Detecting Stress in Dogs using Deep Learning on Pose/Feature Data

Max Jacobsen
University of Minnesota
jaco2017@umn.edu

Benjamin Keeney
University of Minnesota
keene077@umn.edu

Kenny Tran
University of Minnesota
tran0726@umn.edu

Abstract

Traditional techniques for emotion recognition have largely focused on human emotions, as these are more widely useful. However, the value in autonomously detecting emotion in animals is readily apparent for many careers involved in the handling of animals - detecting stress in zoos or dog shelters passively through surveillance to allow a quick response, giving inexperienced handlers a quick heuristic, or for pet owners to quickly diagnose issues. In this paper we present a new dataset with labeled images of both relaxed/happy dogs and of stressed dogs. We also present a new method for emotion recognition in animals using a combination of pose and visual feature data.

1. Introduction

Understanding animal emotion is essential for providing proper care and preventing harm to the animal or harm that it could cause to other animals or caretakers interacting with them through aggressive behavior. While signs of stress/agitation are often easily recognized by professionals trained in animal handling, there is value in automatic recognition of these traits for inexperienced caretakers or for passive observation - such as in a kennel or shelter where security footage could be used to monitor the animals' stress levels.

While much research has been done on the recognition of emotion in humans as a more familiar and more immediately useful challenge, no work (as far as the authors are aware) has been done on recognition of animal emotion through purely visual/auditory cues. In both humans and animals, however, this problem is challenging - it goes beyond classification to detecting often very minute perturbations to a subject. Therefore the accuracy of many models on non-exaggerated examples is often far lower than accuracies seen in classification problems.

This paper addresses the problem of recognition of emotions in animals by first collecting a new dataset of images of dogs labelled as either examples of stressed or of unstressed behavior, the *Dog Stress Dataset* (DOGST). Secondly, we propose a new method based on the unique challenge of classifying animal emotion which utilizes a combination of both pose and feature data.

We believe that existing methods are limited in this use case by their focus on specifically human facial features or by not taking into account the full range of available data (as in methods which exclusively focus on pose).

More specifically, we speculate that a combination of the dog's skeletal pose and features which pose does not capture (such as facial expression or, in the case of our data, tail position) through images. All of these are known elements used by dog to express emotion [21] - and therefore are essential features to capture.

Our approach involves combining pose and bounded image data of the animal through a deep neural network. We show that our model is comparable in performance (though not stronger) than a context-based human emotion model baseline at classifying dog stress. We conclude that more rigorous and expansive data collection is needed, as the images in our dataset both have low confidence in pose estimation and are relatively few in number. We also conclude that both elements of our model (image and pose) improve the performance.

2. Related Work

Specific facial cues have long been the primary method through which emotion recognition has been performed in humans. Older methods like Ekman and Friesen's *Facial Action Coding System* (FACS) [4, 5] track local muscle movements (known as *Action Units* in FACS) of the face based on domain knowledge of the expression of human emotion. Many of these methods classify emotions using the 6 universal facial expressions across culture identified by Ekman and Friesen [3], though some measure across continuous spectra such as the Valence, Arousal, and Dominance Emotional State Model [16].

These types of methods generally rely on handcrafted features or on simpler feature analysis with methods such as HOG [13]. Recent representation-learned approaches using deep neural networks are capable of outperforming manual methods in extracting local movements [1] or in directly classifying emotion from either images [12, 2] or video [24].

Another central area of study is recognizing emotion through pose. Schindler et. al demonstrated a neural network structure inspired by biologically plausible neuron responses to estimate emotion through pose [20]. More recent work includes using Recurrent Neural Networks on KNN-classified pose video data [23].

Techniques using pure convolutional or recurrent neural networks have also been successful, with 3D convolutional networks (C3D) and RNN networks being applied to video of facial expressions [8, 7]. Autoencoders have also been used effectively [22, 9]. Many recent approaches using CNNs focus on features in context, isolating both the target's face or body and feeding it and the surrounding context into separate networks that perform a joint evaluation of emotion [10, 11, 14].

There are also plenty of multimodal techniques, which use some combination of previously mentioned features to evaluate emotion, such as Nicolaou et. al's combination of face, shoulder, and audio cues [19], or that use multiple other methods in an ensemble technique such as EmoNets [1]. Recent work by Mou et. al has analyzed emotion at a group level using a combination of existing methods [18].

No previous approaches, as far as the authors are aware, combines pose and feature data using convolutional neural networks in the way presented. As mentioned previously, we are also unaware of any studies which specifically study animal emotion using computer vision.

3. Dog Stress Dataset

The Dog Stress Dataset (DOGST) consists of images collected from frames of dog rehabilitation videos available on YouTube and of manually collected images from Google using search terms related to emotion and dogs such as ”angry dog”, ”aggressive dog”, ”playful dog”, etc. All of these images show a dog acting in either a stressed pattern of behavior or in a relaxed/playful pattern of behavior. This is a challenging set of images, as most are real world situations, often with poor framing or partial obstruction of the animal.

There are 2,750 images where 1,650 are labeled as examples of a stressed dog and 1,100 are labeled as examples of an unstressed dog (a class label of 1 and 0, respectively). These images were filtered further by classifying the presence of dogs in each image, and not including those with a confidence less than 0.3.

Further preprocessing was done on the dataset, by extracting bounded images of the dogs in each image. This was to avoid processing time during training and testing. Doing this doubled the refined dataset, but not necessarily meaningfully. A flaw in the dataset is the presence of multiple frames from the same video, and as discussed later, may contribute to overfitting and training the models on false features.

4. Baseline Method

There is one baseline of interest [10]. Kosti et. al aims to combine both bounded subject and scene context to determine the presence of various emotional states in a person. To achieve this, a bounding boxed subject is extracted from a scene, and is passed into a CNN. The scene (including the subject) is also passed into a separate CNN. Features are extracted from both images, and fed into a dense NN. The key idea being that context, and not solely subject, are important in properly understanding the emotional state of the subject.

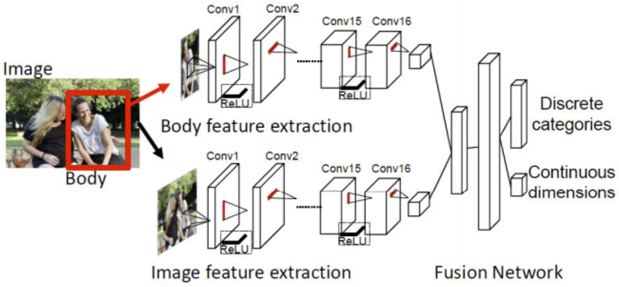

Figure 1. Kosti et. al's proposed end-to-end model for emotion recognition in context

Kosti et. al's method takes into consideration the context of the subject, along with the subject itself. This method may perform well for humans, but there is no indication of pose or focus on domain specific features (such as the tail or ears of a dog).

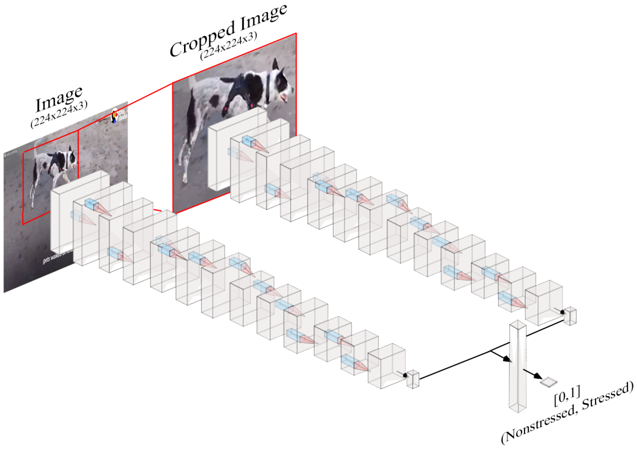

Figure 2. An implementation of Kosti et. al's proposed end-to-end model for the recognition of the presence of stress in a dog

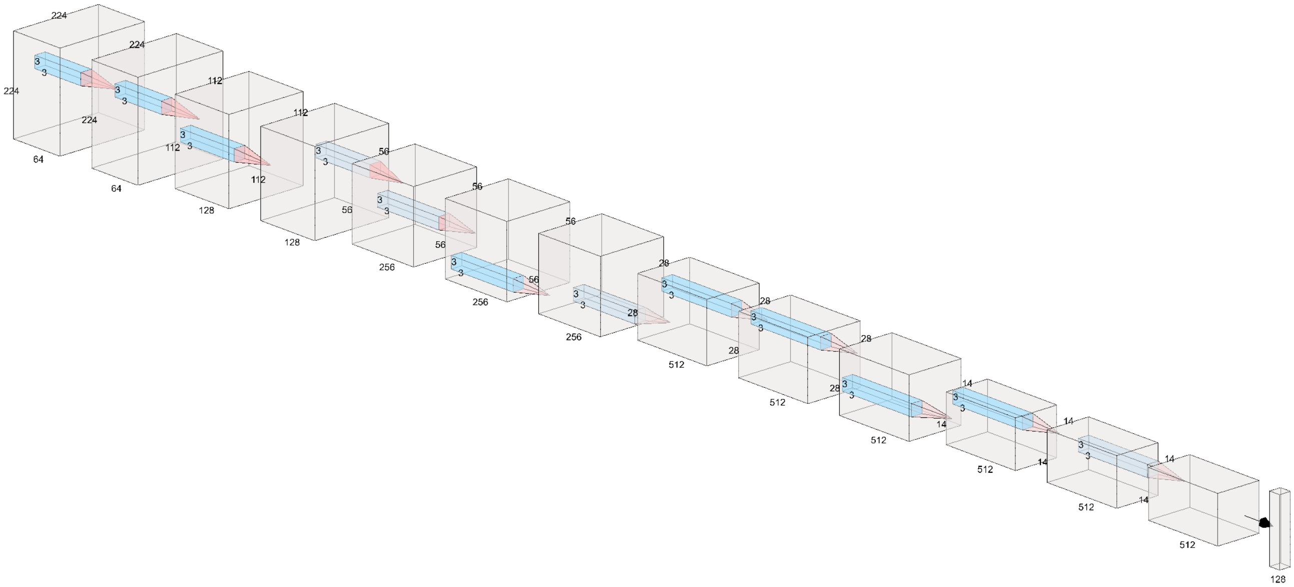

Figure 3. The structure of the CNN layer used in our method and baseline implementation.

5. Proposed Method

5.1. Overview

We propose a model in which pose and visual feature data are combined to achieve a higher quality result. We first derive pose data from our images using DeepLabCut [15] with a pretrained model from DeepLabCut's Model-Zoo [6].

Along with the pose data, we extract a bounded image of the dog present in an image (created in the preprocessing steps) using the python library ImageAI [17]. We then pass the image cropped to the bounding box to a Convolutional Neural Network. After the features are extracted from the CNN, the pose data is combined with the image feature data and passed into a dense neural network. The output of this dense layer is a 0 or 1, or unstressed and stressed, respectively.

A benefit to this model is that we will leverage the pose data, so as not to rely on pose features being learned by the convolutional model. This allows our feature detector to be trained on areas relevant to the presence of stress, such as a low hanging tail or ears close to the head. Of course, the combination of pose and image data is leveraged so that nuances between tail/ear position are recognized to a greater extent. The position of the tail may mean something different depending on the stance of the dog.

We focus only on the bounded image of the dog, as the presence of context might not carry as much weight as it does for humans, hence Kosti et al's intention to include it. Because dogs are domestic, we expect to both playful and stressed dogs in homes. When a dog is outside, this could imply that it is being engaged with, perhaps at a dog park, or being rescued from an abusive owner. Including the entire context will do more to add noise than provide insight, especially given the size of our dataset. If there is useful information to be gained from context, we are optimistic that the context available in a cropped image is suitable.

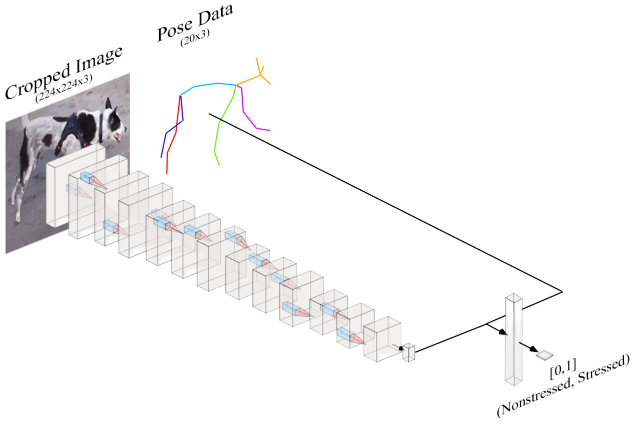

Figure 4. Our proposed model to detect the presence of stress in a dog. The bounded image of the dog is passed into a CNN, and the output of the CNN and the pose data are passed into a dense layer.

5.2. Input

As seen in 4, we see the bounded image of the dog being passed into a CNN, and its output, along with the pose data, are fed into a dense layer. The image data is provided as a 224x224x3 image, and the pose data is provided as 20 pose points, each composed of x and y positions, and a confidence value, totalling 60 values.

5.3. CNN

The CNN we used for both our method and the baseline implementation is shown in 3. There are clusters of layers with similar dimensions, with each cluster separated by a (2,2) max pooling layer (not shown). The clusters contain convolutional layers of the following dimensions: (224x224x64), (112x112x128), (56x56x256), (28x28x512), and (14x14x512). The output of the last layer of the last cluster is input to a (128x1x1) dense layer. All layers (besides the max pooling layers) have relu activation. In each convolutional layer, the kernel size is (3,3).

5.4. Output

The pose data and the output of the CNN are combined and used as input to a (512x1x1) dense layer with relu activation, which is connected to a (1x1x1) dense layer with sigmoid activation, which serves as our output.

6. Results

6.1. Our Method and Baseline

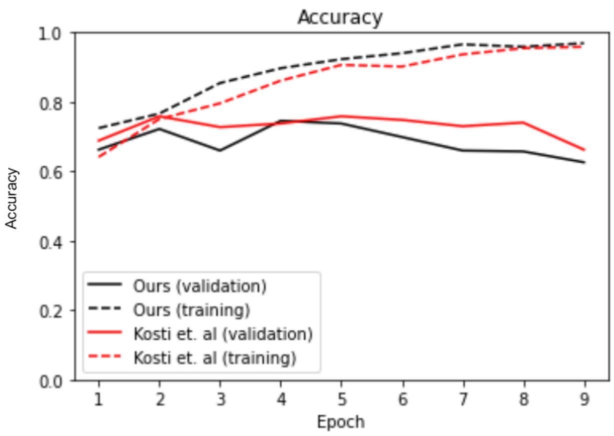

Figure 5. Accuracy of our method and our implementation of Kosti et. al over 10 training epochs.

For both models, we see heavy overfitting relatively quickly into the training phase. This suggests that the models have much greater discriminative capacity than the dataset is able to provide. Our model performed slightly worse than the Kosti et. al implementation. We can see in 5,6 that Kosti et. al has a slightly higher accuracy and lower loss throughout the training epochs. This, however, is within margin of error, so the models are comparable in overall performance. One possible reason for this is the presence of noise in the pose data. The pose data was found to be relatively low confidence, so introducing it into the model may have consequently introduced noise. It is also possible that pose (at a single time) does not provide enough information regarding the stress of a dog, or our dataset is too small and of too little quality to acquire pose data with high confidence.

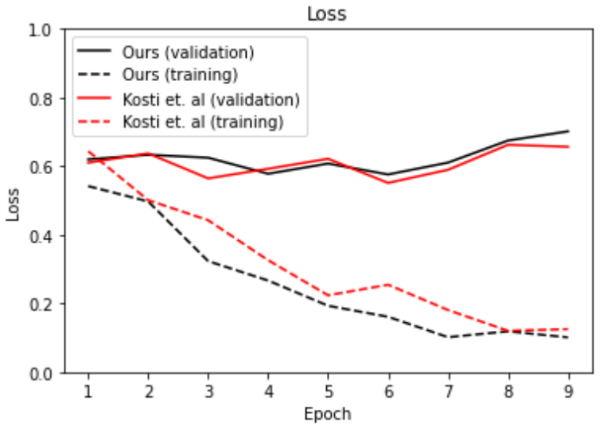

Figure 6. Loss of our method and our implementation of Kosti et. al over 10 training epochs.

Because many frames were extracted from each video we supplied, it was initially possible that the model was learning the backgrounds (or context) to determine the presence of stress, more than features of the dog. To negate this, we ensured that videos did not have frames in both the training and testing sets. With this in mind, we must investigate the impact of removing the pose data and testing only on the image data, and vice versa.

6.2. Ablation Study

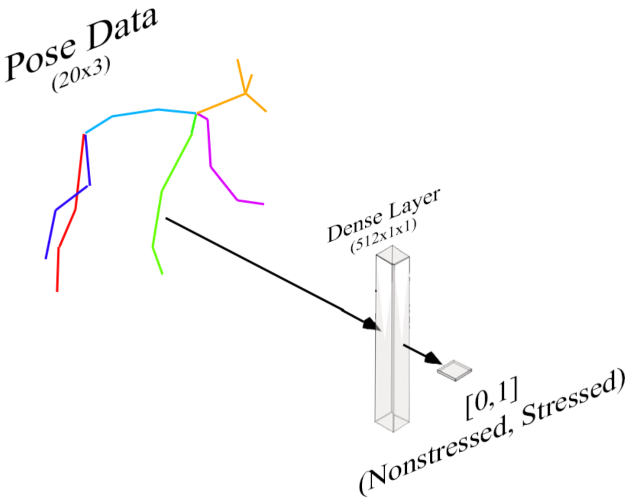

Figure 7. Our model trained on pose data only.

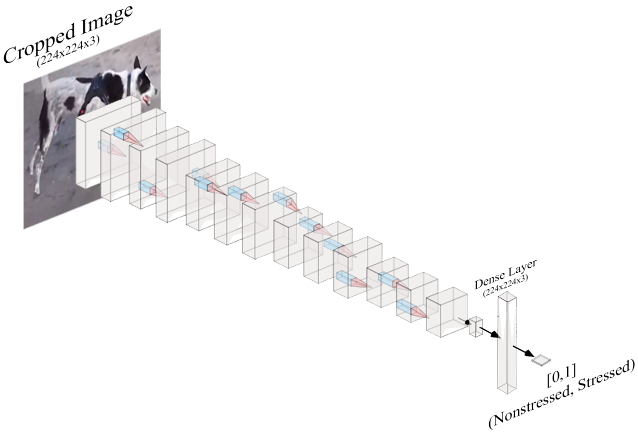

Figure 8. Our model trained on image data only.

6.3. Overall Results

Method	Accuracy	Loss
Ours	0.7714	0.5402
Kosti et. al	0.8130	0.5178
Pose only	0.6208	0.6802
Image only	0.7481	0.6087

Table 1. The best accuracy and loss for our method, Kosti et. al's method, pose only ablation, and image only ablation.

From 1, we see that the pose data by itself performed substantially worse than ours, Kosti et. al's, and the image only ablation. This agrees with our observation of low confidence pose data. Image only, however, is also less accurate than ours. While the pose data is noisy, in combination with the image data the model is better able to learn features associated with stress. Combining the bounded image of the dog with the context does appear to be superior. To determine how significant this is, we will need to acquire a larger, and higher quality, dataset.

Extracting features from the dog using the pose data is also of interest to us. For example, isolating the head/ears, and tail, and passing these into their own CNNs to later combine with the output of the others may yield better results. As it stands, the model must learn by itself that these areas of the dog are fairly clear markers of stress. Passing these in directly would allow the model to extract more nuanced features from the image of the dog, while layers that learn the pose-extracted features only need to focus on those domains.

7. Conclusion

With new emotion recognition technologies appearing for humans, the emotion recognition capabilities for animals does not follow. We propose a model that utilizes image and pose data of a dog to detect the presence of stress. We compare our model to Kosti et. al, and discover that using a bounded image of a dog along with the scene context yields a higher stress classification accuracy. It was found that the pose data used in our model was low confidence, mostly as a result of the relatively small and low quality dataset we collected. Improving the size and quality of this dataset would yield better pose data, which would allow us to more accurately detect the presence of stress in a dog. Going forward, we are also interested in using the pose data to extract domain specific images from the dog, such as the head/ears and tail. Doing this would allow the classifier to learn these directly, and not expend computation trying to learn that these domains are important, implicitly.

References

[1] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570. IEEE, 2016. 1, 2

[2] Hyun-Chul Choi and Se-Young Oh. Realtime facial expression recognition using active appearance model and multilayer perceptron. In *2006 SICE-ICASE International Joint Conference*, pages 5924–5927. IEEE, 2006. 2

[3] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. 17(2):124–129, 1971. 1

[4] Paul Ekman and Wallace V. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Palo Alto*, 1978. 1

[5] Paul Ekman and Wallace V. Friesen. Facial action coding system: The manual on cd rom. 2002. 1

[6] Mathis et. al. Deeplabcut modelzoo. `http://modelzoo.deeplabcut.org`. 3

[7] Yingruo Fan, Jacqueline C. K. Lam, and Victor O. K. Li. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 584–588, New York, NY, USA, 2018. Association for Computing Machinery. 2

[8] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 445–450, New York, NY, USA, 2016. Association for Computing Machinery. 2

[9] Otkrist Gupta, Dan Raviv, and Ramesh Raskar. Multi-velocity neural networks for facial expression recognition in videos. *IEEE transactions on affective computing*, 10(2):290–296, 2019. 2

[10] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1960–1968, 2017. 2

[11] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. 42(11):2755–2766, 2020. 2

[12] J. Li and E. Y. Lam. Facial expression recognition using deep neural networks. In *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2015. 2

[13] Zisheng Li, Jun ichi Imai, and M Kaneko. Facial-component-based bag of words and phog descriptor for facial expression recognition. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 1353–1358. IEEE, 2009. 1

[14] Xiaodong Liu and Miao Wang. Context-aware attention network for human emotion recognition in video. *Advances in multimedia*, 2020, 2020. 2

[15] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. 3

[16] A Mehrabian. Framework for a comprehensive description and measurement of emotional states. 121(3):339–361, 1995. 1

[17] Moses and John Olafenwa. Imageai, an open source python library built to empower developers to build applications and systems with self-contained computer vision capabilities, mar 2018–. 3

[18] Wenxuan Mou, Oya Celiktutan, and Hatice Gunes. Group-level arousal and valence recognition from static images: Face, body and context. 05 2015. 2

[19] M. A Nicolaou, H Gunes, and M Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE transactions on affective computing*, 2(2):92–105, 2011. 2

[20] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9):1238–1246, 2008. 2

[21] Marcello Siniscalchi, Serenella d'Ingeo, Michele Minunno, and Angelo Quaranta. Communication in dogs. *Animals (Basel)*, 8(8):131, 2018. 1

[22] Muhammad Usman, Siddique Latif, and Junaid Qadir. Using deep autoencoders for facial expression recognition. In *2017 13th International Conference on Emerging Technologies (ICET)*, pages 1–6. IEEE, 2017. 2

[23] Zhengyuan Yang, Amanda Kay, Yuncheng Li, Wendi Cross, and Jiebo Luo. Pose-based body language recognition for emotion and psychiatric symptom interpretation. 2020. 2

[24] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*, pages 425–442. Springer International Publishing, Cham, 2016. 2