International Journal of Computer Vision (2023) 131:572–590
https://doi.org/10.1007/s11263-022-01716-3

Going Deeper than Tracking: A Survey of Computer-Vision Based Recognition of Animal Pain and Emotions

Sofia Broomé[1] · Marcelo Feighelstein[2] · Anna Zamansky[2] · Gabriel Carreira Lencioni[3] · Pia Haubro Andersen[6] · Francisca Pessanha[7] · Marwa Mahmoud[4] · Hedvig Kjellström[1,5] · Albert Ali Salah[7,8]

Received: 29 April 2022 / Accepted: 7 November 2022 / Published online: 25 November 2022

Abstract
Advances in animal motion tracking and pose recognition have been a game changer in the study of animal behavior. Recently, an increasing number of works go 'deeper' than tracking, and address automated recognition of animals' internal states such as emotions and pain with the aim of improving animal welfare, making this a timely moment for a systematization of the field. This paper provides a comprehensive survey of computer vision-based research on recognition of pain and emotional states in animals, addressing both facial and bodily behavior analysis. We summarize the efforts that have been presented so far within this topic—classifying them across different dimensions, highlight challenges and research gaps, and provide best practice recommendations for advancing the field, and some future directions for research.

Keywords Affective computing · Non-human behavior analysis · Pain estimation · Pain recognition · Emotion recognition · Computer vision for animals

Communicated by SILVIA ZUFFI.

Sofia Broomé, Marcelo Feighelstein, Anna Zamansky and Gabriel Carreira Lencioni have contributed equally to this work.

✉ Sofia Broomé
sbroome@kth.se

Marcelo Feighelstein
feighels@gmail.com

Anna Zamansky
annazam@is.haifa.ac.il

Gabriel Carreira Lencioni
gabriel.lencioni@gmail.com

Pia Haubro Andersen
pia.haubro.andersen@slu.se

Francisca Pessanha
f.pessanha@uu.nl

Marwa Mahmoud
marwa.mahmoud@glasgow.ac.uk

Hedvig Kjellström
hedvig@kth.se

Albert Ali Salah
a.a.salah@uu.nl

[1] Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden

[2] Tech4Animals Lab, Information Systems Department, University of Haifa, Haifa, Israel

[3] Department of Preventive Veterinary Medicine and Animal Health, School of Veterinary Medicine and Animal Science, University of São Paulo, São Paulo, SP, Brazil

[4] School of Computing Science, University of Glasgow, Glasgow, UK

[5] Silo AI, Stockholm, Sweden

[6] Department of Clinical Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

[7] Department of Information and Computing Sciences, Universiteit Utrecht, Utrecht, The Netherlands

[8] Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

1 Introduction

It is now widely accepted that animals can experience not only negative emotional states and pain (Sneddon et al., 2014), but also positive emotional states (de Vere & Kuczaj, 2016; Birch et al., 2021). Although traditionally, animal welfare science focus has been on pain and suffering, a recent paradigm shift is also addressing quality of life in a broader sense, seeking an understanding of animals' positive affective experiences (Duncan, 1996; Boissy et al., 2007).

There is no common agreement on what constitutes animal emotions (see Paul & Mendl, 2018, Kret et al., 2022 for comprehensive reviews). However, emotions are often described as internal states which are expressed in physiological, cognitive and behavioral changes (Anderson & Adolphs, 2014). Pain, traditionally studied separately from emotions, also has an affective component and is described as "an unpleasant sensory and emotional experience" (Raja et al., 2020). Hereafter, we use the term affective states to include both pain and emotions. Due to the subjective nature of affective states, their identification and measurement is particularly challenging, especially seen the lack of verbal basis for communication in non-human animals (hereon referred to as animals). To address this problem, the mentioned changes are often used as putative indicators (Mendl et al., 2010).

While physiological and cognitive changes are difficult to observe, measuring behavior is considered one of the most promising and least invasive methods for studying affective states. It is widely agreed that facial and body expressions may convey information on emotional states (Descovich et al., 2017; Diogo et al., 2008), including the affective components of pain. These expressions are produced and used for communication by most mammalian species (Diogo et al., 2008; Briefer et al., 2015; Schnaider et al., 2022; Sénèque et al., 2019).

Traditional methods for measuring behavior are based on either direct observation or by video recording analysis of one or more subjects, documenting carefully pre-designed behavioral categories, designed to be as unambiguous as possible (Bateson & Martin, 2021). The categories and ethograms may be designed depending on the research question.

For objective measurement of facial expressions of humans, the Facial Action Coding System (FACS) was developed to describe observable movements of facial muscles in terms of facial action units (AUs) (Ekman & Friesen, 1978). Likewise, the Body Action and Posture Coding System (BAP) (Dael et al., 2012) was designed for coding body movements in human behavioral research. With the same goal, coding systems were developed for other animals for both face (Correia-Caeiro et al., 2021; Waller et al., 2013; Wathan et al., 2015; Caeiro et al., 2017) and body (Huber et al., 2018).

Methods based on human observation and manual coding carry several serious limitations. They often require extensive human training, as well as rater agreement studies for reliable application. Furthermore, they are time consuming, and prone to human error or bias (Anderson & Perona, 2014). Computational tools, and especially tools based on computer vision, provide an attractive alternative (Anderson & Perona, 2014), since they are non-invasive, enable 24 h a day surveillance, save human effort and have the potential to be more objective than human assessments (Andersen et al., 2021).

In the human domain, automated facial and body behavior analysis is a rapidly expanding field of research. Accordingly, many datasets are available with extensive annotations for emotional states are available. Comprehensive surveys cover analysis of facial expressions (Li & Deng, 2022b), and body behaviors (Noroozi et al., 2018), with the recent trend being multi-modal emotion recognition approaches (Sharma & Dhall, 2021), combining, e.g., facial expressions, body behaviors and vocalizations. In the context of pain, numerous works have addressed facial expression assessment in humans (Al-Eidan et al., 2020; Hassan et al., 2021), and, notably, in infants (Zamzmi et al., 2017).

Although research concerned with animal behavior have so far lagged behind the human domain with respect to automation, recently, the field is beginning to catch up. This is owing in part to developments in animal motion tracking with the introduction of general platforms, such as DeepLabCut (Mathis et al., 2018), EZtrack (Pennington et al., 2019), Blyzer (Amir et al., 2017), LEAP (Pereira et al., 2019), DeepPoseKit (Graving et al., 2019) and idtracker.ai (Romero-Ferrero et al., 2019). However, as pointed out in Forkosh (2021), "being able to track the intricate movements of a spider, a cat, or any other animal does not mean we understand its behavior". Similarly, presenting good rater agreement on a given behavior does not mean that the behavior actually measures a given emotion. Automated recognition of pain and emotions is an important and difficult problem that requires going deeper than tracking movements, to assess whether the observable behaviors in fact correspond to internal states. Such analysis of facial expressions and body language brings about challenges that are not present in the human domain, as described in Pessanha et al. (2022). In particular, these are related to data collection and ground truth establishment (further discussed in Sect. 4.1.2), and a great variety of morphological differences, shapes and colors within and across animal species.

Indeed, in recent years the number of vision-based research articles addressing these topics is growing. To promote systematization of the field, as well as to provide an overview of the methods that can be used as a baseline for future work, this paper aims to provide a comprehensive survey of articles in this area, focusing on the visual modality. Based on the survey, we also provide technical best-practice recommendations for future work in the area, and propose future steps to further promote the field of animal affective computing.

2 Survey Scope and Structure

This survey covers works addressing automated recognition of pain and emotions in animals using computer vision techniques. This means that many interesting works related to automation of recognition of behavior in animals are left out of scope, including the large and important fields of ani-

mal motion tracking, precision livestock farming, methods for landmarks detection and 3D modeling of animal shapes. Moreover, due to our vision focus, we only consider research focused on the analysis of images or video, excluding works that are based on audio, wearable sensor data, or physiological signals.

We begin with an overview of relevant background within affective states research in non-human mammals (hereafter referred to as mammals) in Sect. 3. This section will be rather condensed, as previously published work covers this topic well. We provide pointers to this literature and summarize the findings. Section 4 provides an overview of computer vision-based approaches for classification of internal affective states in animals. We also include articles that perform facial AU recognition, since this task is closely connected to pain or affective states assessment. The articles included were identified by web search on Google using the terms 'automated animal pain/affect/emotion recognition/detection', 'pain/affect/emotion recognition/detection in animals', 'computer vision based recognition of pain/affect/emotion in animals', and by tracing references of the different works on Google Scholar.

We dissect these works according to the different workflow stages: data acquisition, annotation, analysis, and performance evaluation, respectively. For each of these steps, we identify and highlight the different approaches taken together with common themes and challenges. Based on our dissection and drawing parallels with human affective computing research, Sect. 5 provides some best practice guidance for future work related to technical issues, such as data imbalance, cross-validation and cross-domain transfer. Section 6 draws further conclusions from our analysis, identifying crucial issues that need to be addressed for pushing the field forward, and reflects on future research directions.

3 Research on Pain and Emotions in Animals

In 2012, the Cambridge declaration on consciousness was signed, stating that "The absence of a neocortex does not appear to preclude an organism from experiencing affective states". This implies that in addition to all mammals, even birds and other species, such as octopuses and crustaceans potentially experience emotions (Birch et al., 2021; Low et al., 2012). Affective states incorporate emotions, mood and other sensations, including pain, that have the property of valence (Mendl & Paul, 2020). Although the expression of emotions has been heavily discussed, there is no clear-cut definition of each of those terms, especially when referring to non-verbal individuals. As an example, a common approach for emotions is to consider them as intense, short-term affective states triggered by events in which reinforcers (positive reinforcers or rewards, and negative reinforcers or punishers) are present or expected (Paul & Mendl, 2018; Dawkins, 2008). Plutchik (1979) further suggested that emotions should provide a function aiding survival, which is relevant, but difficult to use as a working definition.

There are also different approaches for the classification of emotional states, with one of the most prominent being the discrete one. According to this theory, animals have a certain number of fundamental emotion systems, based on neuronal structures of different brain areas homologous across species (Panksepp, 2010), leading to a discrete set of distinct emotional states. Paul Ekman, for instance, described the following six distinct emotions in humans: fear, sadness, disgust, anger, happiness, surprise (Ekman, 1992), but other discrete classifications have been suggested as well (e.g., by Panksepp 2010). An alternative classification system is the dimensional approach, which classifies emotions according to different dimensions, usually implying valence and mostly also arousal (but some authors also describe additional dimensions) (Mendl & Paul, 2020; Posner et al., 2005). Anecdotal evidence that animals can experience secondary emotions as grief, jealousy and more, is compelling (Morris et al., 2008; Uccheddu et al., 2022). According to this view, several affective states could be manifested at the same time, making the recognition of affective states in animals even more challenging.

Pain research developed separately from emotion research in both human and animals, despite the close links between them (Hale & Hadjistavropoulos, 1997). According to the International Association for the Study of Pain, human pain is defined as "an unpleasant sensory and emotional experience associated with actual or potential damage, or described in terms of such damage" (Raja et al., 2020); thus, the emotional dimension of pain can be considered an affective state.

Pain assessment in infants is considered one of the most challenging problems in human pain research (Anand et al., 2007), due to the issue of non-verbality in neonates and older infants, analogously to animals. Historically, even the ability of human neonates to feel pain and emotions was questioned (Grunau & Craig, 1987; Camras & Shutter, 2010). As late as the 1980s, it was widely assumed that neonates did not experience pain as we do, together with a hesitancy to administrate opiates to these patients, and surgery was often performed without anaesthesia (Fitzgerald & McIntosh, 1989). Duhn and Medves (2004) provide a systematic review of instruments for pain assessment in infants, where facial expressions are identified as one of the most common and specific indicators of pain. Facial expression of pain in neonates is defined as the movements and distortions in facial muscles associated with a painful stimulus; the facial movements associated with pain in infants include deepening of the nasolabial furrow, brow lowering, narrowed eyes, chin quiver and more (Zamzmi et al., 2017).

The assessment of pain and emotions in mammals is much less explored than in the human domain, due to the difficulties regarding ground truth and subsequent lack of large databases mentioned above. The pressing need for these assessments in animal health and welfare evaluations, has therefore made researchers resort to addressing measurements of physiological, behavioral, and cognitive components of affective states, which can be measured objectively, and even in many cases automatically (Paul et al., 2005; Kret et al., 2022). This involves physiological (such as heart rate, hormone levels, body temperature) and behavioral parameters (vocalisations, facial expressions, body postures). Naturally, behavioral parameters are particularly relevant when exploring computer vision-based approaches.

Facial expressions, produced by most mammalian species (Diogo et al., 2008) are one important source of information about emotional states (Descovich et al., 2017; Diogo et al., 2008). Behavioral parameters such as facial expressions are not only non-invasive to observe, but have also proved more reliable than physiological indicators (Gleerup et al., 2015). The latter are significantly influenced by diseases and can only be used in controlled settings (Andersen et al., 2021; Gleerup & Lindegaard, 2016). Adaptations of FACS to several other species have been used for measuring facial behavior in non-human primates (e.g., Correia-Caeiro et al., 2021), dogs (Waller et al., 2013), horses (Wathan et al., 2015) and cats (Caeiro et al., 2017). Grimace scales can be less demanding to apply than FACS, since they analyze movements and behavior changes in a small set of facial regions related to pain (McLennan & Mahmoud, 2019; Andersen et al., 2021). Further, facial behavior such as eye blink rates and twitches (Merkies et al., 2019; Mott et al., 2020), as well as yawning, have been related to stress and stress handling. A correlation between positive emotions and animal facial behavior has also been shown; as an example, in cows, the visibility of the eye sclera dropped during positive emotional states (Proctor & Carder, 2015).

In addition to facial expressions, other behavioral indicators have been studied in order to assess affective states in animals. These parameters have also been used to estimate the valence of each affective state, from negative to positive (Ede et al., 2019; Hall et al., 2018; Lansade et al., 2018). Similarly to facial behavior, body posture and movement have been correlated to a range of emotions and pain-related behavior (Walsh et al., 2014; Sénèque et al., 2019; Dyson et al., 2018; Briefer et al., 2015; Schnaider et al., 2022). Further, several protocols have also been developed to assess behavioral indicators such as changes in consumption behaviors (time activity budgets for eating, drinking, or sleeping, etc.) (Oliveira et al., 2022; Auer et al., 2021; Maisonpierre et al., 2019), anticipatory behaviors (Podturkin et al., 2022), affiliative behavior (Clegg et al., 2017), agonistic behaviors, and displacement behaviors, amongst others (Foris et al., 2019).

Further, tests such as Open field, Novel object and Elevated plus maze (Lecorps et al., 2016), as well as Qualitative Behavior Assessment (QBA) have also been used (Kremer et al., 2020) for affective state assessment in animals. Less used is the Body and Posture coding system, which recently was adapted for use in dogs (Waller et al., 2013). The advantage of using an "exhaustive" coding scheme is that the coding can be done without any anticipation of what will be detected. In contrary, in a pain score scale, pain is always the issue.

When carefully coding facial actions of horses in pain and horses with emotional stress from isolation, it was found that pain is associated with some degree of emotional stress (Lundblad et al., 2021). Stress is a physiological (endocrine) reaction to threats, but similarly to pain, it has an emotional component, which is what we refer to here. Additionally, stress without pain may have some similarities to pain during the acute stages (Lundblad et al., 2021). External inputs that may induce stress can therefore influence prototypical facial expressions. The mixing problem holds for other affective states as well, adding to the challenge of recognizing specific internal states. Because of these challenges, assessing emotions such as the ones resultant from stressful situations may not be as direct as pain recognition, for which there are several validated indicators (Lundblad et al., 2021; Mayo & Heilig, 2019). This might also be the reason for the lack of automated methods regarding this matter.

4 Overview of Computer Vision-Based Approaches for Classification of Pain and Emotions

To systematize the existing body of research, in this section, we review and analyze twenty state-of-the-art works addressing assessment, classification and analysis of animal emotions and pain. The works are presented in Table 1, and are classified according to the following characteristics:

- *Species:* We restrict the scope of our review to mammals. The list of included species can be seen in Table 1.

- *State and state classifier:* Our main aim is to cover works focusing on recognition of internal states in animals, falling under the categorization of emotion and pain. Thus, the main type of works we are interested in are those providing a classification method (pain score, or classifying specific emotional states). In the 's column 'State classifier', this is signified by '+'. There are five works in the table marked with '−', which do not provide such a classifier, but nevertheless develop computer vision-based methods explicitly designed to investigate behavior patterns related to pain states or emotions.

- *Stimulus:* This category designates the type of stimulus that the animals have been subject to during data collection to induce a particular emotional or pain state.
- *Focus area:* We restrict our attention here to facial and bodily behavior indicators, or a combination of the two.
- *State annotations:* This column is divided into two options: behavior-based or stimulus-based state annotations, respectively. The behavior-based annotations are purely based on the observed behaviors, without regard to when the stimulus (if there was one) occurred. For stimulus-based annotations, the ground-truth is based on whether the data was recorded during an ongoing stimulus or not.

In this section, the overview of the works in Table 1 is organized according to the different stages of a typical workflow in studies within this domain: data collection and annotation, followed by data analysis (typically, model training and inference) and last, performance evaluation. For each of these stages, we classify the methods and techniques applied in these works, highlight commonalities and discuss their characteristics, limitations and challenges.

4.1 Data Collection and Annotation

Pessanha et al. (2022) highlight some important challenges in addressing automated pain recognition in animals, most of which can also be generalized to recognition of emotions. The first challenge is the lack of available datasets, compared to the vast amount of databases in the human domain (Hassan et al., 2021). This is due to the obvious difficulties of data collection outside of laboratory settings, especially for larger animals—companion as well as farm animals. Secondly, particularly in the case of domesticated species selected for their aesthetic features, there may be much greater variation in facial texture and morphology than in humans. This makes population-level assessments difficult, due to the potential for pain-like facial features to be more present/absent in certain breeds at baseline (Finka et al., 2019). Finally, and perhaps most crucially: there is no verbal basis for establishing a ground truth label of pain or emotional state, whereas in humans, self-reporting of the internal affective state is commonly used[1]. This complicates data collection protocols for animals, sometimes requiring conditions where the induction of a particular affective state and its intensity must be closely controlled and regulated, and/or requires rating by human experts, potentially introducing biases. Below we examine the data collection choices and data annotation methods in the works reviewed here.

[1] It can be noted that exact ground truth is not available for humans either; the labels remain subjective. This means that we can only claim to have detected a given affective state up to some constant of uncertainty.

4.1.1 Data Collection

Recording equipment. Choosing the equipment with which to record visual data is the first step to acquire data. Since this survey concerns vision-based applications, this would either be an RGB, depth or infrared camera. As stated in Andersen et al. (2021), the requirement on resolution for machine learning applications is often not a limiting factor. Some of the most frequently used deep neural network approaches work well with inputs of approximately 200×200 pixels. However, subtle cues, such as muscle contractions, can be difficult to detect reliably in low-resolution images. Infrared cameras are typically used to be able to monitor behavior during night, in order not to disturb the sleep cycle of the animal with artificial light. While clinical annotations are typically done by veterinarians, animal images scraped from the Internet without any expert annotations can also be useful for training computer vision tools. Large object recognition datasets such as MS COCO include animal classes, although from a limited number of species (Lin et al., 2014). Therefore, models trained with those datasets can be helpful for detecting animals and for pose estimation.

Environment. Rodents are typically recorded in observation cages with clear walls to permit their recording (Tuttle et al., 2018). The camera is static and can cover the entire cage, but for facial analysis, only frames that have the rodent in frontal view are selected and used. Infrared cameras placed on top of the face are also used, but these observe movement patterns, and not facial expressions. Equines are recorded in a box (Rashid et al., 2022; Ask et al., 2020) from multi-view surveillance cameras, or in open areas, but with static cameras placed at a distance to capture the animal from the side (Gleerup et al., 2015; Hummel et al., 2020; Broomé et al., 2019, 2022; Pessanha et al., 2022), or frontally, when the animal is next to a feeder (Lencioni et al., 2021). The side view makes observing the bodily behavior easier, but only one side of the face is visible. The presence of a neck collar or bridle is common for recordings, as the animals are often constrained. Sheep are recorded outdoors, in farms, with widely varying background and pose conditions (Mahmoud et al., 2018). Recordings of animals from veterinary clinics, on the other hand, uses static cameras indoors, where the animal can move freely in a room (Zhu et al., 2022). This allows the expert to evaluate behavioral cues during movement.

Participants. Controlling the data for specific characteristics of participants can lead to increased performance and its better understanding. On the other hand, generalizability of such models can be limited. Some studies reviewed here practiced control for color (e.g. white mice Tuttle et al., (Tut-

Table 1 A summary of the reviewed works, categorized by the state in question, potential stimulus, focus area, whether the work includes a state classifier and whether the state annotations are behavior- or stimulus-based

Study	Species	State	Stimulus	Focus area	State classifier	State annotations	
						Behavior-based	Stimulus-based
Tuttle et al. (2018)	Mice	Pain	Vet. procedure	Face	+	✓	✓
Andresen et al. (2020)			Vet. procedure	Face	+	✓	✓
Mahmoud et al. (2018)	Sheep	Pain	Unknown or naturally occurring	Face	+	✓	
Pessanha et al. (2020)			Unknown or naturally occurring	Face	+	✓	
Lencioni et al. (2021)	Horses	Pain	Surgical castration	Face	+	✓	
Hummel et al. (2020)			Unknown or induced pain	Face	+	✓	
Pessanha et al. (2022)			Unknown or induced pain	Face	+	✓	
Broomé et al. (2019)			Induced pain	Body and face	+		✓
Broomé et al. (2022)			Induced pain	Body and face	+		✓
Rashid et al. (2022)			Induced pain	Body	+		✓
Reulke et al. (2018)			Vet. procedure	Body	–		✓
Corujo et al. (2021)		Emotion	Unknown	Body and face	+	✓	
Li et al. (2021)		–	–	Face	–	–	–
Feighelstein et al. (2022)	Cats	Pain	Vet. procedure	Face	+		✓
Morozov et al. (2021)	Macaques	Emotion	Induced behavior	Face	–	✓	
Blumrosen et al. (2017)			Induced behavior	Face	–	✓	
Zhu et al. (2022)	Dogs	Pain	Naturally occurring	Body	+		✓
Franzoni et al. (2019)		Emotion	Unknown	Face	+	✓	
Boneh-Shitrit et al. (2022)			Induced behavior	Face	+	✓	✓
Ferres et al. (2022)			Unknown	Body	+		✓
Statham et al. (2020)	Pigs	Emotion	Induced behavior	Body	–		✓

tle et al., 2018), black mice (Andresen et al., 2020), white pigs (Statham et al., 2020)), breed (e.g. British Short Haired cats (Feighelstein et al., 2022), Labrador Retriever dogs (Boneh-Shitrit et al., 2022)), and sex (e.g. female cats (Feighelstein et al., 2022), male horses (Lencioni et al., 2021)).

4.1.2 Data Annotation

In the human domain, self-reporting is considered one of the most unobtrusive and non-invasive methods for establishing ground truth in pain (Labus et al., 2003) and emotion research (Barrett, 2004). Furthermore, in emotion research the use of actors portraying emotions is a common method for data collection and annotation (Seuss et al., 2019). For obvious reasons, these methods are not usable for animals, making the establishment of ground truth with respect to their internal states highly challenging, and adding further complications to the data annotation stages.

One possible strategy for establishing the ground truth can be based on designing or timing the experimental setup to induce the pain or emotion. In the case of pain, designing can refer to experimental induction of clinical short term reversible moderate pain using models known from human volunteers. In Broomé et al. (2019), e.g., two ethically regulated methods for experimental pain induction were used: a blood pressure cuff placed around one of the forelimbs of horses, or the application of capsaicin (chili extract) on the skin of the horse. Another possibility is to time data collection after a clinical procedure. This is the case in Feighelstein et al. (2022), where female cats undergoing ovariohysterectomy were recorded at different time points pre- and post-surgery.

In the case of emotions, state induction can be performed, e.g., using triggering stimuli and training to induce emotional responses of different valence. For instance, in Boneh-Shitrit et al. (2022) the data was recorded using a protocol provided in Bremhorst et al. (2019), using a high-value food reward used as the triggering stimulus in two conditions—a positive condition predicted to induce positive anticipation, and a negative condition predicted to induce frustration in dogs. In Lundblad et al. (2021), stress was induced by letting out one out of two horses that were normally let out together (herd mates). After between 15 and 30 minutes alone, the horse that was not let out showed a marked stress response. The presence of people (such as the owner of pets) can influence the behavior of the animal, and should be taken into account in analyses. In cases when no control over the state of the animal is exercised, the animal can be recorded in a naturalistic setting, such as farms (Mahmoud et al., 2018; Li et al., 2021) or stables, or even laboratory cages. Data collected from veterinary clinics with "naturally occurring" pain

denotes animals brought to the clinic under pain, as opposed to "induced" pain, which is a more controlled setting.

Cases when data is scraped from the Internet (Franzoni et al., 2019; Ferres et al., 2022) should be treated with caution in this context (and are accordingly labeled as 'unknown' in Table 1), as the degree of control cannot be asserted. When the state is not controlled, the only available option for establishing ground truth is by human annotators. This may introduce bias and error, depending on the annotators' expertise (veterinarians, behavior specialists, laymen), the number of annotators and agreement between them, and also on whether specific measurement are instruments used (e.g., validated grimace scales (Dalla Costa et al., 2014)).

Table 1 includes a classification of the works reviewed here according the data annotation strategies discussed above. In the studies of Tuttle et al. (2018), Andresen et al. (2020) and Feighelstein et al. (2022), the animal participants underwent a surgical procedure. In Tuttle et al. (2018) and Andresen et al. (2020), the obtained images were then rated by human experts based on the mouse grimace scale. In Feighelstein et al. (2022), on the other hand, the images were taken at a point in time where the presence of pain was reasonable to assume (between 30-60 min. after the end of surgery, and prior to administration of additional analgesics). In Broomé et al. (2019, 2022) and Rashid et al. (2022) experimental pain is induced using controlled procedures for moderate and reversible pain. The dataset used in Hummel et al. (2020); Pessanha et al. (2022) is composed from several sources: a clinical study, where pain was experimentally induced, images taken at a home housing older horses, and images provided by horse owners. In Corujo et al. (2021), the data is collected from different private sources where the horse and context of the photo was familiar, guiding annotation. However, the state annotation was performed by laymen. Since the states or contexts are not described for this dataset, we have marked this as 'unknown' in Table 1. Franzoni et al. (2019); Ferres et al. (2022) use images scraped from the web, thus the state control is stated 'unknown'. Boneh-Shitrit et al. (2022) uses images of dogs taken in an experiment where emotional states are induced by food rewards, with no human involved in the annotation loop.

Induced approaches, if performed properly and in a controlled and reproducible manner, have the potential to reduce human bias and error, while the use of data from unknown sources can be problematic in terms of bias, error and noise in ground truth annotation (Waran et al., 2010; Price et al., 2002).

4.2 Data Analysis

The stage of data analysis typically involves developing a data processing pipeline, the input of which is images or videos, and the output of which is a classification of an emotions, or pain classification (either binary yes/no or degree assessment with more than two classes). The pipeline may involve one or more steps, and address body/face as a whole, or process first their specific parts.

4.2.1 Input: Frames Versus Sequences of Frames

Computer vision-based methods operate on data in the form of images or image sequences (videos). This implies the following three main modes of operation with respect to temporality:

- *Single frame basis.* This route is taken by the majority of works reviewed in the survey (those marked with 'frame' in Table 2). This is the simplest and least expensive option in terms of computational resources.
- *Frame aggregation.* Using frame-wise features, some works address classification of videos by aggregating the results of classifiers working with single frames, thus at least partially incorporating information contained in sequences of frames. This is the route taken in Tuttle et al. (2018) and Pessanha et al. (2020) (marked with 'frame(ag)' in Table 2).
- *Using spatio-temporal representations.* A third route is to learn spatiotemporal features from video given as sequential input to a deep network. This is done in Broomé et al. (2019, 2022) and Zhu et al. (2022), and enables the detection of behavioral patterns that extend over time. Apart from presenting computationally heavy training, this method requires more data than frame-wise approaches. On the other hand, having access to video recordings often is synonymous to having access to a lot of data. However, this is relative, and the horse video datasets used in Broomé et al. (2019, 2022), which have a duration of around ten hours each, are comparable in scale to older well-known video datasets such as UCF-101 (Soomro et al., 2012) (30 h), but not to newer ones, such as Kinetics (Kay et al., 2017) (400 h).

As was found in Broomé et al. (2019), for the case of horse pain detection, temporal information is crucial for discriminating pain. Temporal information has also previously been found to improve recognition of human pain (Bartlett et al., 2014). In Rashid et al. (2022), the use of single-frame and sequential inputs for pain classification are also compared, in a multiple-instance learning (MIL) setting. MIL can be seen as lying somewhere in between temporal aggregation and spatiotemporal features, in being a more advanced form of temporal aggregation, within a learning framework. Using single frames gives more control and promotes explainability, but leads to information loss. Working with video input, on the other hand, rather than single-frame input, is costly.

Thus, choosing the mode of operation is ultimately goal dependent. If the goal is to count in how many frames in a certain video segment that a horse has its ears forward (an estimate of the fraction of time with ears kept forward), it suffices to detect forward ears separately for each frame, and subsequently aggregate the detections across the time span of interest. If, on the other hand, the goal is to study motion patterns of the horse, or distinguish between blinks and half-blinks for an animal, it is crucial to model the video segment spatiotemporally. An explorative search for behaviors which potentially extends over time might also be desirable, and the degrees of freedom offered by spatiotemporal feature learning approaches is useful for such a task.

4.3 Parts-Based Versus Holistic Methods

Methods for computer vision-based human facial analysis are commonly divided into local parts-based and holistic methods, differing in the way facial information is processed (Wang et al., 2018; Wu & Ji, 2019). Parts-based methods divide the input data into different areas, e.g., considering different facial features separately, while holistic methods process the information of the input data as a whole, be it at the body or face level.

The idea of dividing the face into regions, or parts, is especially relevant for works on pain assessment that are based on species-specific grimace scales. Such scales typically divide the animal face into at least three parts, including ears, eyes and nose/mouth/nostrils (e.g., Dalla Costa et al., 2014). One example is the work of Lu et al. (2017), providing a multi-level pipeline for assessment of pain level in sheep, based on the sheep facial expression pain scale (SPFES (McLennan et al., 2016)), according to which the sheep face is divided into regions of eyes, ears and nose. Although the cheek and lip profile are also discussed in the SPFES, they are omitted in Lu et al. (2017), because the sheep dataset in question only contains frontal faces, and these features can hardly be seen on a frontal face. The eyes and ears are further split into right and left regions each. Each of these regions correspond to one out of three AUs defined based on the SPFES taxonomy (pain not present (0), pain moderately present (1), or pain present (2)). For instance, the ear region can correspond to one of the following AUs: ear flat (pain level = 0), ear rotated (pain level = 1) and ear flipped (pain level = 2). SVM classifiers predicting the pain level for each of the five regions were then trained separately on each facial feature, using Histogram of Oriented Gradients (HOG), to depict the shape and texture of each feature. To aggregate these results, the scores for symmetric features (eyes, ears) were averaged, and all three feature-wise scores (ear, eye, nose) were averaged again to obtain the overall pain score. It can be noted that SheepFACS has not yet been developed, and these facial expressions are thus referred to as AUs in a broader sense.

Another example of a parts-based approach is provided in Lencioni et al. (2021) in the context of horse pain. Based on the horse grimace scale (Dalla Costa et al., 2014), this work also focuses on three regions of the horse face: ears, eyes, and mouth and nostrils, training three separate pain classifier models based on convolutional neural networks (CNNs) for each of the regions. The outputs of these models are then fused using a fully connected network for an overall pain classification. A parts-based approach to AU recognition is presented in Li et al. (2021), where each AU is recognized on cropped image regions specific to the AU in question. Their results show that such close-up crops of the eye-region or lower-face region are necessary for the performance of the classification in their framework.

In general, as the field of pain and emotion recognition is only beginning to emerge, using a parts-based approach can provide important insights on the role of each of the facial regions in pain expression. Interestingly, the results of Lencioni et al. (2021) indicate that ears provide better indication for pain level in sheep and horses than the other regions (although this should be considered with caution due to the imbalance of the dataset in terms of different parts, see also discussion in Sect. 5). Further exploration of parts-based approaches in additional species can provide insights into the importance of the regions, and thus allow methods to fine-tune the aggregation of a general pain score in future studies. The column ‘Part/Holistic’ in Table 2 classifies the works across the dimension of holistic vs. parts-based approaches.

4.3.1 Hand-Crafted Versus Learned Features

A major focus in computer vision is to discover, understand, characterize, and improve the features that can be extracted from images. Traditional features used in the literature have been manually designed, or ‘hand-crafted’, overcoming specific issues like occlusions and variations in scale and illumination, such as histograms of oriented gradients (Nanni et al., 2017). Traditional computer vision methods, prior to the deep learning era, have typically been based on hand-crafted features. The shift toward automatically learning the feature representations from the data occurred progressively during the 2010s as larger datasets were made public, GPU-computing became more accessible and neural network architectures were popularized in both the machine learning literature and in open-source Python frameworks, such as Tensorflow (Abadi et al., 2015). This new computing paradigm is commonly known as deep learning, where the word deep refers to the hierarchy of abstractions that are learned from data, and stored in the successive layer parameters (LeCun et al., 2015).

The above context has important implications in the context of our domain. The first implication is related to dataset size: methods using hand-crafted features can be applied to

Table 2 An overview of the approaches taken at the analysis stage, categorized according to whether the methods are parts-based or holistic, based on frame-wise or video information, and whether the features are learned or hand-crafted

Study	Species	Part/Holistic	Input	Features
Tuttle et al. (2018)	Mice	Holistic	Frame(ag)	Learned
Andresen et al. (2020)	Mice	Holistic	Frame	Learned
Mahmoud et al. (2018)	Sheep	Parts-based	Frame	Hand-crafted (low-level)
Pessanha et al. (2020)	Sheep	Parts-based	Frame(ag)	Hand-crafted (low & high-level)
Lencioni et al. (2021)	Horses	Parts-based	Frame	Learned
Hummel et al. (2020)	Horses	Parts-based	Frame	Hand-crafted (low-level)
Pessanha et al. (2022)	Horses	Parts-based	Frame	Hand-crafted (low-level)
Broomé et al. (2019)	Horses	Holistic	Video	Learned
Rashid et al. (2022)	Horses	Holistic	Video	Learned
Reulke et al. (2018)	Horses	Holistic	Video	–
Corujo et al. (2021)	Horses	Holistic	Frame	Learned
Li et al. (2021)	Horses	Parts-Based	Frame	Learned
Feighelstein et al. (2022) 1	Cats	Holistic	Frame	Learned
Feighelstein et al. (2022) 2	Cats	Holistic	Frame	Hand-crafted (high-level)
Morozov et al. (2021)	Macaques	Holistic	Frame	Hand-crafted (high-level)
Blumrosen et al. (2017)	Macaques	Holistic	Frame	Hand-crafted (high-level)
Zhu et al. (2022)	Dogs	Holistic	Frame	Mixed
Franzoni et al. (2019)	Dogs	Holistic	Frame	Learned
Boneh-Shitrit et al. (2022)	Dogs	Holistic	Frame	Learned
Ferres et al. (2022)	Dogs	Emotion	Frame	Hand-crafted (high-level)
Statham et al. (2020)	Pigs	Emotion	Frame	–

small datasets, whereas deep learning methods require larger amounts of data. The second implication is the explainability of the approaches: hand-crafted features allows for a clearer understanding of the inner workings of the method, while learned features lead to 'black-box' reasoning, which may be less appropriate for clinical and welfare applications, such as pain assessment.

Hand-crafted features can exist on multiple levels, which we roughly divide into two: *low-level* features are technical and may consist of pre-defined notions of image statistics (such as histograms of oriented gradients, or pixel intensity in different patches of the image). *High-level* features, in our context, are semantically grounded, typically based on species-specific anatomical facial and/or body structure, grimace scales or AUs. As these features promote explainability, we refer to them as intermediate representations; these will be discussed in more detail further down.

The column 'Features' in Table 2 classifies the works across the dimension of learned vs. hand-crafted features. The types of high-level features used in Pessanha et al. (2020), Feighelstein et al. (2022), Ferres et al. (2022), Morozov et al. (2021) and Blumrosen et al. (2017) are further discussed in Sect. 4.4.

4.4 Increasing Explainability: Intermediate Representations

High-level features in this context are features that have semantic relations to the domain of affective states, e.g., through facial or bodily landmarks, grimace scale elements, AUs, or pose representations. A specific, e.g., facial landmark does not need to correspond directly to a given state, but they are on a higher level of abstraction compared to for example image statistics. As such, these are highly valuable for the explainability of the different classification methods. These features are usually used in computational pipelines involving a number of pre-processing steps. They can be built either manually, or using classifiers based either on lower-level hand-crafted or learned features. Below we discuss some important types of intermediate representations used in the works surveyed here, and how they are computed and used:

- *Facial Action Units*. Morozov et al. (2021) and Blumrosen et al. (2017) apply two different approaches to address the recognition of facial actions in macaques as an intermediate step towards automated analysis of affective states. Morozov et al. (2021) addresses six dominant AUs from macaque FACS (MaqFACS (Parr et al.,

2010)), selected based on their frequency and importance for affective communication, training a classifier on data annotated by human experts. Blumrosen et al. (2017) addresses four basic facial actions: neutral, lip smacking, chewing and random mouth opening, using an unsupervised learning approach without the need for annotation of data. Both works use eigenfaces (Donato et al., 1999) as hand-crafted lower level features, an approach which uses PCA analysis to represent the statistical features of facial images. Lu et al. (2017) provide a pipeline for pain level estimation in sheep, in which automated recognition of nine sheep facial expressions is performed using classifiers based on histograms of gradients as lower level hand-crafted features. The AUs are defined within SPFES, a standardised sheep facial expression pain scale (McLennan et al., 2016).

- *Landmarks/Keypoints*. One of the approaches investigated in Feighelstein et al. (2022) in the context of cat pain is based on facial landmarks, specifically chosen for their relationship with underlying musculature, and relevance to cat-specific facial action units (CatFACS (Caeiro et al., 2017)). The annotation of the 48 landmarks was done manually. In the pain recognition pipeline, these landmarks are transformed into multi-region vectors (Qiu & Wan, 2019), which are then fed to a multi-layer perceptron neural network (MLP).

  The approach of Ferres et al. (2022) for dog emotion recognition from body posture uses 23 landmarks on both body and face. The landmarks are automatically detected by a model based on the DeepLabCut framework (Mathis et al., 2018), and trained on existing datasets of landmarks (Cao et al., 2019; Biggs et al., 2020) containing subsets of the 23 landmarks. Two approaches are then examined for emotion classification: (1) feeding the raw landmarks to a neural network, and (2) computing body metrics introduced by the authors and feeding it to simpler decision tree classifiers to promote explainability. For the decision tree approach, the authors use a variety of body metrics, such as body weight distribution, and tail angle. The former is calculated using the slope of the dorsal line, which is a hypothetical line between the withers keypoint and the base of the tail keypoint, and the latter by the angle between the dorsal line, and the hypothetical line between the base of the tail and the tip of the tail.
- *Pose Representations.* In Rashid et al. (2022), multi-view surveillance video footage is used for extracting a disentangled horse pose latent representation. This is achieved through novel-view synthesis, i.e., the task of generating a frame from viewpoint j, given a frame from viewpoint i. The latent pose arises from a bottleneck in an encoder-decoder architecture, which is geometrically constrained to comply with the different rotation matrices between different viewpoints. The representation is useful in that it separates the horse from its appearance and background, to remove any extraneous cues for the task which may lead to overfitting. The representation is subsequently fed to a horse pain classifier. The pain classification is cast as a multiple instance learning problem, on the level of videos. In Zhu et al. (2022), a pose stream is combined with a raw RGB stream in a recurrent two-stream model to recognize dog pain, building on the architectures used in Broomé et al. (2019, 2022). This constitutes an interesting example of mixing intermediate with fully deep representations.

4.5 Going Deep: Black-Box Approaches

As noted above, deep learning approaches are becoming increasingly popular in the domain of human affective computing as they require less annotation efforts if transfer learning is leveraged, and no efforts for hand-crafting features. Yet, the resulting models provide what is called 'black-box' reasoning, which does not lend itself easily for explaining the classification decisions in human-understandable terms (see, e.g., London 2019). This is a crucial aspect, especially in the context of clinical applications and animal welfare.

The convolutional neural network (CNN) is the most popular type of deep model used in the works surveyed in this article. Examples of used CNN architectures include ResNet50 (Corujo et al., 2021; Feighelstein et al., 2022; Boneh-Shitrit et al., 2022; Andresen et al., 2020), InceptionV3 (Tuttle et al., 2018; Andresen et al., 2020) and AlexNet (Franzoni et al., 2019). One work addressing dog emotion (Boneh-Shitrit et al., 2022) compared a CNN (ResNet50) to a Vision Transformer, ViT (Dosovitskiy et al., 2021), a model fully based on attention mechanisms instead of convolutions, finding the latter to perform better. The authors hypothesize that this is due to the sensitivity of such models to object parts (Amir et al., 2021), and suggest that automated emotion classification requires understanding at the object-part level.

Another type of neural network is the deep recurrent video model used in Broomé et al. (2019, 2022), and Zhu et al. (2022), based on the ConvLSTM (Shi et al., 2015) layer. A ConvLSTM unit replaces matrix multiplication by convolution in the LSTM equations, thus allowing for spatial input rather than 1D vectors in a recurrent setting. In this way, spatial and temporal features can be learned simultaneously, instead down-sampling the spatial features prior to temporal modeling. The best performing version of the model in Broomé et al. (2019) takes both RGB and optical flow input in two separate streams with late fusion. In Broomé et al. (2019), this model is compared to a frame-wise InceptionV3 and to a VGG (Simonyan & Zisserman, 2015) network with a standard LSTM layer on top, thus taking sequential input. Even if the VGG+LSTM obtains numerical

results not far from the two-stream ConvLSTM, qualitative examples using Grad-CAM (Selvaraju et al., 2017) indicate that the ConvLSTM learns more relevant features. In Broomé et al. (2022), it is also found that an I3D model (Carreira & Zisserman, 2017) (a deep 3D convolutional neural network) can learn spatiotemporal features for pain recognition, but that it performs weaker in terms of generalization to a new pain type compared to the ConvLSTM model. It is hypothesized that this overfitting behavior of the I3D is due to its large parameter count (around 23M) relative to the ConvLSTM (around 1M), and that smaller video models may be advantageous for this type of fine-grained classification task, where motion cues should matter more than appearance and background of the videos. In Zhu et al. (2022), LSTM and ConvLSTM layers are used in a dual-branch architecture, where one branch processes keypoint-based representations, and the other RGB-based representations.

4.6 Performance Evaluation

Understanding and scrutinizing the methods for measuring performance are key when comparing approaches in recognition of affective states. In this section, we give an overview of the evaluation protocols as well as classification performances of the different approaches listed in Table 1. We emphasize that comparing the performance of classifiers of emotions and pain in animals presents great challenges, and cannot be done solely on the basis of the numbers as measured by performance metrics. This is due to the significant differences in data acquisition (different lighting conditions, camera equipment, recording angles, number of samples), as well as in ground truth annotation (naturalistic vs. controlled setting, induced vs. natural emotional state/pain, degree of agreement between annotators and their expertise). Even when all of these factors are comparable, technical choices such as differences in data balance or validation method greatly affect performance metrics. For these reasons, we have chosen to leave out an explicit discussion on the performance in terms of accuracy. However, we discuss these aspects in the next section and provide some best practice recommendations on the basis of the analyzed works.

Table 3 dissects the results and evaluation protocols of the works surveyed here. However, we have excluded works which do not involve a down-stream classification task, but rather describe pain behavior using computer vision (e.g., Rueß et al., 2019 and Statham et al., 2020), as well as pre-prints, and the three works which address only AU classification. The categories included in Table 3 are explained as follows.

- *Species:* As discussed in Sect. 4.1, it is important to consider the data collection protocols, for variations in conditions of lightning, angle of recording, etc. (e.g., for small laboratory animals such as mice, compared to larger animals such as sheep and horses). Also, differences across breeds, age, color and sex may be an important factor.
- *CrossVal:* the method used for cross-validation. We use the following abbreviations: single train-test-validation split (STTVS) (as opposed to k-fold cross-validation) and leave-one-animal-out (LOAO).
- *SubSep:* whether subject separation (subject exclusivity) was enforced in the splitting between train, test and validation sets.
- *SepVal:* whether the validation set was different than the test set. In general, the test set should be fully held-out, ideally until all the experiments are finished. Since this often is difficult to achieve because of data scarcity, it is good practice to base model selection on a validation set, to then evaluate the trained model on the held-out test set.
- *# cl:* number of classes used for classification. In emotion recognition, the classes correspond to the emotions studied (e.g., relaxed/curious in Corujo et al. (2021), or happy in Ferres et al. (2022); Franzoni et al. (2019)). In pain recognition, there is binary (pain/no pain) or three-level classification (e.g., pain level between 0-2 (Lu et al., 2017)). Thus, either the discrete or dimensional approaches (mentioned in Sect. 3) can be taken for dividing data into different classes within machine learning applications. The number of classes is important, as methods using different number and types of classes are often incomparable in terms of performance. In such cases, e.g., multi-class classifications (such as degree-based classification of pain, e.g., the ternary classification in Lencioni et al. (2021)), can be collapsed to binary classification (pain/no pain) to allow for a comparison.
- *Balancing:* data balancing method used. Data imbalance significantly affects performance metrics, thus using balancing methods in cases of greatly imbalanced datasets is important. We further elaborate on this point in the next section.
- *# frames:* The number of frames designates the number of unique frames used in training. We only report the number of samples pertaining to the affective state recognition task, and not e.g., the number of frames used to train a system for facial recognition (typically higher, e.g., Pessanha et al., 2020). For the approaches using video (Broomé et al., 2019, 2022; Rashid et al., 2022) we still report the number of unique frames here rather than number of unique clips for easier comparison.
- *# augm.:* designates the same quantity as **# frames**, when including the number of augmented samples. In Broomé et al. (2019), data augmentation was attempted, but the best performing approach did not use it. The reason for the estimate regarding the number of augmented sam-

ples for Feighelstein et al. (2022) is because the crops and rotations are made in a continuous manner from the images, and it is thus difficult to say exactly how many augmented frames are possible to obtain.
- *Metrics of accuracy, precision, recall and* F_1. Whenever confusion matrices are provided in the articles, we completed the computations for metrics for precision and recall, if not already given. Without published confusion matrices, we can only rely on the numbers in the articles. Hence, the measures that we could not obtain were simply left as blank (−) in the table.

It should be noted that the practice of subject-exclusive evaluation is important. By separating the subjects used for training, validation and testing respectively, generalization to unseen subjects is enforced, making sure that no specific features of an individual are used for classification. Subject-exclusive evaluation is trivially guaranteed in cases when there is one sample per subject, as is often the case with data scraped from the web, e.g., Franzoni et al. (2019) and Hummel et al. (2020). In datasets of this type, however, there is greater risk for bias and noise, introduced by data collected under unknown conditions. Datasets from private sources (Hummel et al., 2020; Corujo et al., 2021) or from the authors own clinical trials (Tuttle et al., 2018; Lencioni et al., 2021), on the other hand, typically involve a smaller number of individual animals, meaning that the risk is higher for the same animal with a similar expression to be present both in the training and testing set. In such cases the cross-validation method leave-one-animal-out is highly recommended, which also naturally enforces subject-exclusivity. The latter can also be exercised with other cross-validation methods, such as STTVS. We further elaborate on best practices in the context of cross-validation in the next section.

5 Best Practice Recommendations

Based on the landscape of current state-of-the-art works reviewed in this survey, and learning from best practice recommendations from other scientific communities, such as human affective computing, we provide below some technical recommendations for best practices in future research on automated recognition of animal affective states.

5.1 Data Imbalance

In the field of affective computing for animals, and specifically in animal pain recognition, data imbalance problems are particularly acute, due to the difficulty to obtain samples of, e.g., the 'pain' class, as opposed to the more available samples of the baseline. Moreover, pain behavior is a complex concept to learn, for humans and non-humans. This poses a difficulty for learning algorithms, which may collapse and only predict the majority class, when in fact the minority class may carry important and useful knowledge. Both for classic machine learning methods and deep learning methods, the most common remedy is data-driven, where the relevant classes typically are over- or under-sampled (Kulkarni et al., 2020; Buda et al., 2018).

The problem of learning from imbalanced data is well-studied in classical machine learning (Japkowicz & Stephen, 2002; Thabtah et al., 2020). A variety of methods to deal with data imbalance in this context have been proposed, see, e.g., Kulkarni et al. (2020) for a comprehensive list. One of the conclusions reached in Japkowicz and Stephen (2002) is that two essential factors which significantly impact performance are the degree of class imbalance and complexity of the concept to be learned.

In a deep learning context, data imbalance is often studied for datasets with a large number of classes and so called long-tail distributions of the minor classes (Huang et al., 2016; Li et al., 2020; Cui et al., 2019), which is typically less relevant for our setting. However, Buda et al. (2018) studies class imbalance in the context of CNNs on datasets with fewer classes, finding that oversampling does not necessarily contribute to overfitting.

In Buda et al. (2018), it is stated that the most common approach for deep methods is oversampling. Modifying the loss function is another option in a deep learning setting (Buda et al., 2018); this is commonly done for the above mentioned long-tail distribution scenarios. In random undersampling, instances from the negative class or majority class are selected at random, and removed until it matches the count of positive class or minority class, resulting in a balanced data set consisting of an equal number of positive and negative class examples. This method was used, e.g., in Feighelstein et al. (2022), addressing cat pain.

In the horse pain video dataset used in Broomé et al. (2019), there is slight class imbalance (pain is the minority class, by around 40%), when the sequences are extracted as back-to-back windows from the videos. No re-sampling is done in Broomé et al. (2019), but an unweighted F1 average across the two classes is used to present more fair results than accuracy (since this metric requires performance on both the positive and negative class). The same class imbalance is addressed in a follow-up work (Broomé et al., 2022), where video clips are over-sampled for the minority class. This is possible for video sequences, since one can easily adjust the stride of the extracted windows to obtain a larger number of sequences from the same video.

Another possibility is to use data augmentation, for instance by horizontally flipping images, adding noise to images, or randomly cropping images. In Pessanha et al. (2022), a 3D horse head model is used to synthesize 2D faces with different poses to augment the training set. However,

Table 3 Overview of the performance of the published works that do down-stream classification in either pain or emotions

	Species	CrossVal	SubSep	SepVal	# cl	Balancing	# frames	# augm.	Acc.	P	R	F1
Pain												
Tuttle et al. (2018)	Mice	STTVS	No	Yes	2	Oversampling	4577	–	93.2	93.7	93.3	93.5
Andresen et al. (2020)	Mice	10-fold	Yes	No	2	–	18,273	–	89.8	–	–	–
Pessanha et al. (2020)	Sheep	5-fold	No	No	2	–	86	–	78.0	83.0	68.0	73.0
Lu et al. (2017)	Sheep	10-fold	No	No	3	Random Undersampling	380	–	64.0	63.9	59.8	61.8
Lencioni et al. (2021)	Horses	10-fold	No	Yes	3	–	4850	–	75.8	76.2	75.8	76.0
Broomé et al. (2019)	Horses	LOAO	Yes	Yes	2	–	70,292	–	75.4	–	–	73.5
Broomé et al. (2022)	Horses	LOAO	Yes	Yes	2	Oversampling clips w/ half stride	70,292	140,584	–	–	–	58.2
Rashid et al. (2022)	Horses	LOAO	Yes	Yes	2	–	143,559	–	60.9	–	–	58.5
Feighelstein et al. (2022)	Cats	LOAO	Yes	Yes	2	Random Undersampling	464	>10,000	73.6	81.9	70.1	75.5
Emotion												
Corujo et al. (2021)	Horses	5-fold	No	Yes	4	Balanced	440	–	65.0	60.0	65.6	62.7
Franzoni et al. (2019)	Dogs	5-fold	No	No	3	–	231	–	95.3	93.3	93.1	93.1
Ferres et al. (2022)	Dogs	10-fold	No	Yes	4	Selected undersampling	400	–	67.5	68.4	67.5	67.9

such augmentations may change the distribution of the data, and should therefore be used with caution when only applied to one class, to avoid overfitting to an artificially augmented distribution.

Recommendation 1: Data imbalance should be minimized using relevant data balancing techniques, such as oversampling, undersampling or loss modifications.

5.2 Cross-Validation

One crucial observation arising from our survey is that works in this domain typically use highly dimensional datasets (being computer vision-based), which commonly have a small number of samples because of the intrinsic difficulties with data collection involving animal participants. The combination of high dimensionality with a small number of participants (possibly with few repeated samples per participant) has a higher potential of leading to bias in performance estimation.

As shown in Table 3, the cross-validation techniques used in the surveyed papers include single train and test split, k-fold cross-validation (with k = 5 or 10) and leave-one-animal-out methods. As previously mentioned, the latter means that the separation to training, validation and test sets is done on the basis of animal individuals, rather than on the basis of images or videos.

For deep methods, neural networks are typically trained throughout a number of epochs on a dataset, during which one can monitor the performance on a validation set after each epoch. One epoch means one round of training using all samples of the dataset once. This process allows you to choose the epoch where the model performs most optimally on the validation set. On the other hand, if the validation set is your final evaluation set, this amounts to adapting your model to your test set. Therefore, it is important to have a third split of the data—the test set, on which you can evaluate your model, which has not been part of the model selection process.

The training, validation and test splits can be constructed either randomly, or, ideally, in a subject-exclusive manner. In Broomé et al. (2019, 2022), Rashid et al. (2022) and Feighelstein et al. (2022), the presented results are averages of the test set results across a full test-subject rotation (each subject is used as test set once). Last, it can be mentioned that for deep learning methods, the random seed affects the initialization of the networks, when trained from scratch. If training has been carried out with different random seeds, it is important to present results that are averages of repeated such runs, to avoid cherry picking a particularly opportune training instance. Therefore, Broomé et al. (2019) additionally repeats each split five times, and the presented result is the average of five runs times the number of subjects.

For non-deep methods, Varma and Simon (2006) studied validation techniques suggesting that Nested Cross-Validation (Koskimäki, 2015) has minimally biased performance estimates. Vabalas et al. (2019) also recommended this method to be used with datasets with a sample size of up to 1000, to reduce strongly biased performance. In this method, a portion of the data is split at the beginning and in

each cross-validation fold, a model is then developed on the reduced training set from scratch, including feature selection and parameter tuning. This is repeated with splitting a different portion of the data for validation, each time developing a new model for training until all the data is used. Koskimäki (2015) showed that to obtain more confidence on the results, models should be trained and evaluated applying at least using nested or single 10-fold cross-validation or by using double or simple LOAO cross-validation.

In double (or nested) leave-one-person-out cross-validation, bias is avoided by adding an outer loop into cross validation. In a simple leave-one-person-out cross-validation the validation is set randomly. Data from one person at a time is chosen as separate testing data while the data from the remaining N-1 subjects is left for basic leave-one-person-out cross-validation. This approach is, however, the most computationally challenging. Nested 10-fold cross-validation or double leave-one-person-out methods are recommended to reduce biased performance when feature selection or parameter tuning is performed during cross-validation (Table 4).

Therefore, in small datasets (number of samples lower than 1000), which have almost no repetition of subjects, 10-fold cross-validation is recommended to reduce biased performance whenever neither feature selection nor parameter tuning is performed during cross-validation. Otherwise, nested 10-fold cross-validation is recommended. For relatively small datasets with numerous repeated samples of same animal subject, the leave-one-animal-out cross-validation technique is recommended to reduce biased performance. Otherwise, double leave-one-animal-out cross-validation is recommended.

Recommendation 2: To reduce biased performance evaluation, for classical machine learning methods, the choice of cross-validation is recommended according to Table 4, when the dataset is small with repeated samples of the same animal subject. For deep methods, it is recommended to use a fully held-out test set, which ideally is subject-exclusive. It is furthermore recommended to present results from repeated runs on more than one random seed.

5.3 Domain Transfer

The variety of species, affective states and environment conditions lends itself to exploration of cross-database transfer methodologies (Li & Deng, 2022a), i.e., training a model on an original, source dataset and subsequently use this instance to classify samples from a target dataset, presenting some degree of domain shift.

One possible setting for domain transfer is cross-species. Hummel et al. (2020), studies domain transfer from horse-based models to donkeys, reporting a loss of accuracy in automatic pose estimation, landmark detection, and subsequent pain prediction. A further example of domain transfer is cross-state: to transfer between different types of emotions or types of pain. In the study of Broomé et al. (2019), it was shown that a model trained only on a dataset of horses with acute experimental pain can aid recognition of the subtler displays of orthopedic pain. This is useful because training is shown to be difficult on the subtler type of pain expression. A third example of a transfer scenario is cross-environment, or simply cross-domain. Mahmoud et al. (2018) train their model for pain estimation in sheep on a dataset collected on a farm, and then present results transferred to a dataset of sheep collected from the internet.

Learning from cross-database transfer in human facial analysis, one crucial issue is the differences in intrinsic bias of the source and target datasets, related not only to the facial expressions, but also to important factors such as occlusion, illumination, and background, as well as factors related to annotation and balance, which may have significant impact. Li and Deng (2022a) demonstrate such differences in the human domain and propose methods to minimize these types of biases. For animals, these differences are expected to play an even greater role, given the large domain variety, as discussed in Sect. 4.1.

Recommendation 3: Methods to minimize intrinsic dataset bias are recommended for cross-domain transfer studies.

6 Conclusions and Future Work

Although the field of automated recognition of affective states in animals is only beginning to emerge, the breadth and variability of the approaches covered in our survey makes this a timely moment for reflection on challenges faced by the community and steps that can be taken to advance the field.

One crucial issue that needs to be highlighted is the difficulty in comparing the different works. Despite some commonalities in the stage of data analysis (features, models and pipelines), the variety of species, and the ways data are collected and annotated differ tremendously, as discussed in Sect. 4.1. Thus, e.g., the 99% accuracy achieved in Andresen et al. (2020) for pain recognition in laboratory mice, in a small box with controlled lighting and good coverage of the animal's face, cannot be straightforwardly compared to the estimation of pain level with accuracy of 67% achieved in Mahmoud et al. (2018) for sheep using footage obtained in the naturalistic (and much less controlled) setting of a farm.

Drawing inspiration from the huge amount of benchmark datasets existing in the human domain (such as the Cohn-Kanade dataset (Lucey et al., 2010), the Toronto face database (Susskind et al., 2008), and more), the development of benchmarking resources for animals—both species-specific and across-species—can help systematize the field and promote

Table 4 Recommendations for best practices in cross-validation for classical machine learning methods

Repeated samples per subject?	Feature selection or parameter fine-tuning used?	
	No	Yes
No	10-fold	Nested 10-fold
Yes	LOAO	Double LOAO

comparison between approaches. However, this is more challenging than in humans, due to the large variety of species, and environments (laboratory, zoo, home, farm, in the wild, etc.) in which they can be recorded. Another barrier is considerations of ethics and privacy, especially when producing datasets with induced emotions and pain (as explained in Sect. 4.1). Unfortunately, this often makes it difficult to make datasets publicly accessible. Thus, there is a strong need for public datasets in this domain.

Another issue that should be addressed in future research efforts is explainability, which is particularly important for applied contexts related to clinical decision making and animal welfare management. Consistently with the human affective computing literature, our review reveals the tendency towards 'black-box' approaches which use learned features. While using hand-crafted features is indeed less flexible than learning them from data, and may in some cases lead to lower performance, their clear advantage is explainability, having more control over the information extracted from a dataset. Learned features, on the other hand, tend to be more opaque, leading to 'black-box' reasoning, which does not borrow itself easily for explaining the classification decisions in human-understandable terms (see, e.g., London, 2019). It is possible to investigate statistical properties of the various dimensions of the feature maps, and study what type of stimuli specific neurons activate maximally for, but these properties are still not conclusive in terms of how features are organized and what they represent. In neural networks, the features are often *entangled*, which complicates the picture more. This is when a given network unit activates for a mix of input signals (e.g., the face of an animal in a certain pose with a certain facial expression and background), but perhaps not for the separate parts of that signal (e.g., the face of the same animal in a different pose, with the same facial expression, but with a different background). There have been *disentanglement* efforts, predominantly unsupervised (Higgins et al., 2017; Kumar et al., 2018; Kim & Mnih, 2018), within deep learning to reduce these tendencies, since this is, in general, not a desirable feature for a machine learning system. In terms of animal affect applications, the pain recognition approach of Rashid et al. (2022), includes such self-supervised disentanglement in one part of their modeling pipeline. However, much development remains before a neural network can stably display control and separation of different factors of variation. This characteristic of neural networks poses difficulty for research that aims to perform exploratory analysis of animal behavior. In particular, it poses high demands on the quality of data and labels, in order to avoid reliance on spurious correlations.

In summary, in the last five years we are witnessing an impressive growth in the number of studies addressing recognition of affective states in animals. Notably, many of these works are carried out by multi-disciplinary teams, demonstrating the intellectual value of collaboration between biologists, veterinary scientists and computer scientists, as well as the increasing importance of computer vision techniques within animal welfare and veterinary science. We believe in the importance of knowledge exchange between different disciplines since having a common understanding of each other's research approaches and respective needs is essential for progress. Efforts invested in pushing the field of animal affective computing forward will not only lead to new technologies promoting animal welfare and well-being, but will also hopefully provide new insights for the long-standing philosophical and ethical debates on animal sentience.

Acknowledgements The research was partially supported by the grant from the Ministry of Science and Technology of Israel according to the research Project No. 19-57-06007. The second author was additionally supported by the Data Science Research Center (DSRC), University of Haifa. The authors would like to thank Ilan Shimshoni and Shir Amir for their scientific consultations.

Funding Open access funding provided by Royal Institute of Technology.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S.,

Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., & Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org. https://www.tensorflow.org/

Al-Eidan, R. M., Al-Khalifa, H. S., & Al-Salman, A. S. (2020). Deep-learning-based models for pain recognition: A systematic review. *Applied Sciences, 10,* 5984.

Amir, S., Gandelsman, Y., Bagon, S., & Dekel, T. (2021). *Deep ViT features as dense visual descriptors*. arXiv preprint arXiv:2112.05814.

Amir, S., Zamansky, A., & van der Linden, D. (2017). K9-blyzer-towards video-based automatic analysis of canine behavior. In *Proceedings of Animal–Computer Interaction 2017*.

Anand, K. J., Stevens, B. J., McGrath, P. J., et al. (2007). *Pain in neonates and infants: Pain research and clinical management series* (Vol. 10). Philedelphia: Elsevier Health Sciences.

Andersen, P. H., Broomé, S., Rashid, M., Lundblad, J., Ask, K., Li, Z., et al. (2021). Towards machine recognition of facial expressions of pain in horses. *Animals, 11*(6), 1643.

Anderson, D. J., & Adolphs, R. (2014). A framework for studying emotions across species. *Cell, 157*(1), 187–200.

Anderson, D. J., & Perona, P. (2014). Toward a science of computational ethology. *Neuron, 84*(1), 18–31.

Andresen, N., Wöllhaf, M., Hohlbaum, K., Lewejohann, L., Hellwich, O., Thöne-Reineke, C., & Belik, V. (2020). Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *PLoS ONE, 15*(4), 0228059.

Ask, K., Rhodin, M., Tamminen, L.-M., Hernlund, E., & Andersen, P. H. (2020). Identification of body behaviors and facial expressions associated with induced orthopedic pain in four equine pain scales. *Animals: An Open Access Journal from MDPI, 10,* 2155.

Auer, U., Kelemen, Z., Engl, V., & Jenner, F. (2021). Activity time budgets—A potential tool to monitor equine welfare? *Animals, 11*(3), 850.

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology, 87*(2), 266–281.

Bartlett, M. S., Littlewort, G. C., Frank, M. G., & Lee, K. (2014). Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology, 24,* 738–743.

Bateson, M., & Martin, P. (2021). *Measuring behaviour: An introductory guide*. New York: Cambridge University Press.

Biggs, B., Boyne, O., Charles, J., Fitzgibbon, A., & Cipolla, R. (2020). Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision* (pp. 195–211). Springer.

Birch, J., Burn, C., Schnell, A., Browning, H., & Crump, A. (2021). Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans.

Blumrosen, G., Hawellek, D., & Pesaran, B. (2017). Towards automated recognition of facial expressions in animal models. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 2810–2819).

Boissy, A., Arnould, C., Chaillou, E., Désiré, L., Duvaux-Ponter, C., Greiveldinger, L., et al. (2007). Emotions and cognition: A new approach to animal welfare. *Animal Welfare, 16*(2), 37–43.

Boneh-Shitrit, T., Amir, S., Bremhorst, A., Riemer, S., Wurbel, H., Mills, D., & Zamansky, A. (2022). Deep learning models for classification of canine emotional states. Submitted.

Bremhorst, A., Sutter, N. A., Würbel, H., Mills, D. S., & Riemer, S. (2019). Differences in facial expressions during positive anticipation and frustration in dogs awaiting a reward. *Scientific Reports, 9*(1), 1–13.

Briefer, E. F., Tettamanti, F., & McElligott, A. G. (2015). Emotions in goats: Mapping physiological, behavioural and vocal profiles. *Animal Behaviour, 99,* 131–143.

Broomé, S., Ask, K., Rashid-Engström, M., Andersen, P. H., & Kjellström, H. (2022). Sharing pain: Using pain domain transfer for video recognition of low grade orthopedic pain in horses. *PLoS ONE, 17,* e0263854.

Broomé, S., Gleerup, K. B., Andersen, P. H., & Kjellstrom, H. (2019). Dynamics are important for the recognition of equine pain in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12667–12676).

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks: The Official Journal of the International Neural Network Society, 106,* 249–259.

Caeiro, C. C., Burrows, A. M., & Waller, B. M. (2017). Development and application of catfacs: Are human cat adopters influenced by cat facial expressions? *Applied Animal Behaviour Science*.

Camras, L. A., & Shutter, J. M. (2010). Emotional facial expressions in infancy. *Emotion Review, 2*(2), 120–129.

Cao, J., Tang, H., Fang, H.-S., Shen, X., Lu, C., & Tai, Y.-W. (2019). Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9498–9507).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4724–4733).

Clegg, I. L., Rödel, H. G., & Delfour, F. (2017). Bottlenose dolphins engaging in more social affiliative behaviour judge ambiguous cues more optimistically. *Behavioural Brain Research, 322,* 115–122.

Correia-Caeiro, C., Holmes, K., & Miyabe-Nishiwaki, T. (2021). Extending the MaqFACS to measure facial movement in Japanese macaques (*Macaca fuscata*) reveals a wide repertoire potential. *PLoS ONE, 16*(1), 0245117.

Corujo, L. A., Kieson, E., Schloesser, T., & Gloor, P. A. (2021). Emotion recognition in horses with convolutional neural networks. *Future Internet, 13*(10), 250.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. J. (2019). Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9260–9269).

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). The body action and posture coding system (BAP): Development and reliability. *Journal of Nonverbal Behavior, 36*(2), 97–121.

Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E., & Leach, M. C. (2014). Development of the horse grimace scale (hgs) as a pain assessment tool in horses undergoing routine castration. *PLoS ONE, 9*(3), 92281.

Dawkins, M. S. (2008). The science of animal suffering. *Ethology, 114*(10), 937–945.

de Vere, A. J., & Kuczaj, S. A. (2016). Where are we in the study of animal emotions? *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(5), 354–362.

Descovich, K. A., Wathan, J., Leach, M. C., Buchanan-Smith, H. M., Flecknell, P., Framingham, D., & Vick, S.-J. (2017). *Facial expression: An under-utilised tool for the assessment of welfare in mammals*. Altex.

Diogo, R., Abdala, V., Lonergan, N., & Wood, B. (2008). From fish to modern humans-comparative anatomy, homologies and evolution of the head and neck musculature. *Journal of Anatomy, 213*(4), 391–424.

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(10), 974–989.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Duhn, L. J., & Medves, J. M. (2004). A systematic integrative review of infant pain assessment tools. *Advances in Neonatal Care, 4*(3), 126–140.

Duncan, I. J. (1996). Animal welfare defined in terms of feelings. *Acta Agriculturae Scandinavica. Section A. Animal Science*. Supplementum (Denmark).

Dyson, S., Berger, J., Ellis, A. D., & Mullard, J. (2018). Development of an ethogram for a pain scoring system in ridden horses and its application to determine the presence of musculoskeletal pain. *Journal of Veterinary Behavior, 23,* 47–57.

Ede, T., Lecorps, B., von Keyserlingk, M. A., & Weary, D. M. (2019). Symposium review: Scientific assessment of affective states in dairy cattle. *Journal of Dairy Science, 102*(11), 10677–10694.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3–4), 169–200.

Ekman, P., & Friesen, W. (1978). Facial action coding system: A technique for the measurement of facial movement. *Environmental Psychology & Nonverbal Behavior*.

Feighelstein, M., Shimshoni, I., Finka, L., Luna, S. P., Mills, D., & Zamansky, A. (2022). Automated recognition of pain in cats. Submitted.

Ferres, K., Schloesser, T., & Gloor, P. A. (2022). Predicting dog emotions based on posture analysis using DeepLabCut. *Future Internet, 14*(4), 97.

Finka, L. R., Luna, S. P., Brondani, J. T., Tzimiropoulos, Y., McDonagh, J., Farnworth, M. J., et al. (2019). Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. *Scientific Reports, 9*(1), 1–12.

Fitzgerald, M., & McIntosh, N. (1989). Pain and analgesia in the newborn. *Archives of Disease in Childhood, 64,* 441–443.

Foris, B., Zebunke, M., Langbein, J., & Melzer, N. (2019). Comprehensive analysis of affiliative and agonistic social networks in lactating dairy cattle groups. *Applied Animal Behaviour Science, 210,* 60–67.

Forkosh, O. (2021). Animal behavior and animal personality from a non-human perspective: Getting help from the machine. *Patterns, 2*(3), 100194.

Franzoni, V., Milani, A., Biondi, G., & Micheli, F. (2019). A preliminary work on dog emotion recognition. In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume* (pp. 91–96).

Gleerup, K. B., Forkman, B., Lindegaard, C., & Andersen, P. H. (2015). An equine pain face. *Veterinary Anaesthesia and Analgesia, 42*(1), 103–114.

Gleerup, K., & Lindegaard, C. (2016). Recognition and quantification of pain in horses: A tutorial review. *Equine Veterinary Education, 28*(1), 47–57.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., & Couzin, I. D. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife, 8,* 47994.

Grunau, R. V., & Craig, K. D. (1987). Pain expression in neonates: Facial action and cry. *Pain, 28*(3), 395–410.

Hale, C. J., & Hadjistavropoulos, T. (1997). Emotional components of pain. *Pain Research and Management, 2*(4), 217–225.

Hall, C., Randle, H., Pearson, G., Preshaw, L., & Waran, N. (2018). Assessing equine emotional state. *Applied Animal Behaviour Science, 205,* 183–193.

Hassan, T., Seuss, D., Wollenberg, J., Weitz, K., Kunz, M., Lautenbacher, S., et al. (2021). Automatic detection of pain from facial expressions: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43,* 1815–1831.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 24–26, 2017, Conference Track Proceedings. https://openreview.net/forum?id=Sy2fzU9gl

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5375–5384).

Huber, A., Dael, N., Caeiro, C., Würbel, H., Mills, D., & Riemer, S. (2018). From BAP to DogBAP-adapting a human body movement coding system for use in dogs. *Measuring Behavior*.

Hummel, H. I., Pessanha, F., Salah, A. A., van Loon, T. J., & Veltkamp, R. C. (2020). Automatic pain detection on horse and donkey faces. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 793–800). IEEE.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The kinetics human action video dataset*. CoRR arXiv:1705.06950

Kim, H., & Mnih, A. (2018). Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Koskimäki, H. (2015). Avoiding bias in classification accuracy—A case study for activity recognition. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 301–306). https://doi.org/10.1109/SSCI.2015.52

Kremer, L., Holkenborg, S. K., Reimert, I., Bolhuis, J., & Webb, L. (2020). The nuts and bolts of animal emotion. *Neuroscience & Biobehavioral Reviews, 113,* 273–286.

Kret, M. E., Massen, J. J., & de Waal, F. (2022). My fear is not, and never will be, your fear: On emotions and feelings in animals. *Affective Science, 3,* 182–189.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In F. A. Batarseh & R. Yang (Eds.), *Data democracy* (pp. 83–106). Academic Press. https://doi.org/10.1016/B978-0-12-818366-3.00005-8.

Kumar, A., Sattigeri, P., & Balakrishnan, A. (2018). Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings. https://openreview.net/forum?id=H1kG7GZAW

Labus, J. S., Keefe, F. J., & Jensen, M. P. (2003). Self-reports of pain intensity and direct observations of pain behavior: When are they correlated? *Pain, 102*(1–2), 109–124.

Lansade, L., Nowak, R., Lainé, A.-L., Leterrier, C., Bonneau, C., Parias, C., & Bertin, A. (2018). Facial expression and oxytocin as possible markers of positive emotions in horses. *Scientific Reports, 8*(1), 1–11.

Lecorps, B., Rödel, H. G., & Féron, C. (2016). Assessment of anxiety in open field and elevated plus maze using infrared thermography. *Physiology & Behavior, 157,* 209–216.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.

Lencioni, G. C., de Sousa, R. V., de Souza Sardinha, E. J., Corrêa, R. R., & Zanella, A. J. (2021). Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. *PLoS ONE, 16*(10), 0258672.

Li, S., & Deng, W. (2022a). A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing, 13*(2), 881–893. https://doi.org/10.1109/TAFFC.2020.2973158.

Li, S., & Deng, W. (2022b). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215.

Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., & Feng, J. (2020). Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10988–10997).

Li, Z., Broomé, S., Andersen, P.H., & Kjellström, H. (2021). *Automated detection of equine facial action units*. arXiv preprint arXiv:2102.08983

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO and: Common objects in context. In *ECCV*.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21.

Low, P., Panksepp, J., Reiss, D., Edelman, D., Swinderen, B.V., Low, P., & Koch, C. (2012). The Cambridge declaration on consciousness. In *Francis Crick Memorial conference on consciousness in human and non-human animals*. Cambridge. Retrieved from https://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf

Lu, Y., Mahmoud, M., & Robinson, P. (2017). Estimating sheep pain level using facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 394–399). IEEE.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 94–101). IEEE.

Lundblad, J., Rashid, M., Rhodin, M., & Haubro Andersen, P. (2021). Effect of transportation and social isolation on facial expressions of healthy horses. *PLoS ONE, 16*(6), 0241532.

Mahmoud, M., Lu, Y., Hou, X., McLennan, K., & Robinson, P. (2018). Estimation of pain in sheep using computer vision. In R. J. Moore (Ed.), *Handbook of pain and palliative care* (pp. 145–157). Springer.

Maisonpierre, I., Sutton, M., Harris, P., Menzies-Gow, N., Weller, R., & Pfau, T. (2019). Accelerometer activity tracking in horses and the effect of pasture management on time budget. *Equine Veterinary Journal, 51*(6), 840–845.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience, 21*(9), 1281.

Mayo, L. M., & Heilig, M. (2019). In the face of stress: Interpreting individual differences in stress-induced facial expressions. *Neurobiology of Stress, 10,* 100166.

McLennan, K., & Mahmoud, M. (2019). Development of an automated pain facial expression detection system for sheep (*Ovis aries*). *Animals, 9*(4), 196.

McLennan, K. M., Rebelo, C. J., Corke, M. J., Holmes, M. A., Leach, M. C., & Constantino-Casas, F. (2016). Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Applied Animal Behaviour Science, 176,* 19–26.

Mendl, M., Burman, O. H., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences, 277*(1696), 2895–2904.

Mendl, M., & Paul, E. S. (2020). Animal affect and decision-making. *Neuroscience & Biobehavioral Reviews, 112,* 144–163.

Merkies, K., Ready, C., Farkas, L., & Hodder, A. (2019). Eye blink rates and eyelid twitches as a non-invasive measure of stress in the domestic horse. *Animals, 9*(8), 562.

Morozov, A., Parr, L. A., Gothard, K. M., Paz, R., & Pryluk, R. (2021). Automatic recognition of macaque facial expressions for detection of affective states. *eNeuro, 8,* ENEURO-0117.

Morris, P. H., Doe, C., & Godsell, E. (2008). Secondary emotions in non-primate species? Behavioural reports and subjective claims by animal owners. *Cognition and Emotion, 22*(1), 3–20.

Mott, R. O., Hawthorne, S. J., & McBride, S. D. (2020). Blink rate as a measure of stress and attention in the domestic horse (*Equus caballus*). *Scientific Reports, 10*(1), 1–8.

Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition, 71,* 158–172.

Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing, 12*(2), 505–523.

Oliveira, T., Santos, A., Silva, J., Trindade, P., Yamada, A., Jaramillo, F., et al. (2022). Hospitalisation and disease severity alter the resting pattern of horses. *Journal of Equine Veterinary Science, 110,* 103832.

Panksepp, J. (2010). *Emotional causes and consequences of social-affective vocalization*. Handbook of Behavioral Neuroscience: Elsevier.

Parr, L. A., Waller, B. M., Burrows, A. M., Gothard, K. M., & Vick, S.-J. (2010). Brief communication: MaqFACS: A muscle-based facial movement coding system for the rhesus macaque. *American Journal of Physical Anthropology, 143*(4), 625–630.

Paul, E. S., Harding, E. J., & Mendl, M. (2005). Measuring emotional processes in animals: The utility of a cognitive approach. *Neuroscience & Biobehavioral Reviews, 29*(3), 469–491.

Paul, E. S., & Mendl, M. T. (2018). Animal emotion: Descriptive and prescriptive definitions and their implications for a comparative perspective. *Applied Animal Behaviour Science, 205,* 202–209.

Pennington, Z. T., Dong, Z., Feng, Y., Vetere, L. M., Page-Harley, L., Shuman, T., & Cai, D. J. (2019). ezTrack: An open-source video analysis pipeline for the investigation of animal behavior. *Scientific Reports, 9*(1), 1–11.

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S.S.-H., Murthy, M., & Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. *Nature Methods, 16*(1), 117–125.

Pessanha, F., McLennan, K., & Mahmoud, M. (2020). Towards automatic monitoring of disease progression in sheep: A hierarchical model for sheep facial expressions analysis from video. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG) (pp. 670–676). IEEE Computer Society.

Pessanha, F., Salah, A. A., van Loon, T., & Veltkamp, R. (2022). Facial image-based automatic assessment of equine pain. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2022.3177639.

Plutchik, R. (1979). *Emotion, a psychoevolutionary synthesis*. New York: Harper & Row.

Podturkin, A. A., Krebs, B. L., & Watters, J. V. (2022). A quantitative approach for using anticipatory behavior as a graded welfare assessment. *Journal of Applied Animal Welfare Science*, 1–15.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology, 17*(3), 715–734.

Price, J., Marques, J., Welsh, E., & Waran, N. (2002). Pilot epidemiological study of attitudes towards pain in horses. *Veterinary Record, 151*(19), 570–575.

Proctor, H. S., & Carder, G. (2015). Measuring positive emotions in cows: Do visible eye whites tell us anything? *Physiology & Behavior, 147,* 1–6.

Qiu, Y., & Wan, Y. (2019). Facial expression recognition based on landmarks. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* (Vol. 1, pp. 1356–1360). https://doi.org/10.1109/IAEAC47372.2019.8997580

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., et al. (2020). The revised IASP definition of pain: Concepts, challenges, and compromises. *Pain, 161*(9), 1976.

Rashid, M., Broomé, S., Ask, K., Hernlund, E., Andersen, P. H., Kjellström, H., & Lee, Y. J. (2022). Equine pain behavior classification via self-supervised disentangled pose representation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1646–1656).

Reulke, R., Rueß, D., Deckers, N., Barnewitz, D., Wieckert, A., & Kienapfel, K. (2018). Analysis of motion patterns for pain estimation of horses. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6). IEEE.

Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J., & de Polavieja, G. G. (2019). Idtracker.ai: Tracking all individuals in small or large collectives of unmarked animals. *Nature Methods, 16*(2), 179–182.

Rueß, D., Rueß, J., Hümmer, C., Deckers, N., Migal, V., Kienapfel, K., Wieckert, A., Barnewitz, D., & Reulke, R. (2019). Equine welfare assessment: Horse motion evaluation and comparison to manual pain measurements. In *Pacific-Rim Symposium on Image and Video Technology* (pp. 156–169). Springer.

Schnaider, M., Heidemann, M., Silva, A., Taconeli, C., & Molento, C. (2022). Vocalization and other behaviors as indicators of emotional valence: The case of cow-calf separation and reunion in beef cattle. *Journal of Veterinary Behavior, 49,* 28–35.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 618–626).

Sénèque, E., Lesimple, C., Morisset, S., & Hausberger, M. (2019). Could posture reflect welfare state? A study using geometric morphometrics in riding school horses. *PLoS ONE, 14*(2), 0211852.

Seuss, D., Dieckmann, A., Hassan, T., Garbas, J.-U., Ellgring, J. H., Mortillaro, M., & Scherer, K. (2019). Emotion expression from different angles: A video database for facial expressions of actors shot by a camera array. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 35–41). IEEE.

Sharma, G., & Dhall, A. (2021). A survey on automatic multimodal emotion recognition in the wild. In G. Phillips-Wren, A. Esposito, & L. C. Jain (Eds.), *Advances in data science: Methodologies and applications* (pp. 35–64). Springer. https://doi.org/10.1007/978-3-030-51870-7_3.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. CoRR arXiv:1409.1556

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014). Defining and assessing animal pain. *Animal Behaviour, 97,* 201–212. https://doi.org/10.1016/j.anbehav.2014.09.007.

Soomro, K., Zamir, A. R., & Shah, M. (2012). *Ucf101: A dataset of 101 human actions classes from videos in the wild*. CoRR arXiv:1212.0402

Statham, P., Hannuna, S., Jones, S., Campbell, N., Robert Colborne, G., Browne, W. J., et al. (2020). Quantifying defence cascade responses as indicators of pig affect and welfare using computer vision methods. *Scientific Reports, 10*(1), 1–13.

Susskind, J. M., Hinton, G. E., Movellan, J. R., & Anderson, A. K. (2008). Generating facial expressions with deep belief nets. In *Affective computing, emotion modelling, synthesis and recognition* (pp. 421–440).

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences, 513,* 429–441.

Tuttle, A. H., Molinaro, M. J., Jethwa, J. F., Sotocinal, S. G., Prieto, J. C., Styner, M. A., et al. (2018). A deep neural network to assess spontaneous pain from mouse facial expressions. *Molecular Pain, 14,* 1744806918763658.

Uccheddu, S., Ronconi, L., Albertini, M., Coren, S., Da Graça Pereira, G., De Cataldo, L., et al. (2022). Domestic dogs (*Canis familiaris*) grieve over the loss of a conspecific. *Scientific Reports, 12*(1), 1–9.

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE, 14*(11), 0224365.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*(1), 1–8.

Waller, B., Caeiro, C., Peirce, K., Burrows, A., Kaminski, J., et al. (2013). *DogFACS: The dog facial action coding system. Manual*. University of Portsmouth.

Waller, B. M., Peirce, K., Caeiro, C., Scheider, L., Burrows, A. M., McCune, S., & Kaminski, J. (2013). Paedomorphic facial expressions give dogs a selective advantage. *PLoS ONE, 8,* e82686.

Walsh, J., Eccleston, C., & Keogh, E. (2014). Pain communication through body posture: The development and validation of a stimulus set. *PAIN®*, *155*(11), 2282–2290.

Wang, N., Gao, X., Tao, D., Yang, H., & Li, X. (2018). Facial feature point detection: A comprehensive survey. *Neurocomputing, 275,* 50–65.

Waran, N., Williams, V., Clarke, N., & Bridge, I. (2010). Recognition of pain and use of analgesia in horses by veterinarians in New Zealand. *New Zealand Veterinary Journal, 58*(6), 274–280.

Wathan, J., Burrows, A. M., Waller, B. M., & McComb, K. (2015). Equifacs: The equine facial action coding system. *PLoS ONE, 10*(8), 0131738.

Wu, Y., & Ji, Q. (2019). Facial landmark detection: A literature survey. *International Journal of Computer Vision, 127*(2), 115–142.

Zamzmi, G., Kasturi, R., Goldgof, D., Zhi, R., Ashmeade, T., & Sun, Y. (2017). A review of automated pain assessment in infants: Features, classification tasks, and databases. *IEEE Reviews in Biomedical Engineering, 11,* 77–96.

Zhu, H., Salgırlı, Y., Can, P., Atılgan, D., & Salah, A. A. (2022). *Video-based estimation of pain indicators in dogs*. arXiv preprint arXiv:2209.13296