns are more likely to have a healthy body condition and higher fertility [26]. Reduced rumination time is often connected to digestive issues in ruminants. Changes in rumination have been associated with health problems and, in some cases, with changes in productivity, such as milk yield.
- **Udder health:** inflammation of the udder tissue (Mastitis) is one of the leading health issues affecting dairy cows. It can lead to reduced productivity and well-being and can result in severe health complications for the animal [27].
- **Lying time:** changes in lying time can indicate physical discomfort or fatigue, though it may also vary based on environmental and management factors [28].

Most if not all of these parameters influence the behavior of the cows and can be detected with behavior monitoring [29]. Combined with content and hormone analysis of the produced milk, the aim is not only to detect anomalies but to predict them earlier to allow the farmer to react sooner. This allows for fast intervention which reduces treatment costs and prevents production losses.

However, manually monitoring behavior across large herds is challenging, time-consuming, and open to a subjective interpretation. Automated monitoring through sensors offers a potential solution to this problem, enabling continuous and scalable animal behavior monitoring.

3.3 Batch active learning

Batch active learning has been proposed as an alternative to traditional supervised learning, particularly in scenarios where labeling data is expensive or time-consuming. Unlike supervised learning, which relies on randomly selected or fully labeled datasets, batch active learning focuses on selecting a subset or batch of the most informative samples for annotation, given a fixed labeling budget.

The primary goal is to maximize the utility of the labeled data by ensuring that the selected batch contributes significantly to improving the model's performance. This approach is especially beneficial in domains with high data annotation costs, such as medical applications, autonomous driving, or livestock monitoring.

The selection of these informative samples is guided by specific strategies designed to optimize the batch's utility. These strategies often rely on measures of uncertainty, such as prediction confidence, or diversity, which aims to select a diverse set of samples representative of the broader dataset. However, the scope of active learning strategies extends beyond these metrics, incorporating approaches that integrate domain knowledge or even hybrid combinations of multiple criteria to improve the selection of samples [30]. In this research, three distinct active learning strategies are evaluated. A brief explanation of each strategy is provided below:

3.3.1 Max-Entropy

One commonly used active learning strategy is entropy-based sampling. A strategy that is used to select a batch of high-uncertainty samples. For a classification model with K classes, let $p_{\text{softmax}}(y|x;\theta)$ represent the predicted probability distribution of the sample x. The *entropy* of this distribution is then given by:

$$H(y \mid \mathbf{x};\theta) = -\sum_{k=1}^{K} p(y=k \mid \mathbf{x};\theta) \log p(y=k \mid \mathbf{x};\theta),$$

A batch of B samples, referred to as the query $Q_t : \{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_B^*\}$ is selected by maximizing the total entropy for the batch from the unlabeled pool U:

$$\{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_b^*\} = \arg \max_{Q \subseteq U_t, |Q|=B} \sum_{\mathbf{x} \in Q} H(y \mid \mathbf{x};\theta),$$

3.3.2 Cluster-Margin

The Cluster Margin strategy selects a batch of samples that are closest to the decision boundary while using an additional diversity constraint. Let U_t be the unlabeled pool at iteration t, and Q_t the batch of B samples to be selected. Using similar notation and definitions as for Max-Entropy, the uncertainty margin for a sample x is defined as the difference between the two highest predicted probabilities p^1_{softmax} and p^2_{softmax}

$$\text{Margin}(x) = p^1_{\text{softmax}} - p^2_{\text{softmax}}$$

To enforce further diversity constraints, Cluster-Margin uses the embeddings of the unlabeled set ($\mathcal{Z}_{U_t}$) and divides the space into C clusters using hierarchical agglomerative clustering. They select the k_m samples with the lowest margin and iteratively traverse across clusters to select label candidates from these low-margin samples. More details on the implementation are given Citovsky et al.[31].

3.3.3 Core-Set

The Core-Set strategy, proposed by Sener et al. [12] poses the sample selection as a k-center optimization problem. Unlike uncertainty-based methods, Core-Set relies on the idea of minimizing the distance between the selected batch and the rest of the data in a feature space.

Given labeled and unlabeled sets L_t and U_t, let $z_i \in \mathrm{R}^d$ represent a feature embedding of the model's penultimate layer. Core-Set aims to select a batch $Q_t \subseteq U_t$ that minimizes

the maximum distance of any unlabeled sample to its closest selected sample in Q_t. This can be formalized as:

$$\{\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_b^*\} = \arg \min_{Q \subseteq U_t, |Q_t| = B} \max_{\mathbf{x} \in U_t} \min_{\mathbf{q} \in Q} \|z_x - z_q)\|_2,$$

Since solving the above optimization problem exactly is often computationally expensive, a greedy approximation is used. The batch is constructed iteratively by adding one sample at a time that maximizes the minimum distance to the already selected samples. At the i-th step, the next sample is selected as follows, where Q_{i-1} is the set of $i-1$ samples already selected in the batch.

$$\mathbf{q}_i = \arg \max_{\mathbf{x} \in U_t \setminus Q_{i-1}} \min_{\mathbf{q} \in Q_{i-1}} \|z_x - z_q\|_2,$$

Chapter 4

Related work

This work finds its application in precision livestock farming through video-based monitoring (1). Its main focus is to reduce the data annotation effort by combining ideas from two separate but related domains in deep learning research: active learning strategies (2) and contrastive learning for videos (3). We discuss relevant prior work for these topics and identify some of the gaps existing in current literature.

4.1 Video-based livestock monitoring

An extensive body of research exists around recognizing livestock behavior from video [4]. For example, Avola et al. explored skeleton-based action recognition using a two-branch stacked LSTM-RNN's [32]. Guo et al. [16] train a bidirectional GRU model to learn spatiotemporal features from cattle videos and identify common behaviors such as feeding, grooming and walking. Jiang et al. conducted a study on a convolution network for recognizing the anomalies in movement associated with lameness in dairy cows [33]. Lodkaew et al. [34] proposed a pipeline for estrus detection that tracks key points to detect mounting and sniffing behaviors characteristic of cows in heat. Similarly, Wang et al. adapt the YOLOv5 model to detect mounting behaviors in cows with high recognition accuracy [35]. Fuentes et al. developed a deep learning approach for recognizing hierarchical cattle behavior using spatiotemporal information [36]. Later, the same authors extended their work by presenting methods for tracking and re-identification of the individual cows [15]. Wu et al. [17] propose a method for monitoring the respiratory behavior of multiple cows using techniques such as optical flow and phase-based motion magnification. Bai et al. [37] proposed and evaluated the X3DFast model, a lightweight and efficient approach for classifying dairy cow behaviors by combining elements of X3D [38] and SlowFast [39] architectures.

A limiting factor in the existing research is that the datasets used for experimental evaluation are often small, not disclosed, or highly curated to a particular farm or cow breed [14, 40]. Annotating behavioral data for livestock is no trivial task, as recognizing and labeling cattle behavior from video requires a certain degree of domain-specific knowledge. In addition, empirical and theoretical studies in deep learning suggest that performance improvements diminish as models get larger and are trained on more data [5]. Consequently, bridging the gap between experimental- and commercially viable and well-performing solutions requires significant investment of resources. Alternative learning paradigms such as active learning could be more cost-effective by accelerating the development of livestock monitoring solutions and reducing the needed resources.

4.2 Active learning for animal monitoring

To the best of our knowledge, no evaluation of active learning strategies for video-based livestock monitoring has been conducted in the literature. However, different active learning frameworks have been applied to related problems such as *animal activity recognition* [6, 7, 8, 9].

4.2.1 Active learning frameworks

Most research regarding active learning concerns itself with proposing new criteria for evaluating the *value* of unlabeled data. After evaluation, a query strategy decides whether one sample has sufficient value to be considered for labeling by the annotator. Ultimately, an effective query strategy aims to reduce labeling effort by selecting samples that yield the best *improvement* given a data budget. Query strategies can be divided into three categories: *uncertainty-based*, *representation-based* and *hybrid methods*.

Uncertainty-based methods

Some of the earliest and most commonly applied strategies are based on the confidence of the prediction, represented by the softmax probabilities produced as the output of a classifier. There are different ways to select samples based on model confidence: *Least Confidence* selects samples with the lowest posterior probability of the predicted label. The *Margin* method selects samples that have small differences between the top-k predictions [41]. *Entropy*-methods [42] select samples with the highest entropy across predictions. Confidence-based methods have their limitations as models can quickly become over-confident in their predictions during early stages or when limited data is available, resulting in unstable performance [30, 10].

Recognizing this limitation, Gal et al. [43] proposed to use Bayesian CNNs to get a more accurate approximation of the true posterior probabilities. They use Monte Carlo (MC)-dropout to quantify uncertainty in neural networks.

Generative and adversarial methods also made their way into active learning. Zhu et al. use GAN's to generate samples with high prediction uncertainty rather than selecting them from unlabeled data [44]. Similarly, Tran et al. [45] propose to generate data that lies at the intersection of decision boundaries between classes. Other prior work proposes task- and model-agnostic approaches by estimating the influence of a sample on the test loss [46, 47]. While these methods report promising results they do impose higher computational requirements. For example, with MC-dropout, the computational load increases proportionally to the number of forward passes used to derive an estimate for uncertainty. GAN's likewise increase the computational load by requiring an additional model for sample generation.

Representation-based methods

A point of critique for uncertainty-based selection is that they tend to over-signify the selection of *difficult* data. As a result, the query contains a small subset of noisy/outlier data that might not represent the majority of the data. Contrary to this, representation-based methods emphasize the selection of typical and diverse data while avoiding redundancy.

Yin et al. [48] propose to minimize redundancy by calculating distance measures between the features of the query and the labeled set. The authors of CoreSet [12] pose active learning as a k-center clustering problem and use a greedy optimization algorithm to iteratively select samples that are farthest from the labeled samples in the latent space. Other

work proposes similar ideas by optimizing different criteria like probability coverage [49] and Wasserstein-distance [50]. Adversarial learning has also been used in representation-based active learning. Dual-optimization between auto-encoders and discriminators is used to minimize the difference between the labeled data distribution and unlabelled pool distribution [51, 52, 53].

Hybrid methods

Some active learning frameworks combine criteria from both uncertainty- and representation-based strategies. Many authors have proposed to add additional diversity constraints to uncertainty sampling, either through unsupervised clustering or intra/inter-class diversity measures [54, 55, 31]. Kirsch et al. [56] propose an extension of BALD by approximating the mutual information between a batch of points and model parameters and using dependencies within the queried batch to remove redundant samples. Wang et al. [57] propose to exploit the abundance of unlabelled data using a hybrid learning strategy. They select low-confidence samples for active learning while assigning pseudo-labels to samples with high prediction confidence. Gao et al. [58] leverage unlabelled data differently by minimizing a consistency loss on augmented versions of the unlabelled data and selecting samples with low consistency for active learning.

4.2.2 Limitations of active learning frameworks

Uncertainty-based methods rely on a non-biased, truthful estimation of uncertainty by the model. Recent works have shown that uncertainty measures are particularly unsuited in cases where the data budget is low, or when dealing with high dimensional data [13, 12, 11, 49]. Both of these limitations apply to video-based livestock monitoring. Uncertainty metrics are often not as effective when the model has not sufficiently learned the underlying data distribution [10]. While representation-based methods are less dependent on model output, they equally depend on initial knowledge learned by the model to be effective. A prerequisite for these approaches is to learn robust representations that encode information relevant to the downstream task while being robust to variance or biases in the data.

Without this *a priori* knowledge, most active learning strategies often fail to outperform random selection. In most active learning frameworks, this initial knowledge gap is filled by randomly selecting an initial batch for supervised training. However, the effectiveness of active learning strategies is highly inconsistent depending on this initially selected batch [11]. This inefficacy in the early stage of active learning is called the *cold-start* problem. Ultimately, active learning fails to reduce annotation efforts consistently when this problem is not addressed.

4.2.3 The cold start problem in active learning

The *cold-start* problem was originally used to describe the lack of a priori information in recommender systems. However, it extends beyond such systems and has been mentioned as one of the core challenges in active learning as well [30]. Little research into the cold start problem in active learning has been conducted outside of the field of natural language processing [59].

Recent work has studied and addressed the cold start problem in image classification. Some approaches kick-start the learning process without the need for initial supervised training by using contrastive learning on unlabeled images. [11, 13, 60]. The latent space is then clustered and annotated with pseudo-labels, after which representative samples are

selected from each cluster for labeling. This strategy aims to select 'prototypical' training examples with additional diversity constraints by leveraging the grouping of features that automatically form through contrastive learning.

Haconen et al. [10] provide a theoretical analysis of the effectiveness of different active learning strategies under different data budgets and empirically validate a new active learning strategy to reduce cold-start effects. Yehuda et al.[49] propose a new querying strategy for early active learning that maximizes probability coverage of the latent space. Most recently, Samet et al. [61] proposed to find the best initial query by posing the subset selection as a linear optimization problem. They select samples that maximize diversity between and within samples and evaluate this method on the task of 3D semantic segmentation.

In video-based livestock monitoring, inference is often applied to sequences of frames rather than static images. Spatio-temporal features are more capable of capturing long-term, complex behavior than static features derived from images as they can accommodate for motion over time [16, 15, 37]. Self-supervised learning methods for images are less effective when applying them to video sequences due to the additional temporal dimension and redundant information found in video [62]. Image-based augmentation methods risk context bias [18] which are present in fixed environments such as free-stall barns. Addressing the cold-start problem for video data could benefit from an adaptation of the contrastive learning method to learn better video representations.

4.3 Contrastive learning for video representations

A large body of research already exists on contrastive learning for video representations [63]. Contrastive learning provides a self-supervised way to push positive input pairs closer together while pushing negative input pairs farther apart. In images, positive and negative samples were often formed by applying some augmentation to the same image (rotation, cropping, color jitter et cetera) to create positive sample pairs and select unrelated images as negative pairs. While these techniques have also seen effective use in video-based learning, augmentations going beyond static image augmentation can, in some cases, outperform the latter by learning more robust spatiotemporal features [63]. For example, by selecting positive samples from the same video at a random offset from the anchor video.

Some research has proposed to use optical flow as an additional supervisory signal to encode motion cues into the learned embeddings, either by learning to predict optical flow descriptors [64, 65], using optical flow for positive sample mining [66] or by learning to assign RGB and optical flow input to the same consistent prototypes [19]. The downside of such methods is that while they report promising results, they create a large additional computational overhead for the generation of the optical flow vectors or by introducing dual optimization schemes of multiple models.

Others combine video pre-text tasks with a contrastive learning loss. Hu et al. [67] apply temporal matching and ordering to push segments from the same video together while simultaneously predicting the temporal ordering of the clips. Huang et al. [18] exploit the motion vectors and keyframes used in compressed H264 videos and combine context matching and motion prediction with contrastive learning.

Singh et al. [68] apply a contrastive loss to maximize agreement between the same videos at different playback rates. Lorre et al. [69] propose to apply contrastive predictive coding, in which a model is trained to predict the latent representation of future video segments while using a contrastive loss function to ensure that it can distinguish between correct and incorrect future representations.

When reviewing the available work on self-supervised video representation learning, the findings are threefold:

1. Downstream video learning tasks can benefit from spatiotemporal augmentations beyond static image augmentations or introducing a supervisory temporal signal during self-supervised training [63].

2. Representations derived from contrastive learning inherently form clusters with diversity constraints due to the characteristics of the instance discrimination task [70].

3. The maximization of agreement in contrastive learning within clusters provides a quantitative metric to select *diverse* and *prototypical* samples within each subset [13, 10].

4.4 Key findings

Annotating livestock behavior data is difficult and time-consuming, necessitating cost-effective solutions to bridge the gap between experimental and practical applications. While active learning can ease the annotation effort, such strategies suffer from the cold-start problem when limited data is available. As far as is known, no research has been done to investigate such cold-start effects in the context of imbalanced video classification problems. Research that addresses the cold start problem in image classification has achieved promising results by leveraging contrastive learning. When applying such works to videos, the approach could benefit from adapting the contrastive learning method to the temporal dimension inherent in video data.

Chapter 5

Research methods

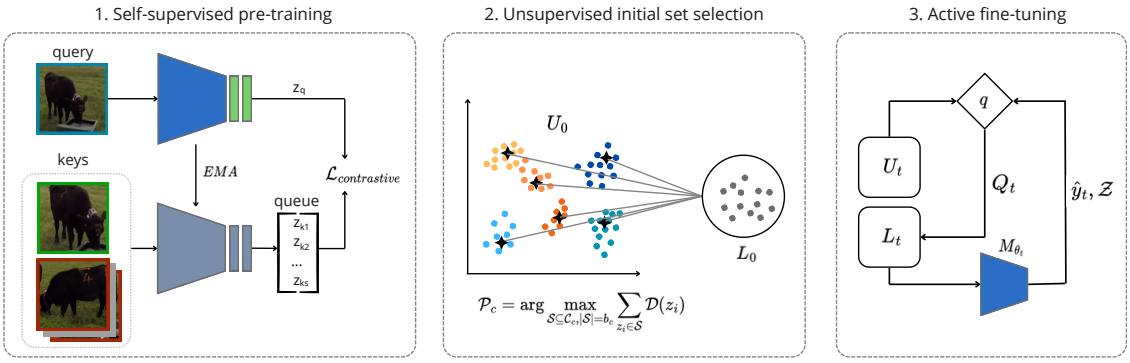


FIGURE 5.1: A high-level overview of the proposed framework. 1) Self-supervised contrastive learning is applied to learn discriminative groupings of features from the input data without downstream annotations. 2) Pseudo-labels are assigned to the embeddings using an unsupervised clustering algorithm. From each cluster, a representative sample is selected by maximizing the typicality (the inverse Euclidean distance to the nearest neighbors). 3) The initial seed is labeled and used as a starting point for different active learning strategies.

5.1 Proposed active learning framework

Inspired by recent findings in active learning research [13, 10] and video representation learning [63], this research extend their methodologies to behavior classification from video and provides an analysis on a new use-case: livestock behavior recognition from video.

A high-level overview of the proposed framework has been visualized in figure 5.1 and consists of the following stages: *1) self-supervised pre-training 2) unsupervised initial seed selection 3) active fine-tuning.*

5.1.1 Self-supervised pre-training

In this framework, self-supervised contrastive learning is used to learn such initial representations of the data, using the data itself as a supervisory signal. This can be achieved through *instance discrimination*. A task in which a model learns to identify augmented versions of the same sample (positive examples) from a set of augmented versions of other samples (negative examples). Positive pairs are generated from augmented views of the

same video, while negative pairs are sampled from augmented views of different videos in the dataset.

Given a dataset D with N videos, e.g. $D = x_1, x_2, ..., x_N$ and some augmentation function $\mathcal{T}(x_i; t)$, applied to the data. Let $z_i \in \mathbb{R}^h$ be the vector learned by the model to represent sample x_i and let z_p be the representation of x_i under some different kind of augmentation (e.g. sampled with an offset, cropped or flipped around the horizontal axis). Now, let $\mathcal{N}_i$ be the set of representations of the negative set for sample x_i. Finally, let τ represent a smoothing factor commonly referred to as the temperature constant of the loss. The contrastive loss can be described using equation 5.1.

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp\left(z_i \cdot z_p/\tau\right)}{\exp\left(z_i \cdot z_p/\tau\right) + \sum_{n\in\mathcal{N}_i} \exp\left(z_i \cdot z_n/\tau\right)}\right] \tag{5.1}$$

The underlying motivation for using self-supervised learning before active learning is that it enables the model to learn initial data representations without relying on behavior annotations as a supervisory signal. This makes it particularly valuable for tasks where the amount of labeled data is limited.

The model is optimized to learn meaningful features from the data to perform the instance discrimination task well. Thus, the hypothesis is that this prior knowledge can serve as a good foundation for the downstream task and potentially reduce cold-start effects as seen in the earlier stages of active learning.

5.1.2 Unsupervised initial seed selection

One of the potentially advantageous effects of self-supervised contrastive learning is that the model is encouraged to construct an embedding space where distances reflect semantically meaningful differences in the data [70]. As a result, the embeddings ideally form groupings of similar data points—given that the model has learned this discrimination task sufficiently.

Within these groups, the point with the highest density or centrality could be described as a representative prototype, capturing the main characteristics of each group [10]. The hypothesis is that by identifying these prototypes, it is possible to select a subset from the data that is both diversified and representative of the underlying data distribution. Both of these are important criteria for selecting an initial training set and could be used as a more informed alternative to randomly sampling the initial training data.

An unsupervised clustering algorithm such as k-Means can be used to divide the embedding space into C clusters, where C can be optimized to capture each of these prototypical groups. The prototypes of each group can then be identified through the use of some adjacency or density measure such as the inverse Euclidean distance to the other members of each cluster (termed the *typicality* of a sample).

This initial sampling strategy can be formalized as follows: Let $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$ represent the representation vectors corresponding to each video, where $z_i \in \mathbb{R}^d$. An unsupervised clustering algorithm (e.g., k-Means) is used to partition $\mathcal{Z}$ into C clusters as seen in equation 5.2

$$\bigcup_{c=1}^{C} \mathcal{C}_c = \mathcal{Z}, \quad \mathcal{C}_i \cap \mathcal{C}_j = \emptyset \quad \text{for } i \neq j. \tag{5.2}$$

For a given cluster $\mathcal{C}_c$, the goal is to identify prototypes $\mathcal{P}_c \subseteq \mathcal{C}_c$ by maximizing the inverse Euclidean distance to the k-nearest neighbors within the same cluster. The inverse

Euclidean distance for a sample $z_i \in \mathcal{C}_c$ is defined in equation 5.3.

$$\mathcal{D}(z_i) = \sum_{z_j \in \mathcal{N}_k(z_i, \mathcal{C}_c)} \frac{1}{\|z_i - z_j\|_2} \tag{5.3}$$

where $\mathcal{N}_k(z_i, \mathcal{C}_c)$ represents the set of k-nearest neighbors. The initial sampling strategy can then be defined using equation 5.4. Where b_c is the number of prototypes to select from cluster $\mathcal{C}_c$. The samples corresponding to the prototypes selected from each cluster are combined into the initial labeled training set L_0 with initial budget size B_0.

$$\mathcal{P}_c = \arg \max_{\mathcal{S} \subseteq \mathcal{C}_c, |\mathcal{S}| = b_c} \sum_{z_i \in \mathcal{S}} \mathcal{D}(z_i), \tag{5.4}$$

Lastly, two additional parameters are used to constrain the optimization problem:

- N_{max} defines the maximum number of nearest neighbors included in the computation of the typicality score. Constraining the number of neighbors ensures that local density variations are not diluted, which allows the identification of multiple high-density (sub-)prototypes within the same cluster.
- $\mathcal{S}_{min}$ is the minimal number of members that a cluster needs to have to be considered for sampling. This parameter is used to reduce the potential accumulation of outliers from small, unrepresentative clusters in the representation space.

5.1.3 Active fine-tuning

At the start of each active learning cycle, a query $Q_t \subseteq U_{t-1}, |Q_t| = B_t$ is selected from the unlabeled dataset and added to the labeled set $L_t \leftarrow L_{t-1} \cup Q$ and $U_t \leftarrow U_{t-1} \setminus Q$. This newly labeled dataset is then used for model training in a supervised manner.

This query set is selected based on several criteria. Depending on the strategy, these criteria are based on either the model output or some intermediate representation of the data, such as the embeddings at the final encoder layer. The downside of using such model-specific criteria is that they can be biased or unreliable if the model is not sufficiently adapted to the data.

As mentioned previously, active learning frameworks therefore rely on an initial supervised training phase to be effective. This initial phase can take up a considerable amount of resources depending on the characteristics of the dataset and the difficulty of the task to be learned. Furthermore, stopping criteria for transitioning from initial supervised training to active learning are not well-defined or task-specific, potentially wasting resources.

With the integration of pre-training and informed initial seed selection, the hypothesis is that the data requirements of this initial supervised training phase can be reduced or even left out due to the additional *a priori* knowledge of the data.

FIGURE 5.2: Representative frames from the CVB dataset showing three distinct behavior classes.

5.2 Dataset

The proposed framework is evaluated on the CVB dataset for cattle behavior recognition [40]. It was proposed as a benchmark for research in livestock management to provide a transparent and fair comparison across research efforts.

The dataset contains 502 videos of multiple cows recorded in a pasture environment from four different camera perspectives. Each frame has been annotated with bounding box annotations as well as a distinct behavior label. Consecutive bounding boxes of the same cow are given a unique identifier to keep track of each animal across a video. Figure 5.2 depicts one of the frames and some examples of the bounding box annotations found in the dataset.

5.2.1 Behavior distribution

The authors have included 11 distinct behavior labels, of which 9 are potentially relevant indicators for cattle health as discussed in Chapter 2.1. These behaviors are *grazing, resting while lying, ruminating while lying, resting while standing, ruminating while standing, walking, running, drinking* and *grooming*. The remaining two behaviors (*hidden* and *other*) are more ambiguous and contain all observations that were either too difficult to assess due to occlusion by other cows or could not be assigned to any of the behaviors mentioned above.

Cattle naturally exhibit some behaviors more frequently than others. This has been reflected in the class distribution of the dataset. Cows in pasture environments tend to spend the majority of their time passively grazing, resting, or ruminating which cumulatively make up 72% of the available annotations. The remainder of the annotations belong to more infrequent types of behavior such as drinking, grooming, or walking. The distribution of these behaviors has been visualized in figure 5.3.

5.2.2 Data pre-processing

From each track, defined as the consecutive bounding box annotations belonging to the same cow, segments of 32 frames are sampled with a stride of 2. At 30 FPS, each segment thus spans roughly 2.1 seconds of video footage. The bounding box annotations vary in size and aspect ratio across a track which leads to inconsistent resolutions between the frames in a segment. Most video classification models except some vision transformers require consistent resolution across frames. To satisfy this requirement each frame is scaled such

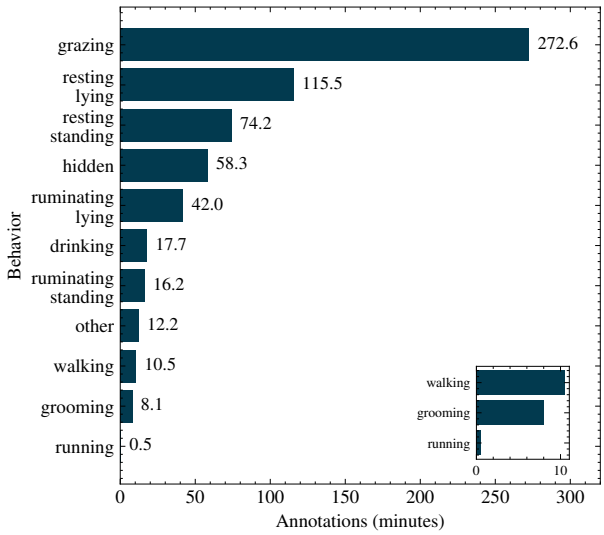


FIGURE 5.3: The class distribution of the CVB dataset reflects the natural imbalance in different behaviors exhibited by cattle.

that the aspect ratio of the original bounding box is maintained while any remaining area is zero-padded.

To be compatible with most video classification models, the resulting video segments are converted to a tensor format ($\mathbb{R}^{C\times T\times H\times W}$) with C channels, T frames, and some height H and width W. The dataset is split into a training- and validation set of roughly 80% and 20% respectively as provided by the research by Zia et al. [40]. This division remains unchanged throughout all experiments.

5.3 Active learning protocol

The following protocol describes a general set of instructions executed identically across all active learning experiments. Three active learning strategies, namely Max-Entropy [42], Cluster-Margin [31] and Core-Set [12] are used to draw samples from the unlabeled pool iteratively. The active learning strategies are then compared against a random sample.

Each model is fine-tuned on an initially chosen sample of varying size after which active learning is applied to select label candidates. The samples selected by the query strategy are added to the labeled set after which the model is reset to its initial state and re-trained on the updated labeled set. The re-initialization step is performed to avoid unwanted side effects such as catastrophic forgetting or any other influence from previous active learning iterations [71]. The budget of each query is kept consistent at 3% of the dataset throughout all experiments.

5.4 Evaluation metrics

In each experiment, model performance is evaluated on a held-out validation set using micro- and macro-average F1 scores. R_{total} and P_{total} refer to the summed recall and precision across all classes and M denotes the set of class labels. Macro averages were chosen due to the nature of the dataset, as reporting global accuracy measures can give a biased representation of model performance when applied to highly imbalanced datasets.

$$F1_{micro} = 2 \cdot \frac{P_{total} \cdot R_{total}}{P_{total} + R_{total}} \tag{5.5}$$

$$F1_{macro} = \frac{1}{|M|} \sum_{i \in M} 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i} \tag{5.6}$$

Chapter 6

Experimental results

6.1 Experimental outline

According to the framework and protocols established in chapter 5, a series of experiments were conducted to provide answers to each of the research questions. Additional information about the practical implementation of each experiment can be found in Appendix B.

6.1.1 Experiment: the influence of the size and selection of the initial training set on active learning outcomes

Initially, the effect of the choice and size of the initial training set on different active learning strategies is assessed in isolation to establish a baseline for comparison. The initial training set is determined by randomly sampling various percentages of the available data. For each initial training size, multiple random samples are taken and compared.

The initial training sizes were 5, 10, 15, 20, and 25 percent respectively. Starting from a model trained on these initial training sets, the active learning protocol, as described in section 5.3.1, was followed in each experiment. The query budget for selecting label candidates was kept consistent at 3% throughout the experiment. Further details regarding the implementation, and choice of hyper-parameters are addressed in appendix B.

We assess the influence of the initial training set across three different dimensions, by analyzing its effect on: *model performance*, *model uncertainty* and *label candidate selection* respectively.

6.1.2 Experiment: the influence of pre-training on active learning outcomes

To assess the effect that contrastive learning has on active learning outcomes, the model is pre-trained on the *instance discrimination* task through self-supervised contrastive learning. The MoCo framework [72] with key/query model optimization and queue mechanism is used to pre-train the model on the instance discrimination task. In this setup, a non-linear projection head is appended to the model's backbone to map the features to an embedding space where the contrastive loss is applied.

The model that performs best during a linear evaluation on the behavior classification task is selected for further evaluation (figure A.3b). The execution of this experiment is identical to the previous experiment and all models are analyzed along the same three dimensions.

6.1.3 Experiment: informed selection of the initial training set

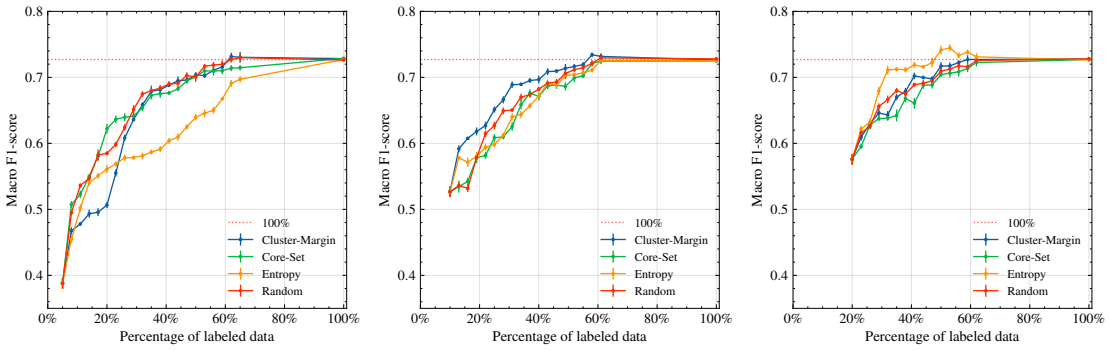
Lastly, the underlying structure of the feature representations derived from self-supervised pre-training is used to draw a representative sample using the unsupervised selection method described in chapter 5. The embedding space is divided into 100 clusters using the k-Means algorithm, after which they are sorted by increasing size. Starting from the smallest cluster, prototypes with the highest typicality score (as defined in chapter 5) are iteratively selected from each cluster. The number of neighbors to be included in the calculation of the typicality score was set to 20 and clusters with less than 5 members were ignored during selection.

This cycle is repeated, starting from the smallest unsaturated cluster until the total budget is exhausted. The proposed initial sampling method is compared to random sampling in terms of class distribution of the initial seed as well as classification performance.

6.2 Experiment: the influence of the size and selection of the initial training set on active learning outcomes

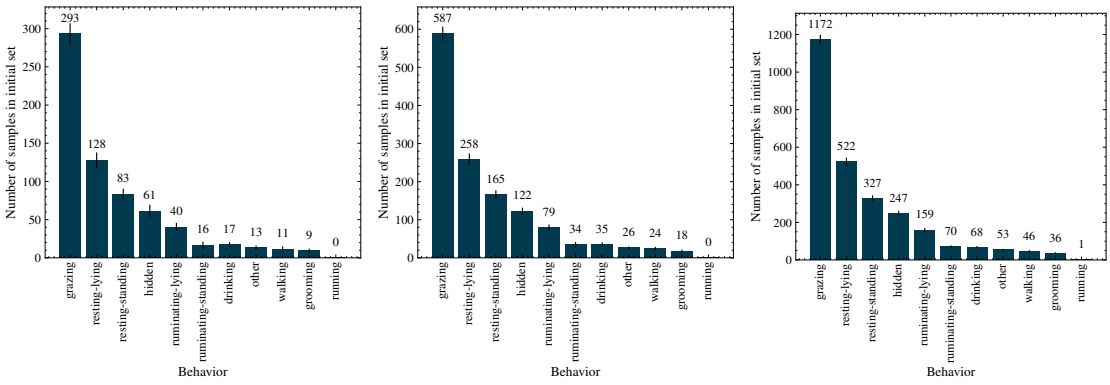
6.2.1 Analysis of model classification performance

Figure 6.1 depicts global macro F1-scores for the evaluated active learning strategies starting from three different sizes and selections of the initial training set. In addition, class-wise F1-scores and averages are reported in table 6.1. Notably, the overall performance on the validation set varies significantly across different initializations. Findings are discussed for each initial budget.



(A) 5% (B) 10% (C) 20%

FIGURE 6.1: Comparison of global macro F1-scores of different active learning strategies when varying the size of the initial training set.



(A) 5% (B) 10% (C) 20%

FIGURE 6.2: Average class-wise distribution for initial training sets of varying budgets

	% of labels used for the initial training set				
Behavior	5%	10%	15%	20%	25%
Grazing	0.88	0.91	0.92	0.92	0.92
Resting (lying)	0.66	0.78	0.79	0.81	0.80
Resting (standing)	0.70	0.73	0.77	0.79	0.80
Hidden	0.30	0.48	0.57	0.56	0.56
Ruminating (lying)	0.44	0.68	0.64	0.70	0.68
Drinking	0.37	0.74	0.72	0.77	0.75
Ruminating (standing)	0.44	0.42	0.48	0.59	0.72
Other	0.10	0.13	0.18	0.20	0.21
Walking	0.33	0.45	0.37	0.51	0.50
Grooming	0.11	0.13	0.23	0.52	0.55
Running	0.00	0.00	0.00	0.00	0.50
Global average (micro)	0.70	0.79	0.79	0.81	0.81
Global average (macro)	0.39	0.53	0.52	0.58	0.63

TABLE 6.1: Comparison of average F1-scores per class for various data budgets when fine-tuning end-to-end on the behavior classification task

Small initial budget (5%)

None of the active learning strategies can outperform random sampling consistently when starting from this initialization. It can be observed that the diversity-based method (CoreSet) achieves competitive performance to random sampling. Cold-start effects occur when using uncertainty-based methods (Max-entropy and Cluster Margin), as both of these strategies perform consistently worse on the validation set compared to random sampling until a large portion of the entire data budget has been exhausted.

Minority classes such as walking, drinking, and grooming are poorly recognized. Observing from figure 6.2a, these classes are relatively underrepresented in the initial training set and often confused with one of the majority classes (figure 6.3). It is evident that, given this initialization, the model shows a consistent bias towards grazing and resting behaviors.

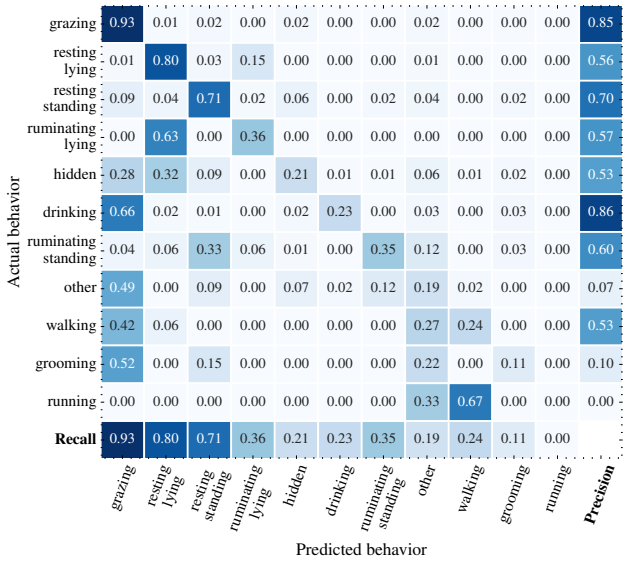


FIGURE 6.3: Confusion matrix at 5% initialization, normalized over the true behavior labels.

Medium initial budgets $(10 - 15\%)$

Given a slightly larger initialization set, a clear difference in the effectivity of different active learning strategies is observed compared to a smaller initial budget. The Cluster-Margin strategy consistently outperforms all other strategies throughout iterations. In addition, the entropy-based strategy, while still inferior to random sampling, exhibits less pronounced detrimental effects on model generalization.

Observing figure 6.2b, minority samples are slightly more represented in the initial training set and the error rate and its spread have been reduced compared to the previous initialization. Not all classes have improved from the addition of data. Resting and rumination in the standing pose are commonly confused (figure 6.4).

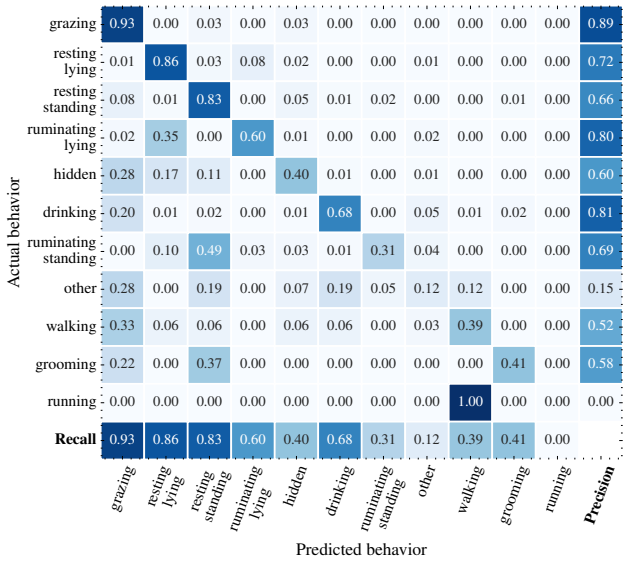


FIGURE 6.4: Confusion matrix at 10% initialization, normalized over the true behavior labels.

Large initial budgets $(20 - 25\%)$

In (relatively) larger initial budgets, the effectivity of uncertainty-based methods over other strategies is apparent. A relative improvement over the 100% baseline is observed for the Max-Entropy strategy (figure 6.1b), indicating that, with sufficient initialization, this strategy can surpass generalization performance compared to when all data is used. However, as the data budget is exhausted further, this relative improvement diminishes.

In observing the confusion in model predictions, as seen in figure 6.5), bias towards majority predictions has been reduced to a large extent. Although many of the minority behaviors in standing pose are still misclassified.

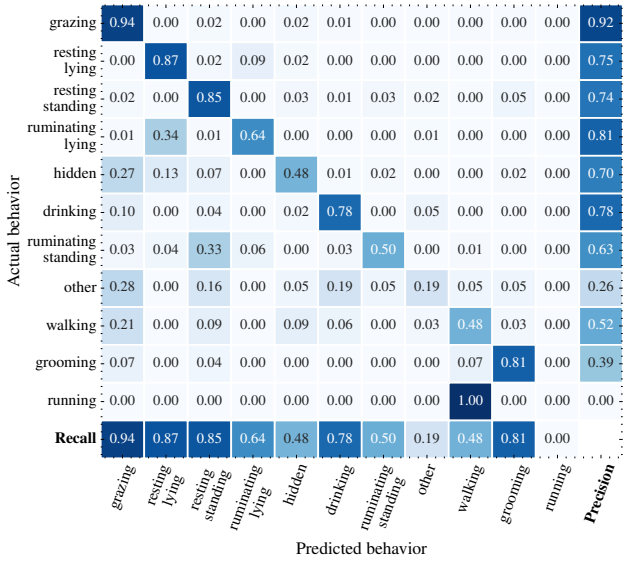


FIGURE 6.5: Confusion matrix at 20% initialization, normalized over the true behavior labels.

6.2.2 Analysis of model uncertainty

A viable stage to initiate uncertainty sampling can be hypothesized as follows: the majority of correct predictions should exhibit very low entropy, indicating high confidence, while the majority of incorrect predictions should have a relatively high entropy in their probability distributions. Furthermore, the distribution of entropy should have stabilized to some extent, such that the addition of a few samples does not result in drastic changes in distribution. Lower-entropy, correct predictions are a prerequisite to minimize redundancy in sample selection, while incorrect predictions with high entropy are necessary for the selection and correction of ambiguous, difficult cases that the model can improve upon.

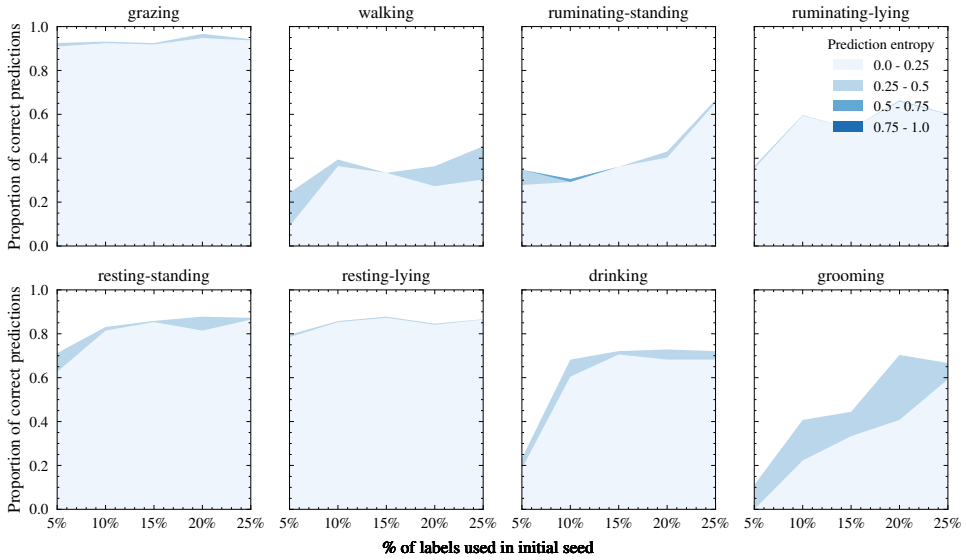


FIGURE 6.6: Initial proportions of correct predictions and their corresponding entropy when observed across different initial training sizes

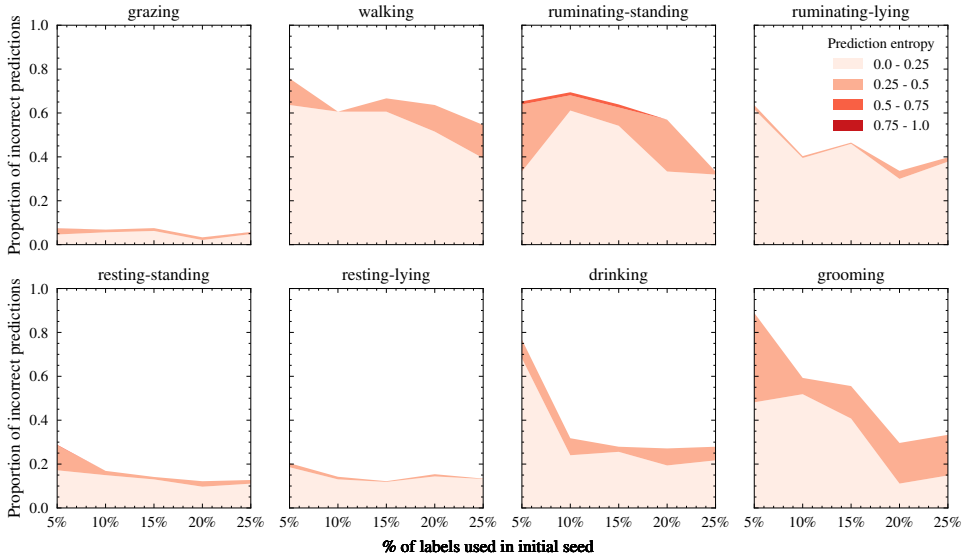


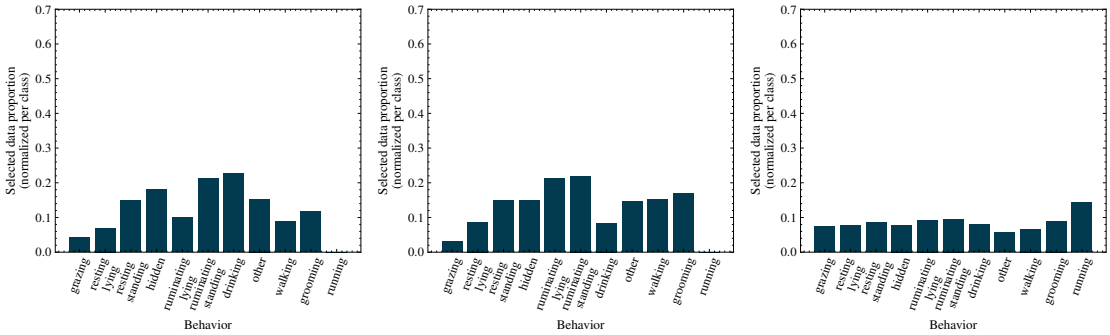
FIGURE 6.7: Initial proportions of incorrect predictions and their corresponding entropy when observed across different initial training sizes

To investigate the dynamics in model uncertainty, the proportions of correctly and incorrectly predicted samples, and their prediction entropy have been visualized in figure 6.6-6.7 across different initial training sets.

The resulting figures indicate that common behaviors, such as grazing and resting are well recognized with relatively low uncertainty, regardless of the initial budget size. On the other hand, minority classes (e.g. drinking, ruminating, grooming, walking), show a relatively large proportion of incorrect predictions with low entropy in the smaller initial data budgets. This signifies that the model is not only predicting these samples poorly at these stages but is highly confident in its incorrect predictions. When starting active learning based on maximizing uncertainty, these minority classes are thus mostly ignored

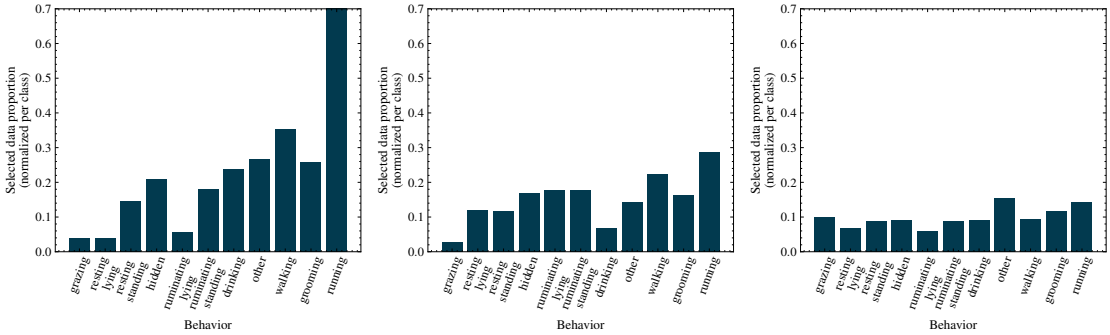
in initial iterations, potentially causing long-term detrimental effects to the recognition of such minority classes.

6.2.3 Analysis of label candidate selection



(A) Max-Entropy (B) Cluster-Margin (C) Core-Set

FIGURE 6.8: Class distribution of selected label candidates in the first 3 active learning iterations (5% initialization)



(A) Max-Entropy (B) Cluster-Margin (C) Core-Set

FIGURE 6.9: Class distribution of selected label candidates in the first 3 active learning iterations (20% initialization)

As shown in figures 6.2a, minority classes are under-represented and poorly recognized when the initial labeled budget is small. Previous analysis on model uncertainty shows that these minority classes are misclassified with high confidence, which could indicate that decision boundaries have not sufficiently developed.

This is reflected in the class distribution of selected samples across iterations, where minority classes are prioritized less in the early stages (figure 6.8). Consequently, uncertainty-based methods fail to identify sufficient minority samples to improve over random or diversity sampling, leading to cold-start effects. With larger initial budgets, minority classes become better represented and decision boundaries more developed, enabling uncertainty-based methods to more effectively identify them for labeling as reflected in figure 6.9.

The diversity-based strategy (Core-Set) is least influenced by the size of the initial training set, as reflected in less pronounced differences between the selected behaviors across different initializations. This finding is in line when observing the performance across active learning iterations, where minor differences are observed (figure 6.1b).

6.2.4 Conclusions

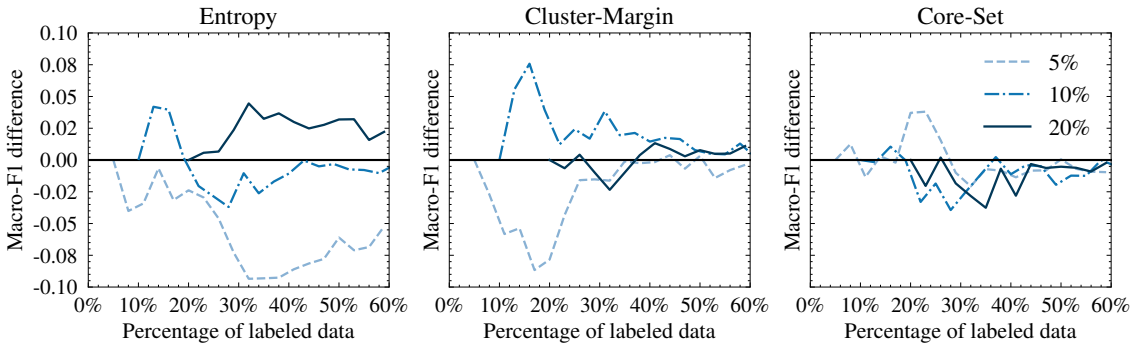


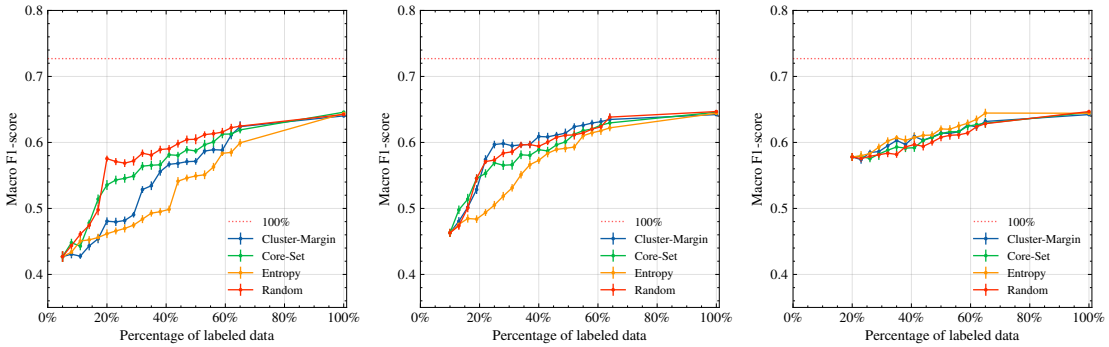
FIGURE 6.10: The impact of varying the initial training set size on the global macro F1-score of various active learning strategies, shown as the difference compared to the randomly sampled baseline.

Figure 6.10 depicts a type of phase transition in which the performance of active learning strategies over the randomly sampled baseline shifts as the size of the initial training set increases. The size of the initial set appears to have a significant impact on the effectivity of the uncertainty-based methods in particular, in which this phase shift is most apparent. In these active learning strategies, the cold-start phenomenon arises when starting from an initial seed of 5% as active learning strategies, especially those based on model uncertainty, perform consistently worse on the validation set compared to random sampling. An analysis of class distribution and model confusion, combined with the model's classification performance (Table 6.1), suggests that the imbalanced nature of the dataset introduces a bias toward majority classes due to insufficient representation in the initial sample.

This bias is evident when observing the distribution of entropy in low-budget initializations. The model produces overly confident predictions in favor of these majority classes, making model uncertainty an unreliable metric for sample selection.

In addition to initial seed size, the *make-up* or distribution of the initial seed appears to have a long-term effect on the effectivity of the active learning strategy. This is reflected by differences in model performance within the same active learning iteration. These differences can be sufficiently large in some cases that the *best* active learning strategy is not consistent across different samples of the same initial starting budget.

Observing either macro or micro average scores alone might give a biased view of the performance increases. Micro F1-score only increases by 15.7% after increasing the training data 5 times, whereas the macro average increases by 74.3%. Furthermore, even under a very low budget of 5%, we already reach 82.6% of the total potential realized when training with 100 percent of the available labels. For this reason, a supplementary analysis based on micro F1-score has been provided in chapter 7.



(A) 5% (B) 10% (C) 20%

FIGURE 6.11: Comparison of global macro F1-scores of different active learning strategies when applied to models that received self-supervised pre-training.

6.3 Experiment: the influence of pre-training on active learning outcomes

6.3.1 Analysis of model classification performance

Behavior	% of labels used for the initial training set					
	5%		10%		20%	
	Baseline	*CL*	*Baseline*	*CL*	*Baseline*	*CL*
Grazing	**0.88**	0.87	**0.91**	0.87	**0.92**	0.89
Resting (lying)	0.66	**0.73**	**0.78**	0.77	**0.81**	0.78
Resting (standing)	0.70	**0.72**	0.73	**0.73**	**0.79**	0.77
Hidden	0.30	**0.41**	**0.48**	0.45	**0.56**	0.45
Ruminating (lying)	0.44	**0.60**	0.68	**0.69**	0.70	**0.70**
Drinking	0.37	**0.63**	**0.74**	0.64	**0.77**	0.64
Ruminating (standing)	**0.44**	0.42	0.42	**0.61**	0.59	**0.77**
Other	0.10	**0.11**	**0.13**	0.10	**0.20**	0.16
Walking	**0.33**	0.05	**0.45**	0.07	**0.51**	0.17
Grooming	0.11	**0.23**	**0.13**	0.11	**0.52**	0.16
Running	0.00	0.00	0.00	0.00	0.00	**1.00**
Global average (micro)	0.70	**0.73**	**0.79**	0.77	**0.81**	0.77
Global average (macro)	0.39	**0.43**	**0.51**	0.46	**0.58**	0.57

TABLE 6.2: Comparison of average F1-scores per class for various data budgets when fine-tuning pre-trained models on the behavior classification task

Table 6.2 provides a detailed overview of class-wise F1-score with- and without applying self-supervised contrastive learning as a pre-train step. In addition, figure 6.11 depicts active learning outcomes for the pre-trained models. Results are discussed for each initial budget.

Small initial budget (5%)

The observed scores indicate that the pre-trained models surpass the baseline model in recognizing most classes during initialization, suggesting better initial adaptation to the behavior classification task. However, pre-trained models show a stronger bias towards the majority class *grazing*, with a larger proportion of minority classes being classified incorrectly.

In addition, the initial improvements are short-lived when observing model performance across active learning iterations. The relative improvement in the first 5 active learning iterations is significantly smaller compared to the baseline model, with only a 32.6% increase in macro-average F1-score. When using all available labeled data, macro-F1 score plateaus at 0.64 versus 0.73 were observed in baseline models, indicating a relatively poor recognition of minority classes.

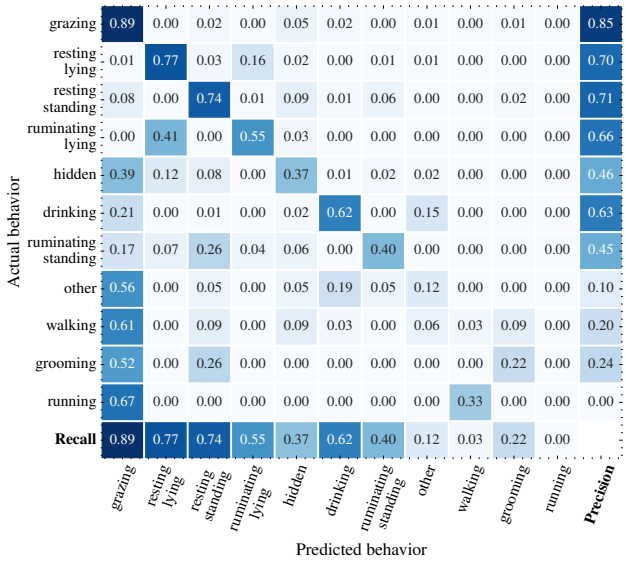


FIGURE 6.12: (Pre-trained) confusion matrix at 5% initialization normalized over the true behavior labels.

Medium initial budgets ($10 - 15\%$)

For the majority classes, both models perform comparably across medium budget sizes with marginal differences in F1 scores. However, models that were pre-trained with self-supervision show poor performance in recognizing minority classes and a larger discrepancy between micro and macro average F1 scores. Similar to previous findings, a stronger bias towards majority classes is observed compared to baseline models 6.13. However, pre-trained models are able to discriminate between rumination and resting in similar postures (standing- or lying) more often compared to baseline models, indicating a better separation of both classes in the representation space.

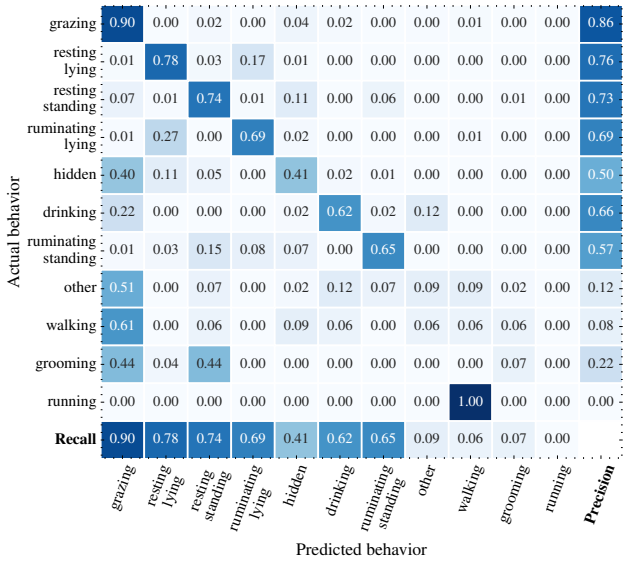


FIGURE 6.13: (Pre-trained) confusion matrix at 10% initialization normalized over the true behavior labels

Large initial budgets $(20 - 25\%)$

Similar observations as in previous budgets are made for larger budgets. Global macro-F1 scores are similar, yet give a biased view of model performance due to the recognition of the running class. Overall recognition of minority behaviors is significantly worse in comparison to baseline models. Figure **??** shows the correct and incorrect classification rates for 20% initializations respectively. It can be observed that the models that received self-supervision are more biased towards predicting the largest majority class (grazing), even for larger initial budgets. The stagnation of improvement in the recognition of minority classes indicates that pre-trained models do not adapt well to additional data compared to baseline models.

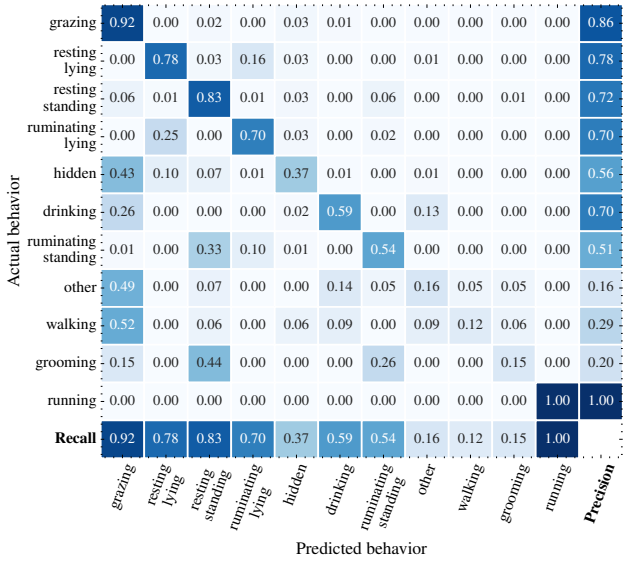


FIGURE 6.14: (Pre-trained) confusion matrix at 20% initialization normalized over the true behavior labels

6.3.2 Analysis of model uncertainty

A striking difference in model uncertainty between models that received self-supervised pre-training and models without pre-training is a larger proportion of high uncertainty in predictions. This uncertainty persists firmly across an increasing initial budget, whereas in baseline models, the proportion of highly uncertain samples often decreases as more labels become available.

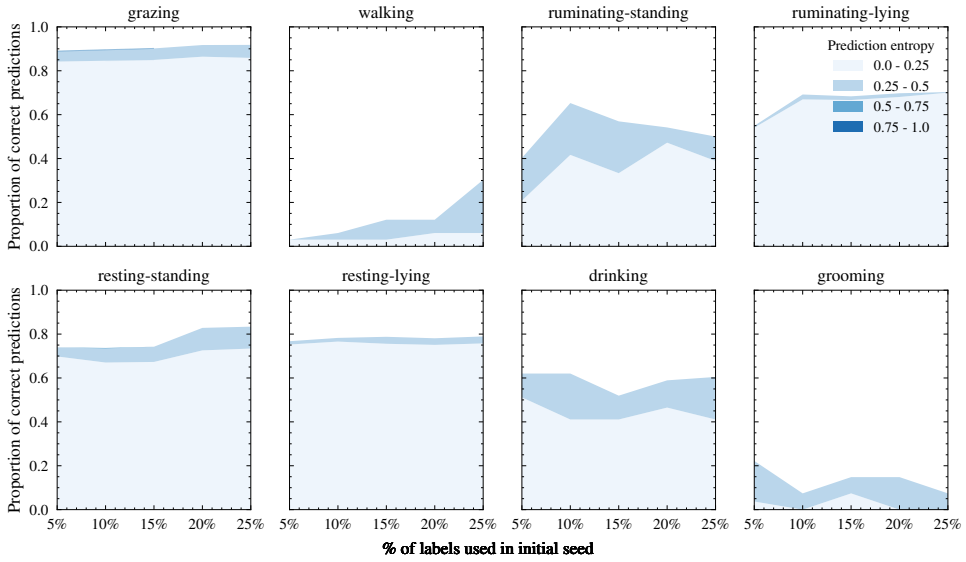


FIGURE 6.15: Initial proportions of correct predictions and their corresponding entropy when observed across different initial training sizes

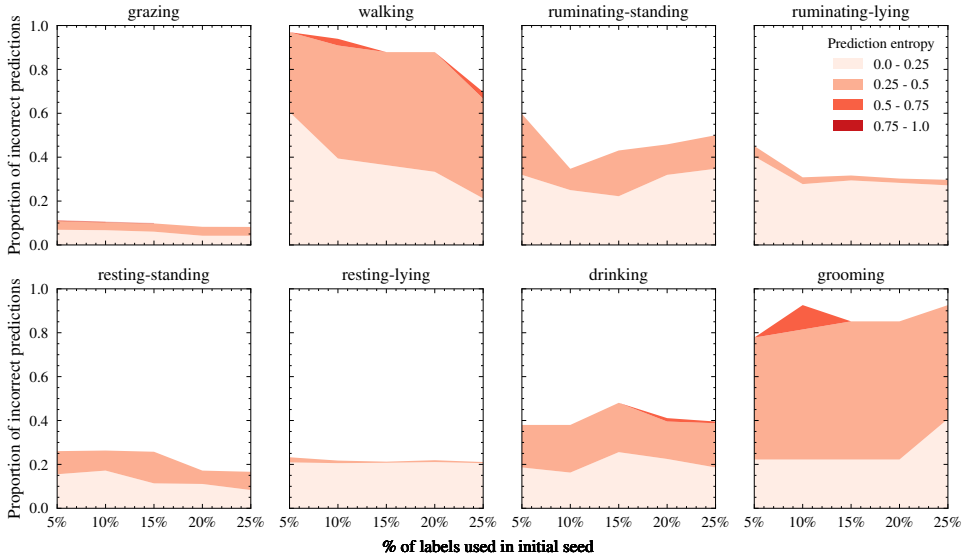


FIGURE 6.16: Initial proportions of incorrect predictions and their corresponding entropy when observed across different initial training sizes

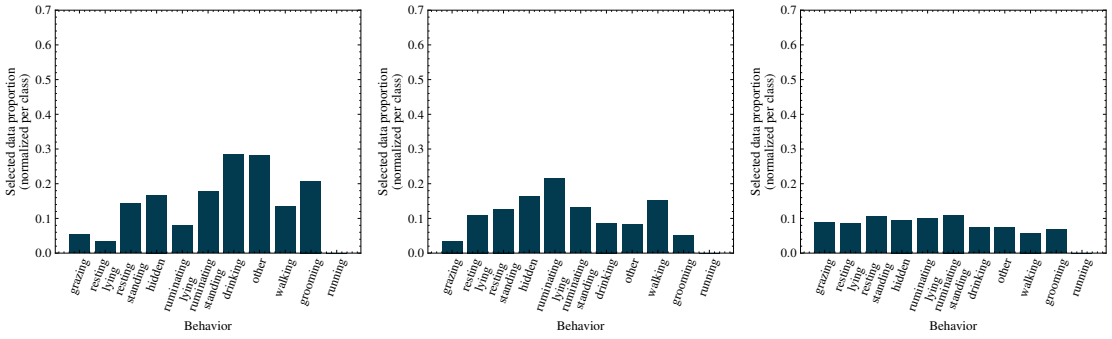
While the relative proportion of incorrect predictions decreases for most classes as more data becomes available, the persistent high uncertainty implies that the pre-train model struggles to learn from additional data effectively. The persisting bias towards the majority

class, as well as the persisting proportion of high-uncertainty samples, could suggest that pre-trained models are under-fitting to the dataset.

6.3.3 Analysis of label candidate selection

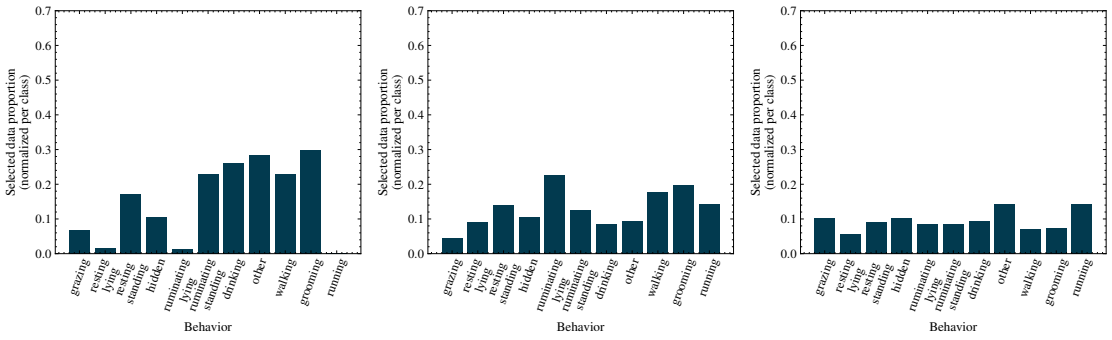
Figure 6.17 and 6.18 depict the label candidates selected by each active learning strategy during the first 3 iterations. Similarly to previous observations (figure 6.8 and 6.9), minority classes are prioritized less when starting from a small initial budget.

One interesting observation that differs from the baseline models is that even in later training stages, all active learning strategies spend a considerable percentage of the total budget on grazing samples. For example, in the first iteration, max-entropy sampling assigns 33% of the budget on grazing samples, which was only 18% on average for models that did not receive pre-training. Likewise, Core-Set spends 54% (previously 45%) towards this class and Cluster-Margin selects 21% versus the previously observed 12%. This suggests a degree of redundancy in sampling, hindering improvement across iterations.



(A) Max-Entropy (B) Cluster-Margin (C) Core-Set

FIGURE 6.17: Class distribution of selected label candidates in the first 3 active learning iterations when applied to pre-trained models (5% initialization)



(A) Max-Entropy (B) Cluster-Margin (C) Core-Set

FIGURE 6.18: Class distribution of selected label candidates in the first 3 active learning iterations when applied to pre-trained models (20% initialization)

6.3.4 Conclusions

Figure 6.19 depicts a similar phase transition as observed in baseline models (figure 6.10) with less pronounced differences between strategies in larger initial budges. No significant

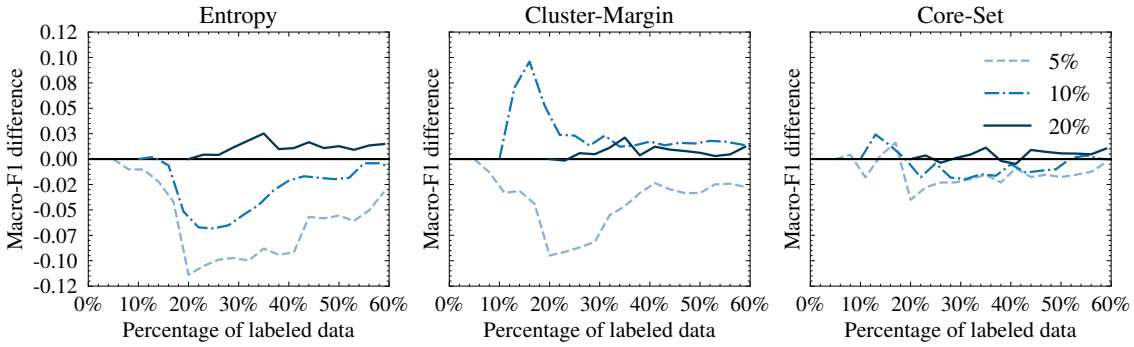


FIGURE 6.19: The impact of varying the initial training set size on the global macro F1-score of various active learning strategies when applied to pre-trained models, shown as the difference compared to the randomly sampled baseline.

differences between the effectivity of different sampling strategies were observed after pre-training, indicating that pre-training did not reduce cold-start effects.

The analysis highlights that self-supervised contrastive pre-training initially benefits recognition of most behavior during initialization when the label budget is small (5%), particularly for the majority classes. However, pre-training results in inferior performance as more data becomes available. In particular, recognition of minority behaviors degrades significantly compared to the baseline models.

In active learning outcomes (figure 6.1c), long-term performance plateaus below the baseline established by previous models that did not receive pre-training. An analysis of the confusion rates as well as the uncertainty in predictions suggests a strong and persisting bias towards majority classes. Such biases are reflected in the selected sample candidates of each active learning strategy, which spend a considerably larger proportion of its budget on majority classes. The early plateau, combined with poor recognition of minority behaviors and high uncertainty in predictions suggests that pre-training causes the models to consistently underfit to the data, a complementary analysis is given in chapter 7.

6.4 Experiment: informed selection of the initial training set

6.4.1 Analysis of model classification performance

Table 6.3 displays classification performance on the validation set when using the proposed initial training set selection method versus random selection. The proposed sampling method, which maximizes typicality across representation clusters, does not show consistent improvements over random sampling of the initial training set and is found to be detrimental in large budgets.

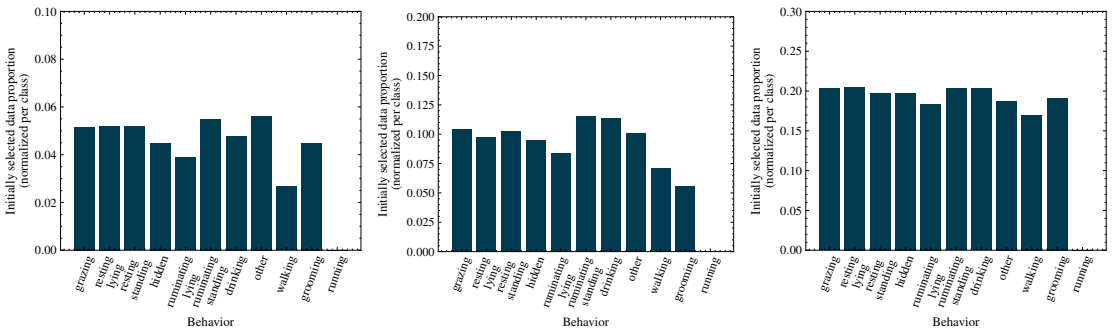
6.4.2 Analysis of behavior distribution of the initial training set

When observing the average distribution of behaviors selected by both random and informed sampling strategies, some notable differences are visible. In small budgets, a relatively larger proportion of *walking* samples are selected by the informed sampling strategy compared to random samples. However, the additional representation in the initial training set results in detrimental performance on class recognition. Meanwhile, despite sampling

similar amounts of the *drinking* samples, models trained on the dataset derived from informed sampling show considerably better recognition in this category.

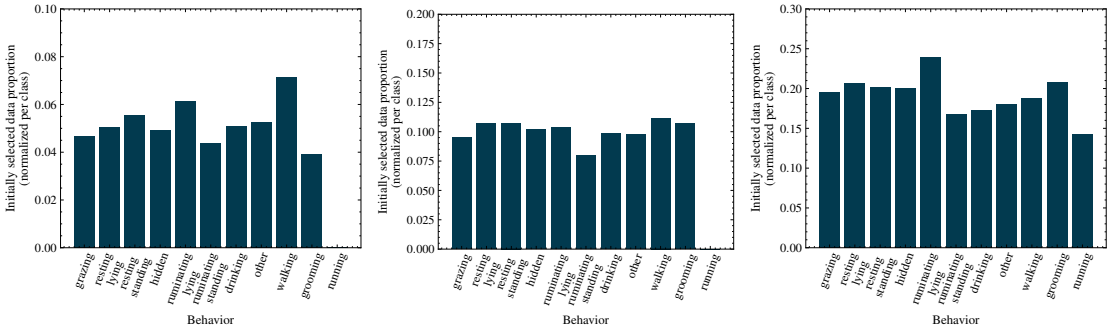
Behavior	**% of labels used for the initial training set**					
	5%		**10%**		**20%**	
	Random	*Typical*	*Random*	*Typical*	*Random*	*Typical*
Grazing	**0.88**	0.87	**0.91**	0.90	**0.92**	0.91
Resting (lying)	0.66	**0.72**	**0.78**	0.77	**0.81**	0.80
Resting (standing)	**0.70**	0.64	0.73	**0.76**	0.79	**0.82**
Hidden	0.30	**0.32**	**0.48**	0.48	**0.56**	0.54
Ruminating (lying)	0.44	**0.58**	**0.68**	0.65	0.70	**0.73**
Drinking	0.37	**0.53**	**0.74**	0.72	0.77	**0.78**
Ruminating (standing)	**0.44**	0.33	**0.42**	0.38	**0.59**	0.48
Other	**0.10**	0.07	0.13	**0.32**	0.20	**0.27**
Walking	**0.33**	0.16	**0.45**	0.26	**0.51**	0.35
Grooming	**0.11**	0.04	0.13	**0.49**	**0.52**	0.33
Running	0.00	0.00	0.00	0.00	0.00	0.00
Global average (micro)	0.70	**0.71**	**0.79**	0.78	**0.81**	0.80
Global average (macro)	0.39	**0.39**	**0.53**	0.52	**0.58**	0.55

TABLE 6.3: Comparison of average F1-scores per class for various data budgets when using random sampling versus informed sampling for initial training set selection.



(A) 5% (B) 10% (C) 20%

FIGURE 6.20: Average normalized class distribution of initial training set when using random sampling.



(A) 5% (B) 10% (C) 20%

FIGURE 6.21: Average normalized class distribution of initial training set when using the proposed typicality sampling.

6.4.3 Conclusions

The proposed informed sampling strategy failed to outperform random sampling for the initial training set consistently. For larger labeling budgets, it often underperformed, indicating that the representation space learned via self-supervised contrastive learning may not sufficiently separate behaviors and prototype-based selection offers little advantage over random sampling. However, only a single parameter configuration was evaluated, highlighting the need for further research on the effects of each parameter.

The effect of informed sampling over random sampling in active learning outcomes was not investigated in this research and would require further investigation. Despite similarities or differences in class distribution, it is difficult to deduce any claims about the relationship between class representation and the realized performance in behavior prediction.

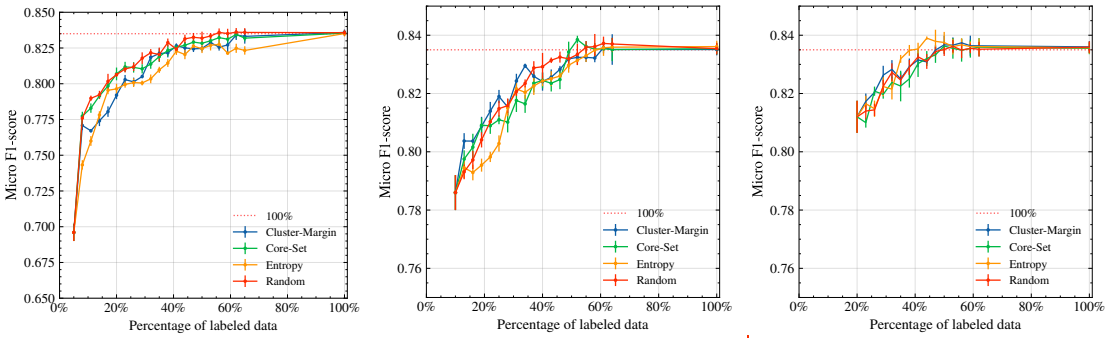
Chapter 7

Discussion

This section discusses complementary insights to the previously obtained results (section 6) through additional analyses.

7.1 Experiment: the influence of the size and selection of the initial training set

7.1.1 Analysis of micro-F1 score in active learning outcomes



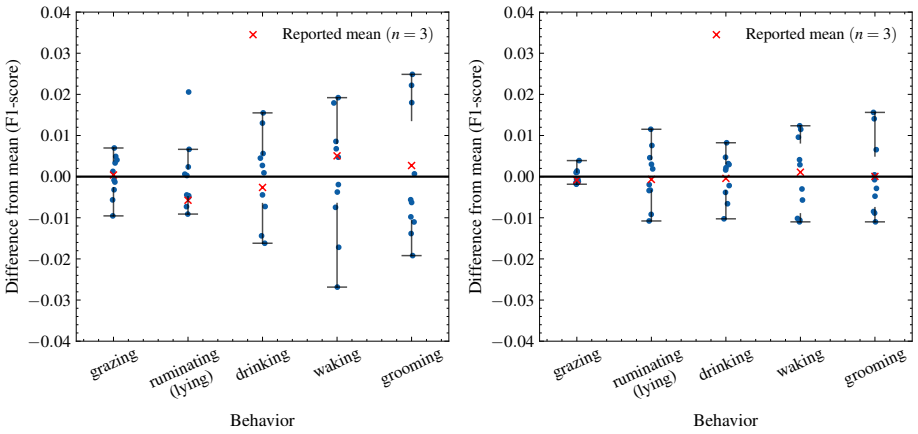
(A) 5% (B) 10% (C) 20%

FIGURE 7.1: Comparison of global micro F1-scores of different active learning strategies across varying initial budgets

Macro-averages can give a more complete view of model performance across all behaviors. Therefore, the macro F1-score was mainly used as an evaluation metric to compare different active learning strategies. However, this metric can give a biased view of performance in certain cases. For example, the macro F1-score may increase considerably while the total error rate across all samples increases.

Figure 7.1 depicts the micro-F1 score for active learning outcomes across different initializations. It can be observed that the overall improvement in the micro F1-score is much smaller than the macro F1-score. This is mainly caused by the majority of classes that are being recognized well across all budgets (table 6.1). Notably, the micro F1 score does follow similar patterns as observed for the macro F1 score in terms of active learning outcomes, albeit at a smaller scale. Improvements over random sampling are relatively small compared to the macro F1-score, which by itself can give an inflated sense of improvement.

7.1.2 Analysis of spread in class-wise performance during initialization



(A) 5% initialization (B) 20% initialization

FIGURE 7.2: The observed averages in class-wise performance change when increasing the number of initial samples ($n = 10$ versus $n = 3$).

The results presented in section 6 are based on only three random initializations per initial training set size. This is a limitation imposed by the high computational cost of training and evaluating active learning strategies. In active learning, where performance can vary based on the choice of the initial labeled set, this limited sampling risks introducing bias into the interpretation of the effectiveness of different strategies. Training the initial model requires only a fraction of the computational resources compared to evaluating multiple active learning iterations. Thus, we choose to analyze the class-wise model performance of the initial model using a greater number of random samples.

Figure 7.2 demonstrates this by comparing performance from ten different random initializations as opposed to three. The previously reported mean (denoted by the red cross) often deviates from the newly observed mean (indicated by the horizontal zero-crossing line), highlighting the influence of sampling variability. Notably, this discrepancy diminishes as the initial budget increases to 20%, suggesting that larger budgets could reduce the impact of random initialization variability. However, even with 10 random initializations, the statistical significance remains limited. The variability introduced by the small sample sizes used in this study should be considered when interpreting the reported results.

7.2 Experiment: the influence of pre-training

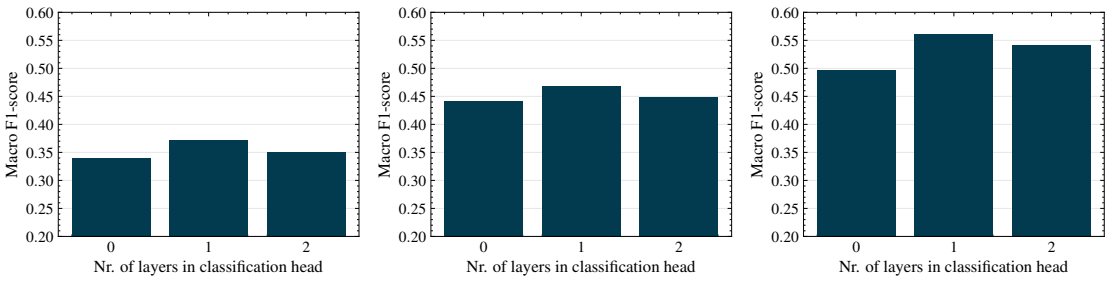
7.2.1 Linear evaluation versus end-to-end fine-tuning of pre-trained models

Table 7.1 summarizes the average model performance on held-out data when trained using different initial budget sizes with and without pre-training. Linear evaluation suggests that some adaptation to the target domain has occurred, as seen by an improvement in micro- and macro-average F1 scores. However, contrastive learning only provides an advantage when limited data is available. The results achieved during linear evaluation seem contradictory to those observed when fine-tuning the entire model. To investigate whether truly informative features have been learned, we vary the depth of the classification

head during evaluation by introducing one or more fully connected layers, followed by a non-linear activation function as suggested by Chen et al. [73]. The results given in figure 7.3 suggest that the features learned by the encoder are useful to the downstream task and can even achieve similar or improved performance compared to fine-tuning the entire model while requiring a fraction of the computation.

	Frozen	**Micro F1-score**		**Macro F1-score**	
% of labels		*Baseline*	*CL*	*Baseline*	*CL*
5%	✓	0.626	**0.634**	0.302	**0.340**
10%	✓	0.678	**0.684**	0.404	**0.428**
20%	✓	0.692	**0.707**	0.425	**0.441**
100%	✓	**0.734**	0.726	**0.510**	0.497
5%	✗	0.696	**0.728**	0.388	**0.427**
10%	✗	**0.786**	0.769	**0.526**	0.463
20%	✗	**0.812**	0.773	**0.577**	0.574
100%	✗	**0.835**	0.812	**0.727**	0.642

TABLE 7.1: Comparison of micro- and macro- F1-score after performing linear evaluation (✓) and end-to-end fine-tuning (✗) with various initial label budgets. *CL* refers to the models that received self-supervised pre-training while *Baseline* refers to the models that did not.



(A) 5% (B) 20% (C) 100%

FIGURE 7.3: Increasing the depth by introducing additional non-linear layers in the classification head improves performance on the validation set during linear evaluation.

7.2.2 Analysis of activation maps across varying initial budgets

Figure 7.4 illustrates the activation maps of the same frame, taken from a sample of each behavior class across different models that were trained using a different percentage of labels. Interestingly, as the data budget increases, the point of high activation appears to shift. For example, in drinking and grooming behaviors, higher-budget models appear to become more focused on the animal, suggesting a change in how models learn to detect these behaviors. Notably, some differences emerge between models that were pre-trained using self-supervised contrastive learning and those that received no such pre-training. This could signify that initial features learned through self-supervision can cause persisting differences in the features learned by the model as it is being fine-tuned to the downstream task. For minority classes such as drinking and grooming, the baseline model predominantly activates regions on the animal itself, whereas models leveraging contrastive pre-training often exhibit activations around the animal rather than directly on it. A more robust, quantitative analysis is required to substantiate these observations and determine their significance in the model's predictions.

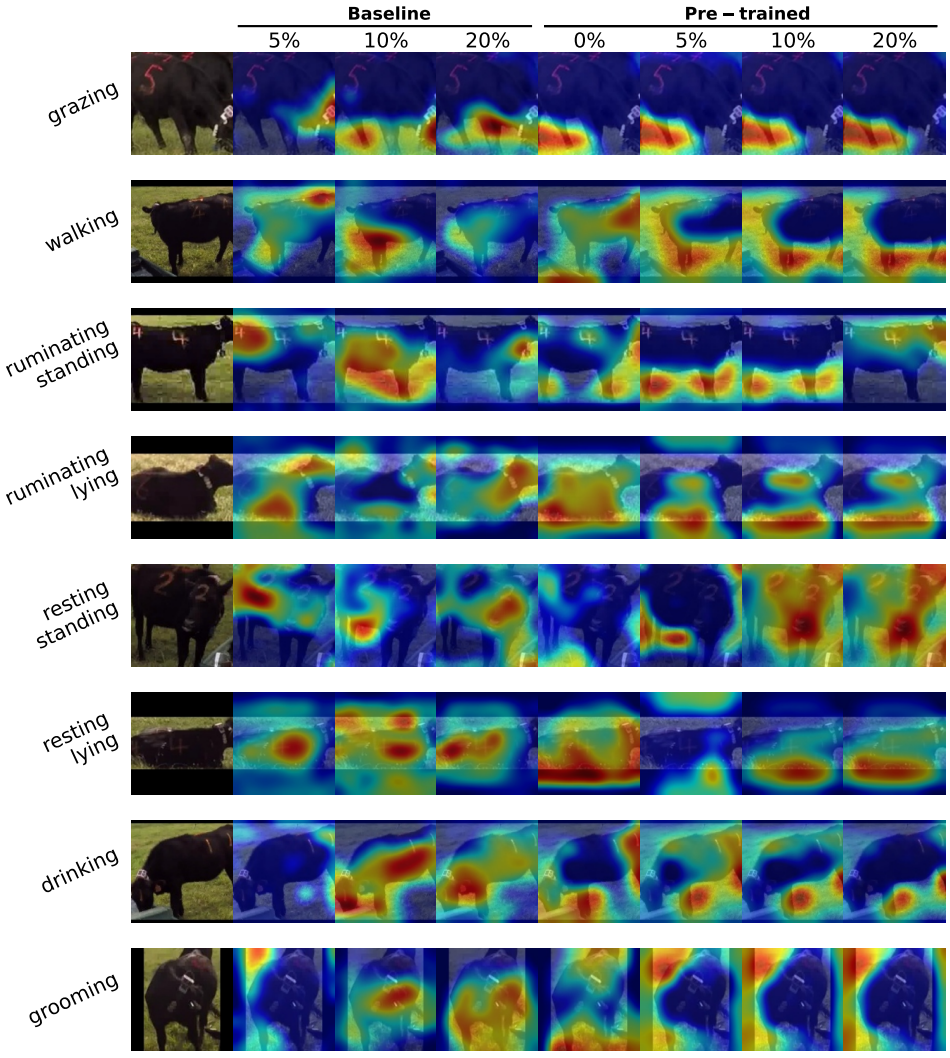
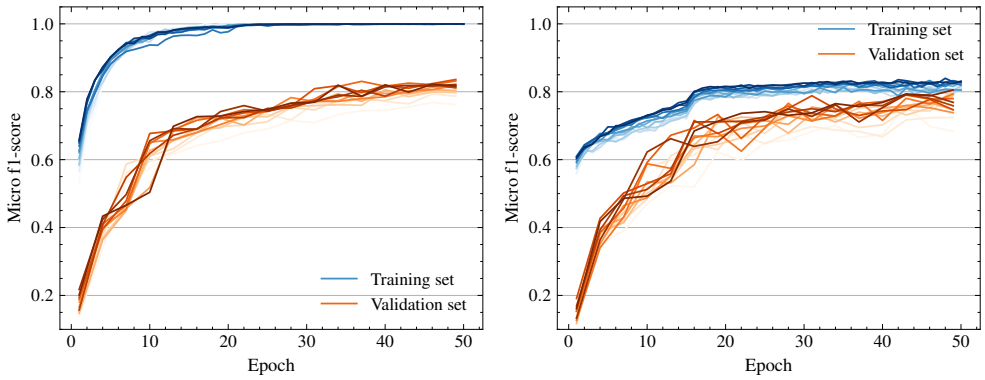


FIGURE 7.4: Comparison of activation maps across different model initializations and behavior classes

In addition, the focus of these activation maps uncovers some potential for data leakage in the dataset. For example, in observing the activations of the ruminating cow (third row), high activation is seen around the number drawn on the coat of the animal, as used for labeling purposes. Such numbers could be exploited given that each cow has a particular distribution of behavior within the dataset. However, in similar cases such as the grazing cow (first row), fixation on these numbers is not observed. Another form of potential data leakage is the water reservoir commonly visible in samples of behavior type *drinking* (seventh row. Strong activations are observed on and around such drinking reservoirs in certain models. This suggests that the visibility of such contextual cues is currently being exploited by the model to some extent.

However, visual inspection of the activation maps does not constitute a strong foundation of proof. A more elaborate analysis is required to determine the potential and degree of data leakage in this dataset.

7.2.3 Analysis of learning dynamics across active learning iterations for baseline and pre-trained models



(A) Baseline (B) Pre-trained

FIGURE 7.5: Learning dynamics of the training and validation set across active learning iterations (starting from an initial budget of 5%)

Figure 7.5 depicts the micro F1 scores on both training and validation sets throughout epochs and label budgets (5% - 60%). Learning dynamics for all active learning strategies can be found in Appendix A (figure A.1). Two observations can be made: (1) learning dynamics of the pre-trained models show a larger degree of instability, as seen in the fluctuation across epochs, and (2) pre-trained models show limited adaptation to the training data and as a result plateau much earlier. Based on these observations, two hypotheses are formalized to explain the observed difference in learning dynamics.

The case for under-fitting pre-trained models

As shown in figure 7.5b, the self-supervised models plateau earlier on both the training and validation data. The early plateau observed during training could suggest that the pre-trained models are unable to fully adapt to the training data, indicating a degree of consistent under-fitting during training. Such claims could be supported by the observation that pre-trained models are consistently more biased towards majority classes in comparison to baseline models (figures 6.12-6.14).

A potential explanation to this under-fitting could lie in the characteristics of the instance discrimination task learned during self-supervised pre-training. In this task, the model is driven to learn general representations that help identify similar, augmented samples. Representations that are helpful to perform this task might not align perfectly with the downstream task. Consequently, the learned representations may lack the fine-grained, task-specific features to discriminate between behaviors.

Addressing this issue may require incorporating mechanisms during pre-training to mitigate such discrepancies, such as class-aware contrastive losses, re-weighting strategies or better positive/negative mining strategies, to ensure that self-supervised representations are better aligned with the downstream classification task.

The case for over-fitting baseline models

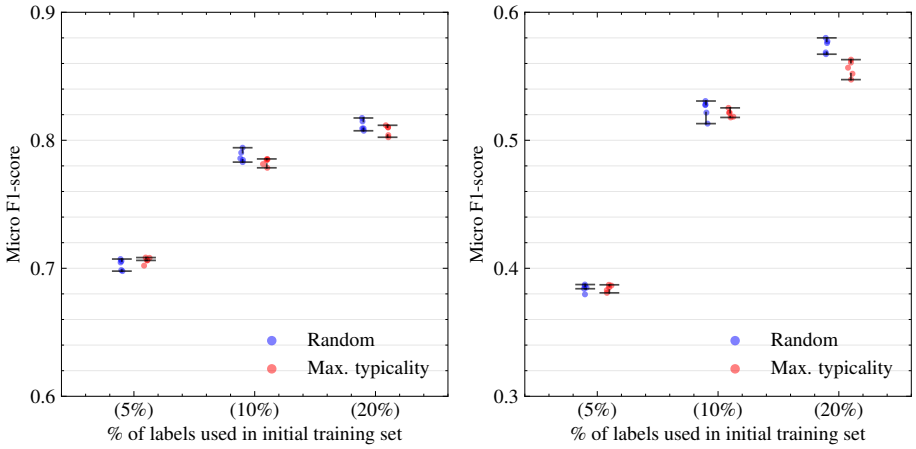
As shown in Figure 7.5a, baseline models can achieve near-perfect recognition of the training data, while the validation accuracy plateaus much earlier. This generalization gap suggests that the model stores information that does not contribute to recognizing unseen data but instead memorizes specific aspects of the training set. Rather than the self-supervised models under-fitting, the baseline models may be overfitting to specific patterns in the training data. If data leakage does occur, such memorization can lead to artificially inflated validation performance, where non-relevant features (e.g., contextual elements) drive predictions. As suggested in the previous analysis of activation maps, examples of these features might include coat numbers, objects, specific textures, or background cues that could correlate with certain behaviors.

In this context, the observed gap between the baseline and self-supervised models might reflect the baseline models' reliance on non-generalizable information. Consequently, the performance achieved by pre-trained self-supervised models, while potentially lower, may be more realistic and indicative of true generalization.

7.3 Experiment: informed selection

7.3.1 Analysis of the spread in performance metrics for informed versus random selections

Figure 7.6 displays the spread in classification performance on the validation set when using the proposed initial training set selection method versus random selection. There is insufficient evidence to suggest that informed sampling reduces variability in initial model performance compared to random sampling. The range of micro- and macro-average F1 scores across initialization is relatively similar to that observed in models where the initial training set was sampled randomly. It should be noted that, due to the small sample size ($n = 5$), results do not bear statistical significance.



(A) Micro F1-score (B) Macro F1-score

FIGURE 7.6: Comparison of model classification performance when sampling the initial training set randomly versus sampling according to the maximization of typicality as proposed in chapter 5

Chapter 8

Limitations and future work

The proposed framework was not able to solve the cold-start problem in active learning when applied to the behavior recognition task. In this section, we discuss some of the potential limitations of the research as well as the potential for future work.

8.1 Limitations of the dataset

The CVB dataset used in this study contains footage recorded exclusively at a single pasture. While this setup simplifies data collection and provides a controlled environment for experimentation, it introduces limitations that must be considered when interpreting the results. One of the key concerns is the potential for data leakage. Since all data originates from the same location and cows, the model might unintentionally learn environment- or animal-specific cues that do not generalize beyond this particular setting. Some of the findings discussed in previous sections suggest the potential for data leakage based on the coat pattern numbers.

In addition, the scale of the available data was limited, only spanning a total of 2 hours of footage. The lack of diversity in such a dataset may limit the ability to learn representative features for some rare behaviors of interest, for which data is quickly depleted in these experiments. This same lack of data could hinder the effective implementation of contrastive learning approaches on the target domain. Empirical research suggests that the approach used for contrastive learning is inherently data-hungry and was originally not designed for problems where large data imbalances occur.

Due to these limitations, the ability of such models to recognize the same behavior across different settings remains uncertain. Future research should address these limitations by collecting footage from multiple pastures or barns. Furthermore, data leakage due to contextual information such as the environment can be minimized through better segmentation of the animals. Well-aligned segmentation masks contain much less background information than bounding boxes used in this research.

Lastly, while not investigated in this work, others researchers have demonstrated the potential of other modalities such as depth information or optical flow to aid in behavior recognition. There is strong evidence in other research to suggest that drawing from multiple data modalities, even besides image- or video based information can drastically improve performance on a variety of different problems.

Although contextual cues, such as background information, have previously been mentioned as a potential risk for data leakage, they also offer a potential to enhance model performance when used appropriately. For instance, the proximity of an animal to a feed bunk can serve as a strong prior for predicting feeding behavior. Similarly, the state of

dormancy and the animal's position in cubicles can be valuable priors for identifying resting or ruminating behaviors. By carefully integrating these environmental cues, models can make more informed predictions.

8.2 Limitations of self-supervised contrastive learning approach

There is a notable instability in the learning dynamics when fine-tuning the model that was adapted to the target domain through contrastive learning (figure 7.5b). Linear evaluation shows that some adaptation to the target domain has occurred, (table 7.1). Yet, models that were pre-trained with a contrastive loss consistently under-fit the training data. The instability in the learning dynamics could be due to a few factors introduced during the contrastive domain adaptation.

Given that adaptation to the target domain has already taken place during the contrastive learning step, fine-tuning to the downstream task could require a much lower, tuned learning rate compared to previous experiments. However, a sensitivity analysis to confirm this hypothesis has not been conducted. Another hypothesis is that features that are learned during instance discrimination do not align well with the downstream task. One observation is that while the classification performance of the majority classes remains relatively stable (table 6.1), the adapted model is still not as capable of discriminating between similar minority classes as our baseline model. The induced bias towards these majority classes introduces an artificial, potentially misleading improvement in summary statistics, while recognition of minority classes remains relatively poor.

Additionally, the current contrastive learning approach uses instance discrimination as a supervisory signal for training. One concern, given the characteristics of the dataset and task, is that we risk over-clustering the representations and learning features that are not helpful to the downstream task. The model does not necessarily learn to push together similar behavior but only similar instances of that behavior. This is a side-effect of the negative sampling strategy which considers all samples outside of a track to be different, even when these two tracks might show the same behavior.

This is less of a concern in large and more diverse datasets such as Kinetics-400. There is more inherent diversity between different activity classes across relatively different environments. In our case, the environment remains fixed and diversity between activities can be minimal. For example, consider the difference between resting while lying versus ruminating while lying. The pose and background remain very similar. This difference lies mostly in a small repetitive movement of the head/jaw while ruminating.

The exact cause of the observed plateau and instability in learning dynamics has not been determined, which makes it difficult to draw conclusions about the effectiveness of the proposed method. Future research should investigate the underlying causes of instability in fine-tuning after contrastive domain adaptation. Another promising direction is to revisit the contrastive learning objective itself. Instead of instance discrimination, future work could explore task-aware or prototype-based contrastive learning strategies that consider the inherent structure of the problem. In addition, there is room for improvement in addressing the imbalance in class representations during contrastive learning. Future work could explore minority-class-aware negative sampling or use some additional prior knowledge to guide the negative and positive sampling selection process.

8.3 Limitations of the proposed selection method for the initial selection set

The approach used to select the initial training set by maximizing typicality across representation clusters relies heavily on the quality of the representations learned during the pre-training stage. Beyond this, its success is influenced by three hyper-parameters, each of which carries potential trade-offs:

- **Number of clusters:** Setting an inappropriate number of clusters can lead to over-clustering or under-clustering of the representation space. Over-clustering risks dividing the data into too many small groups, potentially leading to redundancy. On the other hand, under-clustering can lead to overly broad clusters, reducing diversity in the selected samples and missing finer distinctions between behaviors.
- **Number of neighbors included in the calculation of the typicality score:** The number of neighbors used to determine the typicality can largely influence sample selection. If too many neighbors are considered, density measures may be diluted, making it difficult to identify clear, local prototypes. However, selecting too few neighbors might put too much emphasis on local density variations, which could result in selecting samples that do not effectively represent the entire cluster.
- **Minimum number of samples per cluster:** The threshold for the minimal number of samples in a cluster can have a large impact on the selection process. If set too high, this parameter can exclude rare behaviors that are important in the behavior classification task. However, if the threshold is too low, it risks selecting outliers that are not representative of the broader dataset.

Future work should perform a more systematic analysis of these hyper-parameters and how they influence sample selection for a broader range of problems. Furthermore, methods to optimize or automate the selection of these parameters, based on data distribution or task-specific requirements could further improve the approach.

8.4 Limitations in the evaluation of active learning strategies

A limitation of this research is that the evaluation method used during active learning is computationally expensive and may not be as practical for commercial applications. Specifically, the approach of fully re-training the model at each iteration, while standard in research settings, adds a significant computational cost that reduces the overall cost-effectiveness of active learning.

Furthermore, the hyper-parameters, such as the learning rate and regularization parameters, were determined based on the learning dynamics observed during the initialization phase and remained fixed throughout the active learning process. This approach neglects the fact that as more labeled data is incrementally added, the model's learning dynamics may shift. Consequently, using fixed hyper-parameters may lead to sub-optimal convergence.

Future work could address these limitations by exploring more efficient evaluation strategies during active learning, such as fine-tuning the model from previous iterations instead of full re-training or applying continual learning strategies. Additionally, adaptive hyper-parameter optimization techniques, such as learning rate schedules or warm-ups could improve convergence and model performance across iterations. Implementing such

approaches could make the practical applicability and scalability of active learning more feasible.

Lastly, the active learning strategies evaluated in this research were not explicitly designed to handle scenarios with significant class imbalances, which can hinder their effectiveness in these contexts. These strategies may prioritize majority-class samples due to their higher prevalence in the data, leaving minority classes underrepresented in the labeled dataset.

Future work could focus on incorporating active learning strategies that explicitly address class imbalance. For example, some active learning strategies use weighted uncertainty metrics that give higher priority to uncertain samples from minority classes, ensuring a more balanced representation in the labeled dataset. Another promising approach is diversity-based sampling tailored to identify and prioritize underrepresented samples in the feature space, effectively boosting the representation of minority classes. Furthermore, incorporating domain knowledge about the problem could further mitigate these challenges. For example, prior knowledge about the expected behavior distribution or the relationships between classes could guide the sampling process, ensuring a more representative selection of samples.

8.5 Limitations of active learning as a learning paradigm

Active learning was not able to consistently offer a substantial improvement over regular supervised learning practices during this research. The effectiveness of each active learning strategy is determined by many factors that might not always be known beforehand such as a sufficient amount of pre-training, selection of the best strategy at different stages of model training, and the characteristics of the classification problem. However, active learning is only one branch of a broader set of techniques developed to make machine learning more efficient and sustainable. There has been a surge of interest in alternative approaches such as self-supervised, semi-supervised, and few-shot learning. These methods likewise show considerable potential for reducing the amount of costly annotations. There appears to be a consensus that the field of machine learning as a whole should move to more sustainable practices by leveraging unlabeled data more effectively, using a minimal amount of annotated examples while striving to reduce the computational complexity/footprint of training and maintaining deep learning models.

Chapter 9

Conclusion

This study investigated the cold-start problem in applying active learning to livestock behavior monitoring from video data. A framework for early-stage active learning was proposed and assessed for its impact on reducing cold-start effects. The posed research questions were addressed as follows:

Regarding the main research question, the findings indicate that the proposed framework did not effectively reduce cold-start effects in the behavior classification task. The success of active learning strategies remained heavily dependent on the size and make-up of the initial training set. Furthermore, different strategies demonstrated varying levels of effectiveness at different stages of the learning process.

The size and selection of the initial training set were found to have a large impact on the performance of active learning strategies. Cold-start effects were mostly observed when the initial budget was small, leading to an under-representation of minority classes and a resulting bias toward majority classes in both model predictions and label candidate selection. Additionally, uncertainty-based strategies were found to be poorly calibrated in low-budget scenarios, yielding suboptimal results. Larger initial budgets helped reduce cold-start effects, enabling uncertainty-based strategies to outperform random sampling.

Self-supervised pre-training did not yield improvements over baseline models and, in fact, negatively impacted active learning outcomes. Pre-trained models exhibited a persistent bias toward majority classes, poor recognition of minority classes, underfitting on the training set, and unstable learning curves. These issues remain poorly understood and require further investigation.

Regarding informed initial training set selection, no substantial advantage was observed over random sampling. Both methods performed similarly at lower initial budgets and showed a slight performance decline with larger budgets. While differences in behavior distributions were noted between the methods, no conclusive relationship was found between sample selection and model generalization.

The findings highlight that cold-start effects remain a significant challenge in imbalanced classification problems such as livestock behavior monitoring. Addressing this issue requires further research to develop more robust methods for mitigating cold-start effects. Potential directions include exploring alternative pre-training approaches to improve representation learning for minority classes and refining active learning strategies to handle imbalances more effectively.

Chapter 10

Disclaimer: use of AI tools

During the preparation of this work, generative AI tools such as ChatGPT were consulted to review and improve the clarity of the research paper. No research findings, data analysis, or original contributions were derived using such tools. The conclusions presented in this study are the sole product of the authors' independent work.

Bibliography

[1] A. Frost, C. Schofield, S. Beaulah, T. Mottram, J. Lines, and C. Wathes, "A review of livestock monitoring and the need for integrated systems," *Computers and electronics in agriculture*, vol. 17, no. 2, pp. 139–159, 1997.

[2] A. Beldman, J. Reijs, C. Daatselaar, and G. Doornewaard, *De Nederlandse melkveehouderij in 2030: verkenning van mogelijke ontwikkelingen op basis van economische modellering*. No. 2020-090 in Rapport / Wageningen Economic Research, Netherlands: Wageningen Economic Research, 2020. Project code 2282200565.

[3] D. Berckmans, "General introduction to precision livestock farming," *Animal Frontiers*, vol. 7, no. 1, p. 6 – 11, 2017. Cited by: 270.

[4] C. Chen, W. Zhu, and T. Norton, "Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning," *Computers and Electronics in Agriculture*, vol. 187, p. 106255, 2021.

[5] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos, "Beyond neural scaling laws: beating power law scaling via data pruning," 2023.

[6] M. Lorbach, R. Poppe, and R. C. Veltkamp, "Interactive rodent behavior annotation in video using active learning," *Multimedia Tools and Applications*, vol. 78, July 2019.

[7] J. Li, M. Keselman, and E. Shlizerman, "Openlabcluster: Active learning based clustering and classification of animal behaviors in videos based on automatically extracted kinematic body keypoints," 2023.

[8] J. Li, T. Le, and E. Shlizerman, "Al-sar: Active learning for skeleton-based action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–9, 2023.

[9] C. Harris, K. R. Finn, M.-L. Kieseler, M. M. R., and P. U. Tse, "Deepaction: a matlab toolbox for automated classification of animal behavior in video," *Scientific Reports*, vol. 13, Febuary 2023.

[10] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," 2022.

[11] Q. Jin, M. Yuan, S. Li, H. Wang, M. Wang, and Z. Song, "Cold-start active learning for image classification," *Information Sciences*, vol. 616, pp. 16–36, 2022.

[12] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2018.

[13] L. Chen, Y. Bai, S. Huang, Y. Lu, B. Wen, A. L. Yuille, and Z. Zhou, "Making your first choice: To address cold start problem in vision active learning," 2022.

[14] H. W. e. a. Kuo Li, Daoerji Fan, "A new dataset for video-based cow behavior recognition," 01 2024.

[15] A. Fuentes, S. Han, M. F. Nasir, J. Park, S. Yoon, and D. S. Park, "Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning," *Animals*, vol. 13, no. 12, 2023.

[16] Y. Guo, Y. Qiao, S. Sukkarieh, L. Chai, and D. He, "Bigru-attention based cow behavior classification using video data for precision livestock farming," *Transactions of the ASABE*, vol. 64, no. 6, pp. 1823–1833, 2021.

[17] D. Wu, M. Han, H. Song, L. Song, and Y. Duan, "Monitoring the respiratory behavior of multiple cows based on computer vision and deep learning," *Journal of Dairy Science*, vol. 106, no. 4, pp. 2963–2979, 2023.

[18] L. Huang, Y. Liu, B. Wang, P. Pan, Y. Xu, and R. Jin, "Self-supervised video representation learning by context and motion decoupling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13886–13895, 2021.

[19] M. Toering, I. Gatopoulos, M. Stol, and V. T. Hu, "Self-supervised video representation learning with cross-stream prototypical contrasting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 108–118, January 2022.

[20] P. M. Post, L. Hogerwerf, E. A. Bokkers, B. Baumann, P. Fischer, S. Rutledge-Jonker, H. Hilderink, A. Hollander, M. J. Hoogsteen, A. Liebman, M.-J. J. Mangen, H. J. Manuel, L. Mughini-Gras, R. van Poll, L. Posthuma, A. van Pul, M. Rutgers, H. Schmitt, J. van Steenbergen, H. A. Sterk, A. Verschoor, W. de Vries, R. G. Wallace, R. Wichink Kruit, E. Lebret, and I. J. de Boer, "Effects of dutch livestock production on human health and the environment," *Science of The Total Environment*, vol. 737, p. 139702, 2020.

[21] G. Doornewaard, A. Kok, C. Daatselaar, and A. Beldman, *Verkenning economische prestatie melkveehouderij in relatie tot duurzaamheidsdoelen: Gaat een betere score op biodiversiteit, klimaat en grondgebondenheid samen met een betere economische prestatie?* No. 2023-106 in White paper / Wageningen Economic Research, Netherlands: Wageningen Economic Research, Sept. 2023.

[22] H. M. T. Boer, R. F. Veerkamp, B. Beerda, and H. Woelders, "Estrous behavior in dairy cows: identification of underlying mechanisms and gene functions," *Animal*, vol. 4, no. 3, p. 446–453, 2010.

[23] L. Polsky and M. A. von Keyserlingk, "Invited review: Effects of heat stress on dairy cattle welfare," *Journal of Dairy Science*, vol. 100, no. 11, pp. 8645–8657, 2017.

[24] M. B. Sadiq, S. Z. Ramanoon, W. M. Shaik Mossadeq, R. Mansor, and S. S. Syed-Hussain, "Association between lameness and indicators of dairy cow welfare based on locomotion scoring, body and hock condition, leg hygiene and lying behavior," *Animals*, vol. 7, no. 11, 2017.

[25] L. González, B. Tolkamp, M. Coffey, A. Ferret, and I. Kyriazakis, "Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows," *Journal of Dairy Science*, vol. 91, no. 3, pp. 1017–1028, 2008.

[26] S. Paudyal, "Using rumination time to manage health and reproduction in dairy cattle: a review," *Veterinary Quarterly*, vol. 41, pp. 1–14, 10 2021.

[27] C. Petersson-Wolfe, K. Leslie, and T. Swartz, "An update on the effect of clinical mastitis on the welfare of dairy cows and potential therapies," *Veterinary Clinics of North America - Food Animal Practice*, vol. 34, pp. 525–535, 11 2018.

[28] C. B. Tucker, M. B. Jensen, A. M. de Passillé, L. Hänninen, and J. Rushen, "Invited review: Lying time and the welfare of dairy cows," *Journal of Dairy Science*, vol. 104, no. 1, pp. 20–46, 2021.

[29] M. B. M. Bracke and H. Hopster, "Assessing the importance of natural behavior for animal welfare," *Journal of Agricultural and Environmental Ethics*, vol. 19, pp. 77–89, Feb 2006.

[30] M. Wu, C. Li, and Z. Yao, "Deep active learning for computer vision tasks: Methodologies, applications, and challenges," *Applied Sciences*, vol. 12, no. 16, 2022.

[31] G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar, "Batch active learning at scale," 2021.

[32] D. Avola, M. Cascio, L. Cinque, G. Foresti, C. Massaroni, and E. Rodolà, "2d skeleton-based action recognition via two-branch stacked lstm-rnns," *IEEE Transactions on Multimedia*, vol. PP, pp. 1–1, 12 2019.

[33] B. Jiang, X. Yin, and H. Song, "Single-stream long-term optical flow convolution network for action recognition of lameness dairy cow," *Computers and Electronics in Agriculture*, vol. 175, p. 105536, 2020.

[34] T. Lodkaew, K. Pasupa, and C. K. Loo, "Cowxnet: An automated cow estrus detection system," *Expert Systems with Applications*, vol. 211, p. 118550, 2023.

[35] R. Wang, Z. Gao, Q. Li, C. Zhao, R. Gao, H. Zhang, S. Li, and L. Feng, "Detection method of cow estrus behavior in natural scenes based on improved yolov5," *Agriculture*, vol. 12, no. 9, 2022.

[36] A. Fuentes, S. Yoon, J. Park, and D. Park, "Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information," *Computers and Electronics in Agriculture*, vol. 177, p. 105627, 10 2020.

[37] Q. Bai, R. Gao, R. Wang, Q. Li, Q. Yu, C. Zhao, and S. Li, "X3dfast model for classifying dairy cow behaviors based on a two-pathway architecture," *Scientific Reports*, vol. 13, 11 2023.

[38] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," 2020.

[39] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," 2019.

[40] A. Zia, R. Sharma, R. Arablouei, G. Bishop-Hurley, J. McNally, N. Bagnall, V. Rolland, B. Kusy, L. Petersson, and A. Ingham, "Cvb: A video dataset of cattle visual behaviors," 2023.

[41] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *Machine Learning: ECML 2006* (J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, eds.), (Berlin, Heidelberg), pp. 413–424, Springer Berlin Heidelberg, 2006.

[42] Z. Liu, H. Ding, H. Zhong, W. Li, J. Dai, and C. He, "Influence selection for active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9274–9283, October 2021.

[43] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," 2017.

[44] J.-J. Zhu and J. Bento, "Generative adversarial active learning," 2017.

[45] T. Tran, T.-T. Do, I. Reid, and G. Carneiro, "Bayesian generative active deep learning," 2019.

[46] D. Yoo and I. S. Kweon, "Learning loss for active learning," 2019.

[47] Z. Liu, H. Ding, H. Zhong, W. Li, J. Dai, and C. He, "Influence selection for active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9274–9283, October 2021.

[48] C. Yin, B. Qian, S. Cao, X. Li, J. Wei, Q. Zheng, and I. Davidson, "Deep similarity-based batch mode active learning with exploration-exploitation," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 575–584, 2017.

[49] O. Yehuda, A. Dekel, G. Hacohen, and D. Weinshall, "Active learning through a covering lens," 2022.

[50] R. Mahmood, S. Fidler, and M. T. Law, "Low-budget active learning via wasserstein distance: An integer programming approach," in *International Conference on Learning Representations*, 2022.

[51] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," 2019.

[52] K. Kim, D. Park, K. I. Kim, and S. Y. Chun, "Task-aware variational adversarial active learning," 2020.

[53] C. Shui, F. Zhou, C. Gagné, and B. Wang, "Deep active learning: Unified and principled method for query and training," 2020.

[54] J. Wu, J. Chen, and D. Huang, "Entropy-based active learning for object detection with progressive diversity constraint," 2022.

[55] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision*, vol. 113, pp. 113–127, Jun 2015.

[56] A. Kirsch, J. Van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," *Advances in neural information processing systems*, vol. 32, 2019.

[57] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, p. 2591–2600, Dec. 2017.

[58] M. Gao, Z. Zhang, G. Yu, S. Ö. Arık, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), Springer International Publishing, 2020.

[59] M. Yuan, H.-T. Lin, and J. Boyd-Graber, "Cold-start active learning through self-supervised language modeling," 2020.

[60] L. Chen, Y. Bai, S. Huang, Y. Lu, B. Wen, A. Yuille, and Z. Zhou, "Making your first choice: To address cold start problem in medical active learning," in *Medical Imaging with Deep Learning* (I. Oguz, J. Noble, X. Li, M. Styner, C. Baumgartner, M. Rusu, T. Heinmann, D. Kontos, B. Landman, and B. Dawant, eds.), vol. 227 of *Proceedings of Machine Learning Research*, pp. 496–525, PMLR, 10–12 Jul 2024.

[61] N. Samet, O. Siméoni, G. Puy, G. Ponimatkin, R. Marlet, and V. Lepetit, "You never get a second chance to make a good first impression: Seeding active learning for 3d semantic segmentation," 2023.

[62] W. Lin, X. Ding, Y. Huang, and H. Zeng, "Self-supervised video-based action recognition with disturbances," *IEEE Transactions on Image Processing*, vol. 32, pp. 2493–2507, 2023.

[63] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Comput. Surv.*, vol. 55, jul 2023.

[64] L. Tao, X. Wang, and T. Yamasaki, "An improved inter-intra contrastive learning framework on self-supervised video representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5266–5280, 2022.

[65] R. Li, Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Motion-focused contrastive learning of video representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2105–2114, October 2021.

[66] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 5679–5690, Curran Associates, Inc., 2020.

[67] K. Hu, J. Shao, Y. Liu, B. Raj, M. Savvides, and Z. Shen, "Contrast and order representations for video self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7939–7949, October 2021.

[68] A. Singh, O. Chakraborty, A. Varshney, R. Panda, R. Feris, K. Saenko, and A. Das, "Semi-supervised action recognition with temporal contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10389–10399, June 2021.

[69] G. LORRE, J. Rabarisoa, A. Orcesi, S. Ainouz, and S. Canu, "Temporal contrastive pretraining for video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[70] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 8547–8555, May 2021.

[71] P. Munjal, N. Hayat, M. Hayat, J. Sourati, and S. Khan, "Towards robust and reproducible active learning using neural networks," in *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022* (K. Dana, G. Hua, S. Roth, D. Samaras, and R. Singh, eds.), Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (United States of America), pp. 223–232, IEEE, Institute of Electrical and Electronics Engineers, 2022. Publisher Copyright: © 2022 IEEE.; IEEE Conference on Computer Vision and Pattern Recognition 2022, CVPR 2022 ; Conference date: 19-06-2022 Through 24-06-2022.

[72] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019.

[73] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," NIPS '20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[74] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Computer Vision – ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 318–335, Springer International Publishing, 2018.

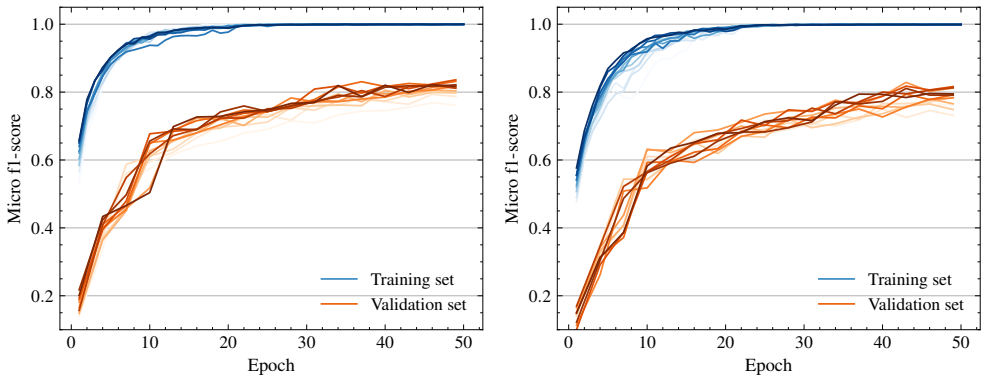
[75] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[76] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," in *Neurips*, 2020.

Appendix A

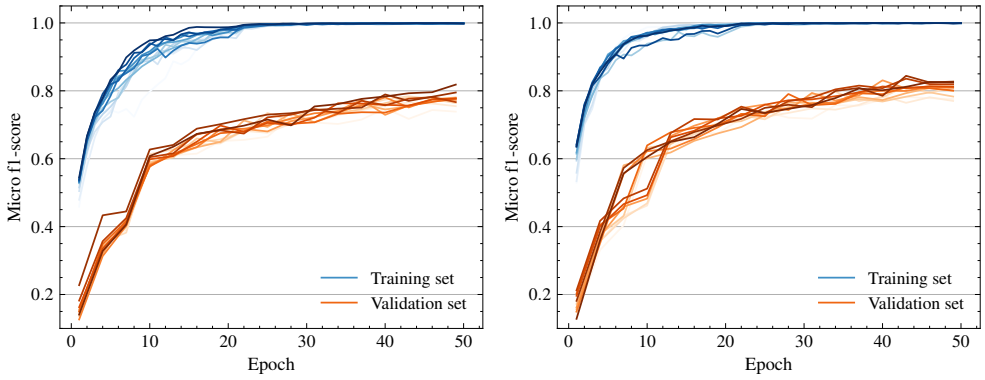
Learning dynamics of active learning- and self-supervised pre-training experiments

A.1 Active learning (baseline models)



(A) Random

(B) Max-Entropy

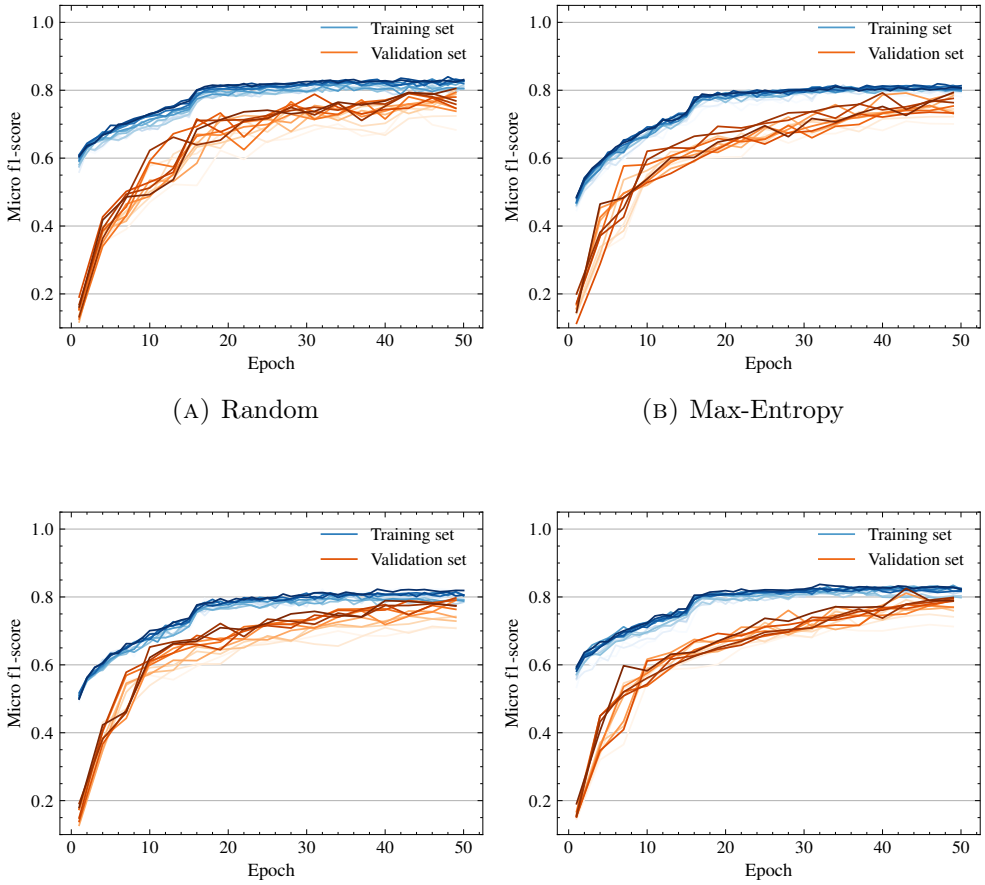


(C) Cluster-Margin

(D) Core-Set

FIGURE A.1: Learning dynamics of the evaluated active learning strategies when applied to baseline models (starting from an initial budget of 5%)

A.2 Active learning (pre-trained models)

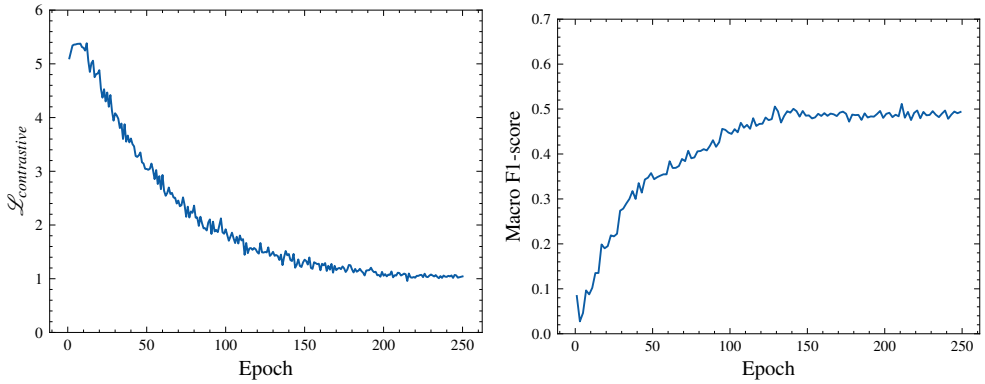


(A) Random (B) Max-Entropy

(C) Cluster-Margin (D) Core-Set

FIGURE A.2: Learning dynamics of the evaluated active learning strategies when applied to self-supervised pre-trained models (starting from an initial budget of 5%)

A.3 Self-supervised contrastive pre-training



(A) Contrastive loss (B) Macro F1-score

FIGURE A.3: Learning dynamics observed during self-supervised contrastive pre-training. The contrastive loss objective has been visualized on the left and on the right the macro-average F1-score when performing a linear evaluation of the behavior classification task

Appendix B

Practical implementation

B.1 Backbone architecture

The S3D [74] architecture was chosen as the backbone architecture for all experiments in this research. This backbone was chosen due to its ubiquitous use in prior works.

B.2 Pre-training

All pre-training experiments were conducted using a single NVIDIA A100 GPU. We use momentum contrastive learning (MoCo) [72] with InfoNCE loss as our objective function. All models were trained for 250 epochs with a batch size of 32 and a queue size of 512. Gradient descent with the Adam optimizer [75] was applied, starting from a learning rate of $1 \cdot 10^{-3}$ and a weight decay of $5 \cdot 10^{-4}$. Other hyper-parameters were kept identical to those used in previous research by Han et al. [76]. For more details about the implementation, we refer to their work.

B.3 End-to-end fine-tuning

All fine-tuning experiments were conducted using a single NVIDIA Tesla T4 GPU. In the experiments that do not use contrastive learning, we apply transfer learning and fine-tune the model trained on the Kinetics-400 dataset[74]. Each model was fine-tuned for 50 epochs using a batch size of 16. SGD was used as an optimizer with an initial learning rate of $1 \cdot 10^{-3}$ and weight decay of $1 \cdot 10^{-4}$. Dropout was applied before the final classification layer with a probability of 0.3. Each sample was augmented by cropping, flipping, or adjusting the brightness, contrast, and saturation values. For more implementation details we kindly refer to the repository attached to this work.

B.4 Linear evaluation

Similarly to end-to-end fine-tuning, experiments were conducted using a single NVIDIA Tesla T4 GPU. We follow a consistent linear evaluation protocol in which we connect one or more dense layers to the output of the frozen encoder. In the case of multiple dense layers, we introduce non-linearity by applying ReLU as an activation function between dense layers. We apply global batch normalization and train the classification head on the downstream task for 50 epochs. Batch size was set to 16 and SGD was used as an optimizer with an initial learning rate of $1 \cdot 10^{-2}$ and weight decay of $1 \cdot 10^{-4}$.