Equine Pain Behavior Classification via Self-Supervised Disentangled Pose Representation

Maheen Rashid[1,2] Sofia Broomé[3] Katrina Ask[4] Elin Hernlund[4] Pia Haubro Andersen[4]
Hedvig Kjellström[3,5] Yong Jae Lee[1,6]
[1] UC Davis, USA [2] Univrses AB, Sweden maheen.rashid@univrses.com
[3] KTH Royal Institute of Technology, Sweden sbroome, hedvig@kth.se
[4] SLU, Sweden katrina.ask, elin.hernlund, pia.haubro.andersen@slu.se
[5] Silo AI, Sweden [6] UW Madison, USA yongjaelee@cs.wisc.edu

Abstract

Timely detection of horse pain is important for equine welfare. Horses express pain through their facial and body behavior, but may hide signs of pain from unfamiliar human observers. In addition, collecting visual data with detailed annotation of horse behavior and pain state is both cumbersome and not scalable. Consequently, a pragmatic equine pain classification system would use video of the unobserved *horse and weak labels. This paper proposes such a method for equine pain classification by using multi-view surveillance video footage of unobserved horses with induced orthopaedic pain, with temporally sparse video level pain labels. To ensure that pain is learned from horse body language alone, we first train a self-supervised generative model to disentangle horse pose from its appearance and background before using the disentangled horse pose latent representation for pain classification. To make best use of the pain labels, we develop a novel loss that formulates pain classification as a multi-instance learning problem. Our method achieves pain classification accuracy better than human expert performance with 60% accuracy. The learned latent horse pose representation is shown to be viewpoint covariant, and disentangled from horse appearance. Qualitative analysis of pain classified segments shows correspondence between the pain symptoms identified by our model, and equine pain scales used in veterinary practice.*

1. Introduction

Timely detection of pain in horses is important for their welfare, and for the early diagnosis and treatment of underlying disease. While horses express pain through changes in facial and body behavior, equine pain classification is a challenging problem, with expert human performance on video data at just 58% accuracy for pain or no pain classification for acute pain [4], and 51% accuracy for orthopaedic pain [3]. While self-evaluation can be used as a gold standard for determining pain in human subjects, horses, being non-verbal, lack a gold standard for pain [1]. In addition, as prey animals, horses may hide signs of pain from human observers [11]. It is therefore difficult to ascertain if a horse is experiencing and expressing pain.

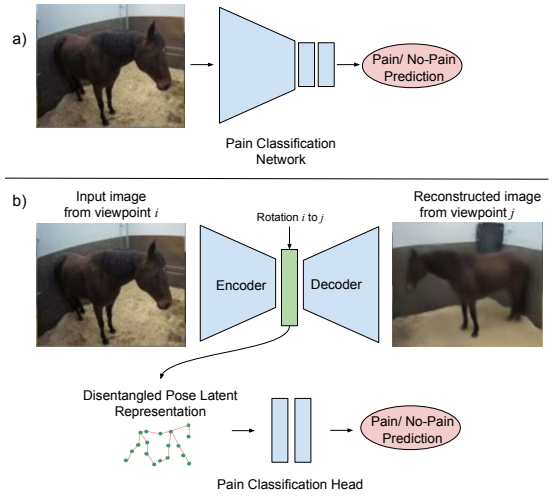

Figure 1. **Main Idea** (a) Directly training a pain classifier on video frames can result in a model that is not interpretable, and would overfit to limited training data. (b) Our method first uses self-supervised multi-view synthesis to learn a latent horse pose representation, that is then used to learn a light weight pain classifier.

Consequently, even though pain scales relying on facial and body behaviors are used in veterinary practice [42, 40, 51, 18, 16], determining the pain level of horses remains challenging. Furthermore, outside of an animal hospital or clinic, horse owners may underestimate the prevalence of disease, and thereby impede contact with the veterinarian [27]. A computer vision system capable of determining

horse pain, therefore, has great potential for improving animal welfare by enabling round the clock classification of pain and consequent diagnosis of illness.

Human pain datasets with clear closeup facial footage, and detailed Facial Action Coding System (FACS) [14] annotation have been used for training human pain classification systems [33]. However, a similar approach is not practical for equine pain classification. First, obtaining similar annotation is also costly and time consuming for horses as it is for humans. Equine FACS [64] annotators must be trained and certified, and spend upwards of 30 minutes to annotate a single minute of video [43]. Second, with pain scales including entries like 'interactive behavior' there is not always a mapping between pain attribute and obvious visual behavior [16] making detailed annotation of video an ambiguous task. Third, horses behave differently in the presence of humans [11, 58], which calls to question the applicability of datasets with observed horses to more natural settings when the horse is alone. Relatedly, a vision based monitoring system for horses would be more pragmatic if it could operate off unobtrusive surveillance cameras, rather than require horses to be close to the camera, with the face clearly visible, as is true for human datasets.

In 2018-2019, a dataset of horses with induced orthopaedic pain was collected at Swedish University of Agricultural Sciences [2]. It includes multi-view surveillance video footage of unobserved horses with pain labels from periodic pain assessments by expert veterinary researchers. Since the dataset contains few subjects (i.e. 8), and lacks temporally detailed annotation, a fully and strongly supervised network is likely to overfit to the training data. At the same time, the network predictions would not be interpretable, and may use superfluous but correlated information, such as the lighting to determine the pain state of the horse.

On the other hand, self-supervised methods have been shown to disentangle semantically and visually meaningful properties from image data without the use of labels. Examples include disentangling the 3D normals, albedo, lighting, and alpha matte for faces [53], and disentangling pose and identity for facial recognition [59].

Our **key idea** is to use self-supervision to disentangle the visual properties a pain classification system should focus on, and then use the disentangled representation to identify pain. In this manner, we can reduce the likelihood of the model learning extraneous information to determine pain, and prevent overfitting to the training data (Figure 1). Additionally we use the observation that a painful horse may also exhibit non-painful behavior to formulate a novel loss that makes best use of the sparse pain annotation.

We use a two step process for pain classification. In the first stage, we train a view synthesis model that, given an image of a horse from one angle learns to synthesize the scene from a different viewpoint [48]. We use an encoder-decoder architecture, and disentangle the horse pose, appearance, and background in the process. In the second stage, we use the learned pose representation to classify video segments as painful. As we lack detailed temporal annotation for pain, we use weak supervision to train the pain classification module, and propose a modified multiple instance learning loss towards this end. Our system is able to learn a viewpoint aware latent pose representation, and determine the pain label of video segments with 60% accuracy. In comparison, human performance on a similar equine orthopaedic dataset is at 51% accuracy [3].

We present pain classification results of our model alongside ablation studies comparing the contribution of each module. In addition, we analyze the features of pain detected by our system, and note their correspondence with current veterinary knowledge on equine pain behavior.

Our contributions are:

- Creating a disentangled horse pose representation from multi-view surveillance video footage of horses in box stalls using self-supervised novel view synthesis.
- Presenting a method for video segment level pain classification from the learned disentangled horse pose representation that is trained using weak pain labels and a novel multiple instance learning (MIL) loss.
- Extensive experiments including analysis of our automatically detected pain video segments in terms of cues used for pain diagnosis in veterinary practice.

2. Related work

Our work is closely related to novel view synthesis: given a view of a scene, the task is to generate images of the scene from new viewpoints. This is challenging, as it requires reasoning about the 3D structure and semantics of the scene from the input image. Rhodin et al. [48] train an encoder-decoder architecture on the novel view synthesis task using synchronized multi-view data to create a latent representation which disentangles human pose and appearance. Our work uses the same approach to learn a disentangled horse pose representation. However, while their method uses the learned pose representation for the downstream and strongly related task of 3D and 2D body keypoint estimation, our work uses the latent representation to classify animal pain in a weakly supervised setting.

Others achieve novel view synthesis assisted by either noisy and incomplete [10, 56, 38, 22], or ground truth depth maps, in addition to images during training [31, 60, 52, 37].

Similar to our work, generative models have been used with emphasis on learning a 3D aware latent representation. Of note are deep voxels [55], HoloGAN [36], and Synsin [65]. Although we share the aim to create a 3D

aware latent representation, these methods emphasize the generation of accurate and realistic synthetic views, while our work focuses on using the latent representation for the downstream task of pain classification. While we do make use of multi-view data, the different viewpoints are few – 4 – and are separated by a wide baseline, unlike the above mentioned novel view synthesis works.

Generative models with disentangled latent representations have been developed for a wide range of purposes such as to discover intrinsic object properties like normals, albedo, lighting, and alpha matte for faces [53], fine grained class attributes for object discovery and clustering [54], and pose invariant identity features for face recognition [59], and have been a topic of extensive research [57, 8, 23, 13, 24]. Our work relies on a disentangled pose representation from multi-view data, and places emphasis on utilizing the learned representation for a downstream task. Self-supervised disentangled pose representations have been used for 2D and 3D keypoint recognition [48, 47, 9, 7, 19], but no previous work has used them for the behavior related task of pain recognition, particularly in animals.

There is a growing body of work on deep visual learning for animal behavior and body understanding. This includes work on animal body keypoint prediction [6, 34], facial keypoint prediction [66, 44, 28], and dense 3D pose estimation via transfer from human datasets [50] and fitting a known 3D model to 2D image [69, 68]. Of note are [67] which uses a synthesized zebra dataset to train a network for predicting zebra pose, and [30] which develops a horse specific 3D shape basis and applies it to the downstream task of lameness detection.

Beyond animal keypoint and pose prediction, there is a growing body of research on detecting animal affective states from images and videos. Of great relevance is Broomé et al.'s [4] work on horse pain recognition that uses a fully recurrent network for pain prediction on horse videos. The method uses strong supervision and a dataset with close up videos of horses with acute pain. Follow up work [3] uses a similar network architecture and domain transfer from between acute and orthopaedic pain for horse pain classification. Sheep [32], donkey [25], and mouse [61] pain have also been explored with promising results. However, previous methods use either facial data, strong supervision, or additional information such as keypoints, segmentation masks, or facial movement annotation to learn the pain models. In contrast, we use weak supervision, with no additional annotation or data, and video data with the full horse body visible rather than just the face.

3. Approach

The equine pain dataset comprises time aligned videos of horses from multiple views, and pain labels from periodic observations. We use a two-step approach for pain classification. In the first stage, we train an encoder-decoder architecture for novel view synthesis, and learn an appearance invariant and viewpoint covariant horse pose latent representation in the process. In the second stage, we train a pain classification head using the trained pose aware latent representation as input to diagnose pain in video sequences. Since the dataset does not have detailed temporal annotation of horse pain expression, we use a MIL approach for pain classification for video sequences. In the following sections, we first describe the dataset, followed by details of our view synthesis and pain classification methods.

3.1. Dataset

The Equine Orthopaedic Pain (EOP) Dataset [2] comprises of 24-hour surveillance footage of 8 horses filmed over 16 days before and during joint pain induction. The experimental protocol was approved by the Swedish Ethics Committee in accordance with the Swedish legislation on animal experiments (diary number 5.8.18-09822/2018). As few horses as possible were included and a fully reversible lameness induction model was used. Hindlimb lameness was induced by injecting lipopolysaccharides (LPS) into the hock joint, leading to inflammatory pain and various degrees of joint swelling. To decrease such pain, the horse tries to unload the painful limb both at rest and during motion, and movement asymmetry can then be measured objectively. The horses were stalled individually in one of two identical box stalls, with four surveillance cameras in the corners of each stall capturing round the clock footage. Starting 1.5 hours after joint injection, horses were periodically removed from the box stall to measure movement asymmetry during trot. Measurements were discontinued once horse movement asymmetry returned to levels similar to before LPS injection. In addition, pain assessments using four different pain scales [12, 16, 62, 5] were performed by direct observation 20 minutes before and after each movement asymmetry measurement. The dataset can be accessed for collaborative research by agreement with the authors of [2].

3.1.1 Data preprocessing

We use video data from two hours of pre-induction baseline as our no-pain data. Pain level was determined by averaging the Composite Pain Scale (CPS) score [5] of three veterinary experts during each pain assessment session. The session with the highest average CPS score was selected as peak pain period. Two hours closest to the pain observation session were used as our pain data. Four (two pain, two no-pain) hours per horse, from four surveillance cameras for eight horses results in a total of 128 hours of film. We only use the time periods when no humans are present in the stall

or the corridor outside the stall, reducing the likelihood of interactivity leading to changes in the horses' behavior.

Videos were cut in to 2 minute long segments (see imp. details in Sec. 4), and all segments belonging to the two hour peak pain period were labeled as pain, and pre-induction period were labeled as no-pain.

Note that we extrapolate pain labels from the pain observation session to the closest unobserved horse videos. These videos have not been viewed and annotated as containing pain behavior by experts, and the typical duration and frequency of horse pain expression is not known. Similarly, the no-pain videos were not manually verified as not containing any pain expression. As a result, our video level pain labels are likely noisy and contribute to the difficulty of our task.

3.2. Multi-view synthesis

The multi-view synthesis network uses the same architecture and training methodology as original work by Rhodin et al. [48]: a U-Net architecture [49] learns a disentangled horse appearance and pose representation in its bottleneck layer. This is done by training the model to synthesize an input scene from a different viewpoint.

Specifically, given an input video frame $x_{i,t}$, from viewpoint i, at time t, the encoder, f_E, outputs are a latent representation of the horse pose, $p_{i,t}$, and of the horse appearance, $h_{i,t}$:

$$p_{i,t},\, h_{i,t} \,=\, f_E\,(x_{i,t}). \tag{1}$$

During training, both $p_{i,t}$ and $h_{i,t}$ are manipulated before being input to the decoder, f_D. The pose representation is multiplied by the relative camera rotation, $\mathbf{R}$, of viewpoint i and j, so that the pose representation input to f_D is in the same space as pose representations for viewpoint j:

$$p_{i\to j,t} \,=\, \mathbf{R}_{v\to j}\, p_{i,t}. \tag{2}$$

The appearance representation is swapped by the appearance representation of an input frame of the same horse from the same viewpoint, but from a different time, t', and hence likely with a different pose. This appearance representation swap encourages the network to disentangle pose and appearance. In addition, a background image for each viewpoint, b_i, is input to the decoder so that the network does not focus on learning background information for synthesis and instead focuses on the horse:

$$x'_{i\to j,t} \,=\, f_D\,(p_{i\to j,t},\, h_{i,t'},\, b_i). \tag{3}$$

As in [48], the synthesized image, $x'_{i\to j,t}$, is supervised by both a pixel-level mean squared error (MSE) loss compared with the ground truth image, $x_{j,t}$, as well as a perceptual loss on ImageNet pretrained ResNet18 [21] penultimate layer's features.

$$\begin{aligned} L_{MVS} = ||x'_{i\to j,t} \,-\, x_{j,t}||^2 + \\ \alpha\, ||\theta_{RN}(x'_{i\to j,t}) \,-\, \theta_{RN}(x_{j,t})||^2, \end{aligned} \tag{4}$$

where α is a loss weighting parameter, and $\theta_{RN}(x)$ outputs ResNet18's penultimate feature representation for input image x. The multi-view synthesis loss, L_{MVS}, is averaged across all instances in a batch before backward propagation.

During testing, the synthesized image is generated without swapping of the appearance representation.

Similar to [48], we detect and crop the horse in each frame to factor out scale and global position. $\mathbf{R}_{i\to j}$ is calculated with respect to the crop center instead of the image center. The crop is sheared so that it appears taken from a virtual camera pointing in the crop direction. In more detail, the crop is transformed by the homography induced by rotating the camera so that the ray through its origin aligns with the crop center.

3.2.1 Refining multi-view synthesis for horses

Unlike the Human 3.6 dataset [26] used in [48], which features actors that move around constantly, the EOP dataset features the horse standing or grazing in similar pose for long periods of time. As a result, randomly selecting a frame for the appearance feature swap on EOP can lead to suboptimal appearance disentanglement. Therefore, we pre-select time sequences with a variety of horse poses for training, based on the optical flow magnitude.

The EOP dataset was collected over multiple months, during which the cameras, although fixed as firmly as possible, were nudged by chewing from curious horses. As a result, background images calculated as the median image over the entire dataset were blurry. We therefore extracted separate higher quality background images per month.

3.3. Classifying pain

The self-supervised base network (Section 3.2) provides us with a means to disentangle the horse pose from its background, and appearance from a given input image. This representation is now used to train a pain classification head.

We treat video pain classification as multiple instance learning (MIL) problem. Every video comprises time segments that are independently classified as pain or no pain. These time segment level predictions are collated to obtain a video level pain prediction.

We classify pain classification with both frame and, following insight from earlier works [4, 46] showing pain classification from instantaneous observations to be unreliable, clip level inputs. For the latter, we concatenate per-frame latent representations into a clip level volume used as the atomic unit for pain classification during both training and testing.

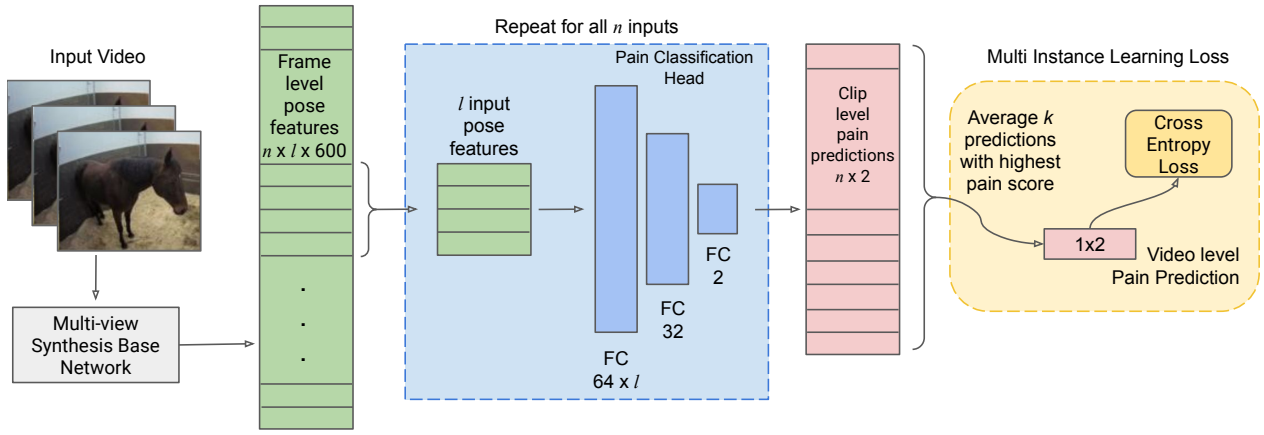

Figure 2. **Pain Classification Model.** Pose representations from the trained multi-view synthesis model are extracted for each frame of an input video, and collated into l length clips. The pain classification head, comprising three fully-connected (FC) layers, predicts pain for each clip. Clip level predictions are collated and supervised by the proposed MIL loss.

The network architecture, shown in Figure 2, comprises two hidden linear layers with ReLU and dropout, followed by a classification layer with two dimensional output for pain vs no-pain. Frame level predictions from the first linear layer are concatenated in to a $64 * l$ vector, where l is the number of frames in each input segment, that is then forwarded through the network.

More specifically, each video sequence s, comprises n time segments indexed by t. The pain head, θ, provides a two-dimensional output that is softmaxed to obtain the pain, and no pain confidence values for time segment t, where $\mathbf{p}$ represents the set of pose representations of all frames in t:

$$y_{i,t}^{NP}, y_{i,t}^{P} = \text{softmax}\left(\theta\left(\mathbf{p}_{i,t}\right)\right). \tag{5}$$

The k time segments with the highest *pain* predictions are averaged to obtain the video level pain prediction:

$$y_{i,s}^{P} = \frac{1}{k}\sum_{t\in S} y_{i,t}^{P}, \quad y_{i,s}^{NP} = \frac{1}{k}\sum_{t\in S} y_{i,t}^{NP}, \tag{6}$$

where S is the set of k time segments' indices with the highest pain prediction:

$$S = \{j \mid y_{i,j}^{P} \in \max_{K}\{y_{i,1}^{P}, y_{i,2}^{P}, \ldots, y_{i,n}^{P}\}\}. \tag{7}$$

The video level pain predictions are then supervised with a cross-entropy loss.

Parameter k is set to $\lfloor \frac{n}{d} \rfloor$, where d is randomly selected from $\{1, 2, 4, 8\}$ at every training iteration, and set to 8 during testing. d correlates with the proportion of a video that is predicted as an action class. As we do not know the proportion of time a horse in pain will express pain in a video, varying this parameter randomly is likely to provide the most robust performance, as shown in previous work [45].

Our key insight in designing the loss is that a horse in pain need not express pain constantly, and the absence of its pain behavior can be rightly classified as no-pain. By collating only the predictions for the top k *pain* segments to obtain the video level prediction, we do not penalize the network for classifying no-pain time segments within a pain video. We hence require only that the pain predictions have high confidence in pain videos, and that no time segments have high pain confidence in no-pain videos. Our loss is different from the MIL loss used in literature (e.g. [63, 39, 35], which would have averaged the highest k predictions for both pain and no-pain class independently. Section 4.2 compares these loss formulations.

4. Experiments

Implementation details. Optical flow was calculated using Farnebäck's method [15] on video frames extracted at 10 frames per second. Time segments that had optical flow magnitude in the top 1%, 143559 frames, were used to train the multi-view synthesis module. We use leave-one-out, subject exclusive training. Multi-view synthesis (backbone) models were trained for 50 epochs at 0.001 learning rate using Adam optimizer. The perceptual loss is weighted 2 times higher than the MSE loss during training. The same U-Net based architecture as in [48] is used. The 600 dimensional pose representation is reshaped to 200×3 for multiplication with the rotation transformation $\mathbf{R}$.

The pain classification dataset comprises video segments, s, with the maximum length of 2 minutes. MaskR-CNN [20] was used to detect, crop and center the horse in each frame. Missing MaskRCNN detections can reduce the video segment length, but no instances less than 10 seconds in length are included. Note that all data with valid horse detections were included in the pain classification dataset, and not just the high motion frames.

Pain is predicted for short clips, of length l, that are collated for video level pain prediction. We show results when

using clips of length 1 frame (frame based), and with clips of length five seconds at 2 fps. The five second clip length is set following past research [4]; additionally, [46] suggests it to be the duration of time a horse pain expression lasts.

The backbone network is frozen when training the pain classification head, which is trained for 10 epochs at 0.001 learning rate. Leave-one-out training is again used, excluding the same test subject as for the backbone network. In addition, pain classification performance on a validation set is calculated after each epoch, and the model at the epoch with the highest performance is used for testing. Data from the non-test subject with the most balanced pain/no-pain data distribution is used as the validation data. Further details are included in the supplementary.

4.1. Disentangled representation learning

Disentangled pose representation. We explore the quality of the latent representation qualitatively. The ideal pose representation would be able to cluster the same horse pose regardless of viewpoint. In addition, this representation would be disentangled from horse appearance.

Given the pose representation of a test input image at time t from viewpoint i, $p_{i,t}$, we find its top 3 nearest neighbors in the training data after rotation to viewpoint j. That is, we find the nearest neighbors of $\mathbf{R}_{i\to j}\, p_{i,t}$. Some qualitative results are shown in Figure 3, where the second columns show the actual image from viewpoint j.

The top 3 neighbors are consistent with the expected ground truth, showing that the network has learned a pose representation that is viewpoint covariant. One exception is the second nearest neighbor in the third row, left set, that is quite different from the ground truth image. The backgrounds of the retrieved images are often different from the query background, for example in the middle set, showing background disentanglement. Moreover, the retrieved horses may be physically different from the query horse, showing appearance disentanglement: a black horse is retrieved in the fourth row, left set, and a horse with a white blaze is retrieved in the second row, right set. Interestingly, when the horse head and neck is occluded in the second row, right set, the nearest neighbors suggest that the model hallucinates a reasonable – though not entirely accurate – neck and head position.

Disentangled identity representation. In Figure 4, we show results of swapping the appearance representation. As explained in Section 3.2, the decoder uses an appearance and pose representation to reconstruct an image. We compare reconstructed test images with and without swapping the appearance representations with the appearance representation of a training horse with a black coat. Good disentanglement would show a horse with the same pose as the input image, but with a black rather than a brown coat.

	True Performance		Oracle Performance	
	F1 Score	Accuracy	F1 Score	Accuracy
Ours-Frame	**58.5±7.8**	**60.9±5.7**	60.8±6.4	62.3±5.4
Ours-Clip	55.9±5.1	57.8±4.4	**65.1±6.7**	**65.6±6.5**
Ours-Clip-HaS	56.5±5.0	58.6±4.3	63.6±6.2	64.6±5.8
Scratch	54.5±9.1	57.3±6.5	61.7±8.1	63.2±7.7
Broomé '19 [4]	53.0±8.1	55.2±7.0	60.0±8.3	59.4±8.3

Table 1. Comparison of frame and clip based pain heads against a models trained from scratch with early stopping using a hold out dataset (True Performance) and best case (Oracle).

	F1 Score	Accuracy
Ours-Frame	**58.5±7.8**	**60.9±5.7**
NoApp	57.1±7.9	59.7±6.0
NoBG	53.2±8.3	55.9±7.4
NoApp+NoBG	55.6±5.2	57.8±5.7
NoMod	56.6±7.4	58.4±5.5

	F1 Score	Accuracy
Ours-MIL	**58.5±7.8**	**60.9±5.7**
CE-Clip	52.2±10.2	57.0±6.4
CE-Frame	49.1±10.9	55.2±5.9
MIL-OG	47.7±12.7	55.0±8.2

Table 2. *Left*: Pain classification performance with backbones of varying pose disentanglement. *Right*: Comparison of cross-entropy loss and multi instance learning loss variations.

The model trained with uniformly sampled video frames is not able to disentangle appearance and pose and reconstructs horses with more or less the same color, both with and without appearance swapping. High motion sampling during training increases the chance that the swapped appearance frame features a horse in a different pose than the input frame. As seen by the darker coats in the last row, it hence leads to better appearance disentanglement. It also leads the network to learn a variety of poses, as can be seen by the crisper reconstructions around the head and legs, in the fifth and sixth columns. Lastly, the background is crisper in the last two rows. This is due to our use of crisper background images derived from the same month as the input that results in better background disentanglement.

4.2. Pain classification

We present F1 score and accuracy for pain classification results, taking the unweighted mean of the F1 score across both classes. Both metrics are averaged across all training folds, and presented here alongside the standard deviation.

In Table 1 we compare variants of our pain classification head. The 'True Performance' column shows the performance of the model selected by early stopping based on performance on a holdout validation set. 'Oracle Performance' shows results if the early stopping criteria aligned with the epoch with the best testing performance and shows the upper limit performance. 'Ours-Frame' uses frame level inputs ($l = 1$), 'Ours-Clip' uses 5 second clips. The 'Scratch' model has the same architecture as the encoder part of the base network and the pain head and is trained on frame level inputs, supervised by our MIL loss, from scratch.

Both our frame and clip based models have better true performance than the scratch model; the frame based model shows 4 percentage points (pp.) higher F1 score. Additionally, the oracle performance is either comparable or better than the scratch model, even though the latter learns more

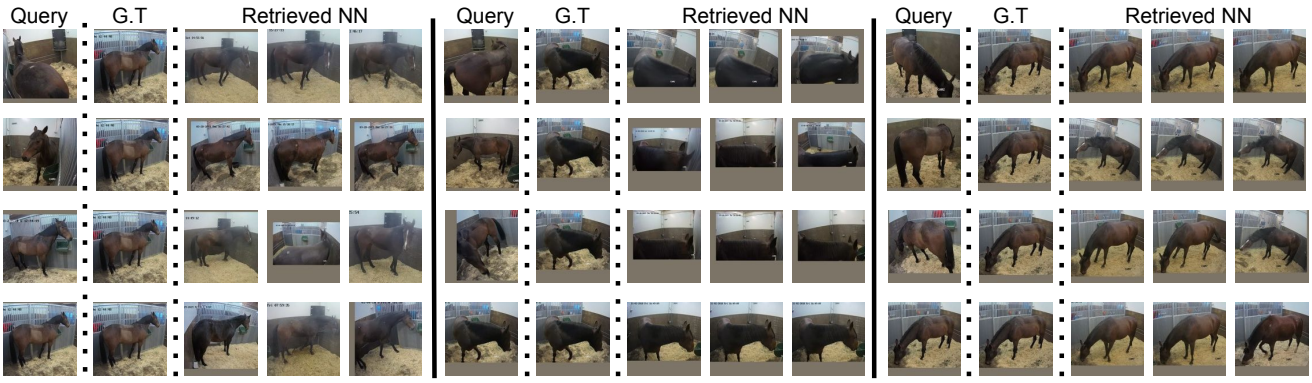

Figure 3. Nearest neighbor retrieval on latent pose representation. The pose representation of the query image is rotated before nearest neighbor retrieval. The nearest neighbors match the pose in the actual ground truth image from the rotated view.

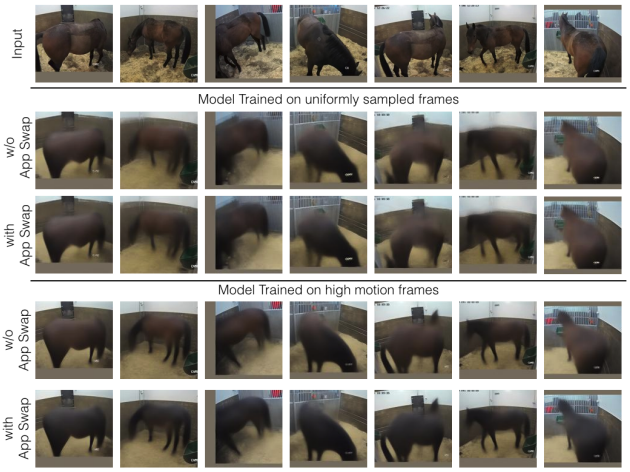

Figure 4. Appearance (App) swapping on different models. Each column shows the decoder output for the corresponding input image from the first row. In the third and fifth rows, the appearance representations are swapped with the appearance representation from a training horse with a black coat.

parameters specific to the pain classification task. At the same time, our models' use of a disentangled pose representation ensures that pose rather than extraneous information is used to deduce the pain state. These results indicate that using a disentangled pose representation is useful for a dataset such as ours with limited training subjects.

All models suffer from some degree of overfitting, as can be seen from the difference between the true and oracle results. The clip based and scratch models exhibit a high degree of overfitting: with oracle F1 score at $\sim$ 10 and $\sim$ 7 pp. higher than the true performance. This can be expected since these models learns more parameters, and points to the need for a light weight pain classification head.

However, the use of temporal information through clip level inputs results in a much higher oracle performance compared to frame level inputs (5 pp. higher F1). We therefore add more regularization by using random adversarial erasing via Hide-and-Seek [29] ('Ours-Clip-HaS') resulting in 1 pp higher accuracy. Altogether, clip based pain prediction is most promising, but requires more regularization to compete with the simple frame based results.

Finally, 'Broomé '19' shows results of the convolution LSTM model proposed in [4] on EOP dataset. As we do not use optical flow frames in other experiments, we use the one-stream (RGB) variant of their model. Following the original work we evaluate and train on 10 frame clips at 2 FPS, supervised with binary cross entropy loss. The method achieves performance slightly worse than the 'Scratch' model, and shows similar overfitting behavior.

Backbone Ablation study. In Table 2 (left) we show the results on an ablation study on our backbone model. Each model variant has the same pain detection head as 'Ours-Frame', but uses different backbone trainings. 'NoBG' means that the backbone was trained without providing a background frame to the decoder and has low background disentanglement. 'No-App' means that the appearance feature swapping was not used during training, and appearance disentanglement is low. Finally, 'No-Mod' does not use our modifications mentioned in Section 3.2.1 and shows the importance of motion based sampling, and a better background, but is otherwise identical to 'Ours-Frame'.

Each component of the backbone is important for final pain results, and shows the importance of a well disentangled pose representation for pain classification. As shown in Figure 4, 'No-Mod' has little appearance disentanglement and shows similar results to 'No-App', but is slightly worse probably because it also uses worse background images. Background disentanglement is most important for our performance, with 'NoBG' achieving 5 pp. poorer performance than our full model. Interestingly, when both background image and appearance swapping are removed ('NoBG+NoApp'), the model does better than when just background image is removed ('NoBG'). This may be because appearance swapping prevents the model from encoding any background knowledge in the appearance latent features, making it harder for the model to disentangle back-

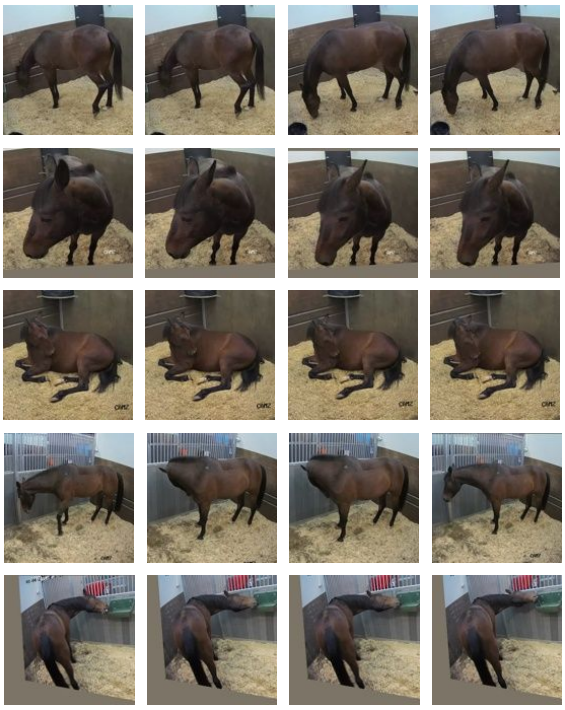

Figure 5. Video segments correctly detected as painful by our model. The detected segments display signs of pain such as avoiding weight bearing (1st row), backwards ears (2nd row), lying down (3rd row), looking at the painful leg (4th row), and stretching (last row).

ground on its own.

Weakly supervised learning. In Table 2 (right), we evaluate the importance of multi-instance learning. As discussed in Section 3.3, our version of MIL loss (Ours-MIL), averages the pain and no-pain predictions of the top k time segments (or frames) with the highest *pain* prediction. We contrast this against the original MIL loss (MIL-OG), which averages the top pain and no-pain predictions separately. Lastly, we compare against a simple cross-entropy loss (CE), where each frame or clip is separately supervised during training. The test results are still obtained by averaging the top k clip level predictions to keep results comparable.

Firstly, by comparing Ours-MIL against CE we see that using a MIL setting is essential, and that pain (or no-pain) behavior is not present in each frame from a pain (or no-pain) video. In fact, the results are lower than random for a CE-trained model (49.1% F1 score). The use of dynamic information with clip based inputs leads to improved performance, although the overall performance is lower than for weakly supervised training. Secondly, compared to MIL-OG, we see that Ours-MIL is necessary to learn a reasonable model for pain. In fact, MIL-OG leads to a worse model of pain than random guessing, at 47.7% F1 score. This bolsters our underlying reasoning that clips with no-pain features may exist in pain videos and should not be penalized during training in order to develop a good pain model.

4.3. Attributes of pain

Figure 5 shows example clips that our model classifies as painful. The clips contain classic signs of pain such as 'lowered ears' [17] (second and fourth row), a lifted left hind limb (first row), corresponding to 'non-weight bearing' [5], 'lying down' (third row), 'looking at flank' (third and fourth row) [41], and an example of gross pain behavior, 'stretching' (last row) [16]. These results show a good correspondence between the visual attributes our model focuses on, and pain scales used by veterinary experts. While we only expected body behavior to be picked up by the pain model, interestingly, subtle facial behavior, specifically, ear movements, are also picked up and learned. More results are shown in the supplementary.

5. Discussion

Equine pain detection is an intrinsically challenging problem. Recent work [3] shows that equine orthopaedic pain is particularly difficult to detect, as it has high signal to noise ratio and therefore requires transfer learning from cleaner acute pain data for reasonable performance (49.5% F1 before transfer learning, and 58.2% after). EOP dataset presents additional challenges as pain labels are sparse and noisy. Despite these challenges, our method can achieve 60% accuracy which is better than human expert performance on equine orthopaedic pain (average of 51.3% accuracy across three pain levels [3]), and on par with human expert performance on acute pain [4].

Our method is scalable and pragmatic as we use unobtrusive surveillance footage, and sparse pain labels. With few training subjects, and noisy labels, we ensure that pain is learned from horse body language alone by use of a self-supervised generative model to disentangle the horse pose from appearance and background. The resulting pose representation is used to learn a pain prediction model, weakly supervised with a novel pain specific multiple instance learning loss. We qualitatively analyze our model's disentangled pose and appearance features, and show quantitative and qualitative results on pain classification.

Future work should include means to exclude pain predictions on videos without a clear enough view of the horse. As we do not know the typical frequency and duration of horse pain expression, the video length used in this work may be sub-optimal. Research on the optimal duration of a video segment to guarantee the observation of pain behavior, and further regularization of the pain classification head are promising directions of future improvement.

Acknowledgements. This work was supported in part by NSF IIS-1812850.

References

[1] Pia H Andersen, KB Gleerup, J Wathan, B Coles, H Kjellström, S Broomé, YJ Lee, M Rashid, C Sonder, E Rosenberg, and D Forster. Can a machine learn to see horse pain? an interdisciplinary approach towards automated decoding of facial expressions of pain in the horse. In *Measuring Behavior*, 2018.

[2] Katrina Ask, Marie Rhodin, Lena-Mari Tamminen, Elin Hernlund, and Pia Haubro Andersen. Identification of body behaviors and facial expressions associated with induced orthopedic pain in four equine pain scales. *Animals*, 10(11), 2020.

[3] Sofia Broomé, Katrina Ask, Maheen Rashid, Pia Haubro Andersen, and Hedvig Kjellström. Sharing pain: Using domain transfer between pain types for recognition of sparse pain expressions in horses. *arXiv preprint arXiv:2105.10313*, 2021.

[4] Sofia Broomé, Karina Bech Gleerup, Pia Haubro Andersen, and Hedvig Kjellström. Dynamics are important for the recognition of equine pain in video. In *CVPR*, 2019.

[5] G Bussieres, C Jacques, O Lainay, G Beauchamp, Agnès Leblond, J-L Cadoré, L-M Desmaizières, SG Cuvelliez, and E Troncy. Development of a composite orthopaedic pain scale in horses. *Research in veterinary science*, 85(2), 2008.

[6] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, 2019.

[7] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3D pose estimation with geometric self-supervision. In *CVPR*, 2019.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[9] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *CVPR*, 2019.

[10] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *ICCV*, 2019.

[11] Britt Alice Coles. No pain, more gain? evaluating pain alleviation post equine orthopedic surgery using subjective and objective measurements. *Swedish University of Agricultural Sciences, Masters Thesis*, 2016.

[12] Emanuela Dalla Costa, Michela Minero, Dirk Lebelt, Diana Stucke, Elisabetta Canali, and Matthew C Leach. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS one*, 9(3), 2014.

[13] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, 2017.

[14] Paul Ekman. Facial action coding system (FACS). *A human face*, 2002.

[15] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*. Springer, 2003.

[16] KB Gleerup and Casper Lindegaard. Recognition and quantification of pain in horses: A tutorial review. *Equine Veterinary Education*, 28(1), 2016.

[17] Karina B Gleerup, Björn Forkman, Casper Lindegaard, and Pia H Andersen. An equine pain face. *Veterinary anaesthesia and analgesia*, 42(1), 2015.

[18] Claudia Graubner, Vinzenz Gerber, Marcus Doherr, and Claudia Spadavecchia. Clinical application and reliability of a post abdominal surgery pain assessment scale (paspas) in horses. *The Veterinary Journal*, 188(2), 2011.

[19] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2D features and intermediate 3D representations. In *CVPR*, 2019.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask rcnn. In *ICCV*, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[22] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (TOG)*, 37(6), 2018.

[23] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2016.

[24] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *CVPR*, 2018.

[25] Hilde I Hummel, Francisca Pessanha, Albert Ali Salah, Thijs JPAM van Loon, and Remco C Veltkamp. Automatic pain detection on horse and donkey faces. In *FG*, 2020.

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[27] JL Ireland, PD Clegg, CM McGowan, SA McKane, KJ Chandler, and GL Pinchbeck. Comparison of owner-reported health problems with veterinary assessment of geriatric horses in the united kingdom. *Equine veterinary journal*, 44(1), 2012.

[28] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *CVPR*, 2020.

[29] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017.

[30] Ci Li, Nima Ghorbani, Sofia Broomé, Maheen Rashid, Michael J Black, Elin Hernlund, Hedvig Kjellström, and Sil-

via Zuffi. hSMAL: Detailed horse shape and pose reconstruction for motion pattern recognition. In *CV4Animals Workshop, CVPR*, 2021.

[31] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

[32] Yiting Lu, Marwa Mahmoud, and Peter Robinson. Estimating sheep pain level using facial action unit detection. In *FG*. IEEE, 2017.

[33] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *FG*, 2011.

[34] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *CVPR*, 2020.

[35] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, 2018.

[36] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *ICCV*, 2019.

[37] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3D ken burns effect from a single image. *ACM Transactions on Graphics (TOG)*, 38(6), 2019.

[38] David Novotny, Ben Graham, and Jeremy Reizenstein. Perspectivenet: A scene-consistent image generator for new view synthesis in real indoor environments. In *Advances in Neural Information Processing Systems*, 2019.

[39] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, 2018.

[40] Jill Price, Seago Catriona, Elizabeth M Welsh, and Natalie K Waran. Preliminary evaluation of a behaviour–based system for assessment of post–operative pain in horses following arthroscopic surgery. *Veterinary anaesthesia and analgesia*, 30(3), 2003.

[41] Lori C Pritchett, Catherine Ulibarri, Malcolm C Roberts, Robert K Schneider, and Debra C Sellon. Identification of potential physiological and behavioral indicators of postoperative pain in horses after exploratory celiotomy for colic. *Applied Animal Behaviour Science*, 80(1), 2003.

[42] Marja Raekallio, Polly M Taylor, and M Bloomfield. A comparison of methods for evaluation of pain and distress after orthopaedic surgery in horses. *Veterinary Anaesthesia and Analgesia*, 24(2), 1997.

[43] M Rashid, S Broomé, PH Andersen, KB Gleerup, and YJ Lee. What should i annotate? an automatic tool for finding video segments for equifacs annotation. In *Measuring Behavior*, 2018.

[44] Maheen Rashid, Xiuye Gu, and Yong Jae Lee. Interspecies knowledge transfer for facial keypoint detection. In *CVPR*, 2017.

[45] Maheen Rashid, Hedvig Kjellström, and Yong Jae Lee. Action graphs: Weakly-supervised action localization with graph convolution networks. In *WACV*, 2020.

[46] Maheen Rashid, Alina Silventoinen, Karina Bech Gleerup, and Pia Haubro Andersen. Equine facial action coding system for determination of pain-related facial responses in videos of horses. *Plos one*, 15(11), 2020.

[47] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *CVPR*, 2019.

[48] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *ECCV*, 2018.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[50] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S. McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *CVPR*, 2020.

[51] Debra C Sellon, Malcolm C Roberts, Anthony T Blikslager, Catherine Ulibarri, and Mark G Papich. Effects of continuous rate intravenous infusion of butorphanol on physiologic and outcome variables in horses after celiotomy. *Journal of Veterinary Internal Medicine*, 18(4), 2004.

[52] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3D scene reconstruction with multi-layer depth and epipolar transformers. In *ICCV*, 2019.

[53] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017.

[54] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019.

[55] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3D feature embeddings. In *CVPR*, 2019.

[56] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019.

[57] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6), 2000.

[58] Catherine Torcivia and Sue McDonnell. Equine discomfort ethogram. *Animals*, 11(2), 2021.

[59] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[60] Shubham Tulsiani, Saurabh Gupta, David F Fouhey, Alexei A Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3D scene. In *CVPR*, 2018.

[61] Alexander H Tuttle, Mark J Molinaro, Jasmine F Jethwa, Susana G Sotocinal, Juan C Prieto, Martin A Styner, Jeffrey S Mogil, and Mark J Zylka. A deep neural network to assess spontaneous pain from mouse facial expressions. *Molecular pain*, 14, 2018.

[62] Johannes PAM van Loon and Machteld C Van Dierendonck. Monitoring acute equine visceral pain with the equine utrecht

university scale for composite pain assessment (EQUUS-COMPASS) and the equine utrecht university scale for facial assessment of pain (EQUUS-FAP): a scale-construction study. *The Veterinary Journal*, 206(3), 2015.

[63] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017.

[64] Jen Wathan, Anne M Burrows, Bridget M Waller, and Karen McComb. EquiFACS: The Equine Facial Action Coding System. *PLoS one*, 10(8), 2015.

[65] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020.

[66] Heng Yang, Renqiao Zhang, and Peter Robinson. Human and sheep facial landmarks localisation by triplet interpolated features. In *WACV*, 2016.

[67] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, 2019.

[68] Silvia Zuffi, Angjoo Kanazawa, and Michael J Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, 2018.

[69] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, 2017.