



OPEN A segment-based framework for explainability in animal affective computing

Tali Boneh-Shitrit¹, Lauren Finka², Daniel S. Mills³, Stelio P. Luna⁴, Emanuella Dalla Costa⁵, Anna Zamansky^{1✉} & Annika Bremhorst^{6,7}

Recent developments in animal motion tracking and pose recognition have revolutionized the study of animal behavior. More recent efforts extend beyond tracking towards affect recognition using facial and body language analysis, with far-reaching applications in animal welfare and health. Deep learning models are the most commonly used in this context. However, their “black box” nature poses a significant challenge to explainability, which is vital for building trust and encouraging adoption among researchers. Despite its importance, the field of explainability and its quantification remains under-explored. Saliency maps are among the most widely used methods for explainability, where each pixel is assigned a significance level indicating its relevance to the neural network’s decision. Although these maps are frequently used in research, they are predominantly applied qualitatively, with limited methods for quantitatively analyzing them or identifying the most suitable method for a specific task. In this paper, we propose a framework aimed at enhancing explainability in the field of animal affective computing. Assuming the availability of a classifier for a specific affective state and the ability to generate saliency maps, our approach focuses on evaluating and comparing visual explanations by emphasizing the importance of meaningful semantic parts captured as segments, which are thought to be closely linked to behavioral indicators of affective states. Furthermore, our approach introduces a quantitative scoring mechanism to assess how well the saliency maps generated by a given classifier align with predefined semantic regions. This scoring system allows for systematic, measurable comparisons of different pipelines in terms of their visual explanations within animal affective computing. Such a metric can serve as a quality indicator when developing classifiers for known biologically relevant segments or help researchers assess whether a classifier is using expected meaningful regions when exploring new potential indicators. We evaluated the framework using three datasets focused on cat and horse pain and dog emotions. Across all datasets, the generated explanations consistently revealed that the eye area is the most significant feature for the classifiers. These results highlight the potential of the explainability frameworks such as the suggested one to uncover new insights into how machines ‘see’ animal affective states.

In psychology, affect refers to the fundamental experience of feelings, emotions, attachment, or mood¹. Affective computing is a broad interdisciplinary research field that integrates computer science, psychology, physiology, and neuroscience to computationally model, monitor, and classify emotions and affective states^{2–4}. Recent trends in the field emphasize multi-modal emotion recognition^{5–7} and advancements that enable real-time processing and deployment of these models^{8,9}. The identification and measurement of affective states is particularly challenging due to their internal nature, and in the case of animals, the absence of verbal communication makes this even more difficult. However, since all mammals are known to produce facial expressions that convey affective information related to pain and emotions¹⁰, detecting subtle changes in these expressions presents a promising non-invasive approach for studying animal affective states. While studies on animal behavior have traditionally lagged behind those focused on humans in terms of AI and automated behavior analysis, this gap is starting to narrow. This progress is largely driven by advancements in deep learning platforms such as DeepLabCut¹¹, EZtrack¹², Blyzer¹³, LEAP¹⁴, DeepPoseKit¹⁵, and idtracker.ai¹⁶. These platforms are specifically designed for

¹Information Systems Department, University of Haifa, Haifa, Israel. ²Cats Protection, National Cat Centre, Chelwood Gate, Sussex, UK. ³School of Life&Environmental Sciences, Joseph Bank Laboratories, University of Lincoln, Lincoln, UK. ⁴School of Veterinary Medicine and Animal Science, São Paulo State University (Unesp), São Paulo, Brazil. ⁵Department of Veterinary Medicine and Animal Sciences, University of Milan, Milan, Italy. ⁶Dogs and Science, Bern, Switzerland. ⁷Department for Clinical Veterinary Science, Vetsuisse Faculty, University of Bern, Bern, Switzerland. ✉email: annazam@is.haifa.ac.il

tracking animal movement and posture recognition, but their application to the study of facial expressions remains underexplored. The number of studies addressing the automation of animal affect recognition tasks is, however, increasing significantly. In a comprehensive review by Broome et al.¹⁷ the majority of the studies reviewed there employ deep learning techniques. It is not surprising, as these techniques are adept at successfully extracting high-level contributing features from data¹⁸ and have demonstrated superior performance compared to other machine learning methods in most domains of affective analysis computing¹⁹. Despite their success, deep learning techniques have a major limitation: their complexity leads them to function as what is commonly referred to as “black-box reasoning.” The intricate, non-linear structure of deep neural networks makes it challenging to break them into intuitive and easily interpretable components, complicating the understanding of their decision-making processes. This limitation can lead to skepticism among researchers, veterinarians, animal experts, and other stakeholders, making them hesitant to rely on the models’ results. Moreover, researchers seek to gain deeper insights from these models and are often unsatisfied with merely knowing what the model predicts rather than why it predicts it. This highlights the need for methods and techniques to “open the black box” and provide a clearer understanding of how deep learning models arrive at their conclusions.

In an effort to gain a deeper understanding of deep learning techniques, eXplainable Artificial Intelligence (XAI) techniques have emerged as a valuable tool²⁰. Visual explanations are a widely-used approach in XAI^{20–25}, offering insights into a deep learning model’s decision-making process through visualizations. A common method is *saliency maps*, where each pixel’s value reflects its importance to the model’s output. Prominent techniques in this category include Class Activation Maps (CAM)²⁶ and its advanced versions like Grad-CAM²⁷ and Grad-CAM++²⁸. However, these methods also have limitations, including subjectivity and inconsistency. Different techniques might produce different explanations for the same input, leading to confusion about which one is correct²⁹. Moreover, individuals may interpret the same visualization differently due to cognitive biases. For example, confirmation bias can cause people to focus on areas of the heatmap that match their expectations or hypotheses. Prior experience also plays a role, guiding attention to certain regions. Furthermore, the resolution of the heatmap can impact interpretation. In high-resolution maps, for instance, some individuals may zoom in on specific details, perceiving small intensity variations as more important^{30,31}. These problems combined with the lack of standardization and quantitative metrics makes it difficult to compare different models or explanations^{32–34}. Studies addressing these challenges have proposed techniques for quantitatively evaluating and comparing different saliency maps, primarily focusing on general classification^{32,33,35,36}.

In animal affective computing, XAI techniques are just beginning to be explored and no domain-specific methods have yet been introduced in this field. Saliency maps are only qualitatively used in Boneh-Shitrit et al.³⁷ and Broome et al.³⁸. Feighelestein et al.³⁹ was the first to present a quantitative approach in this domain by computing average heat of facial landmarks in cat facial analysis for pain detection. However, facial landmarks represent only specific points on the face and considering their heat only may not be sufficiently informative. Animal affective computing, particularly in the context of automated facial analysis, currently lacks structured XAI approaches that are tailored to its specific needs. Such approaches could help relate deep learning reasoning to concepts grounded in behavioral meanings, drawing inspiration from frameworks like AnimalFACS^{40–42}. Additionally, the field is lacking frameworks for comparing XAI techniques and clear guidelines for selecting the most appropriate approach for specific domains or cases.

This paper takes a step towards systematization of visual XAI approaches for animal affective computing. To this end, we make the assumption that in this domain, explanations are closely related to specific semantic segments of interest which constitute facial or body parts of interest. In other words, we can ‘explain’ classifications in this domain in terms of importance of specific body parts (e.g., ‘ears are most important’, ‘mouth is less important’). Of course, each classification task can induce its own domain-specific (and species-specific) set of such segments. Visual ‘explanations’ are expressed in terms of their importance for classification which can be extracted from saliency maps. For instance, for the task of cat pain recognition, ears, eyes and mouth may constitute three semantic segments of interest, as they are anchored in the Feline Grimace Scale commonly used by human experts for assessing pain in cats⁴³. Therefore, we expect explanations to be grounded in these segments of interest. This view also allows for the introduction of ‘heatmap quality metric’ which considers how much of the ‘heat’ is focused on the semantic segments (as opposed to the full animal body or face). The proposed framework is generic in the sense that it can be applied with different classifier architectures and any heatmap extraction technique that is applicable for that architecture and can be converted into a probability map. Our method uses the probability maps to calculate a ‘normalized grade’ for each relevant segment. This grade reflects the importance of the segment for the classifier (grades lower than 1 signify low importance). This approach enables the quantification of the relative importance of each relevant segment, and/or their different combinations in the classification. Moreover, by using ‘normalized grade’ metrics, it also allows for the comparison between different classifiers, segmentation methods and heatmap generation methods.

Figure 1 demonstrates the core idea behind the proposed concept of heatmap quality: high-quality heatmaps, as shown in column B, are more focused on biologically meaningful regions, whereas the low-quality heatmaps in column C tend to lack focus or emphasize less relevant areas.

In this study we evaluate the proposed framework using three case studies related to facial expressions analysis in different species: (i) cat pain recognition⁴⁴, (ii) horse pain recognition⁴⁵ and (iii) emotion recognition in dogs⁴⁶. In these case studies, we compare various classifier architectures and saliency maps algorithms within our framework, concluding that ViT⁴⁷ pretrained with DINO⁴⁸ weights combined with Grad-CAM++²⁸ provides better quality explanations in all of these three cases.

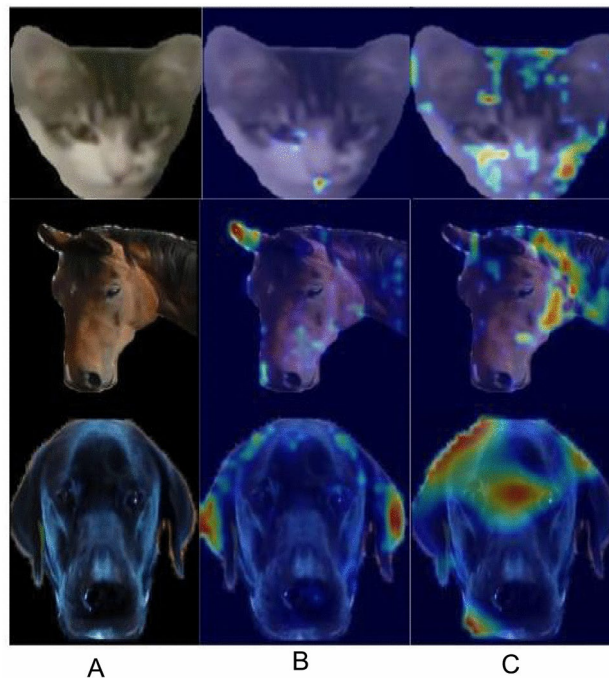


Fig. 1. Quality of Heatmaps. Column A shows the original image after background masking; Column B shows higher quality heatmaps focusing on facial parts such as ears and mouth; Column C shows lower quality heatmaps which are unfocused or focus on less relevant facial areas such as neck or back of the head.

Results

The explainability framework

We propose a conceptual framework for generating explanations to animal affective computing, focusing on specific semantic regions like facial or body parts. In this framework, explanations are based on evaluating the contribution of each semantic part to the classification decision. To simplify, we center our analysis on image-based tasks, although extending to video is feasible. Each image is assumed to depict an animal with k semantic parts which are considered relevant for the classification task. We consider a pre-trained classifier with satisfactory performance on the task and aim to explore its explainability. In addition, for each image I , we assume the ability to obtain segments $\{S_1^I, \dots, S_k^I\}$ representing the semantic parts, as well as S_{full}^I , which corresponds to the entire animal or a specific whole body region (e.g., the face without background). Each segment is assumed to have an associated mask $mask(S_i^I)$, marking the pixel locations of the segment. Lastly, we suppose a technique that is chosen for producing a heatmap $H(I)$ for each image I . It should be noted that saliency maps are assumed to contain only non-negative values, which can be ensured through rescaling or the application of the ReLU function. The heatmaps are then converted into probability maps by dividing each element by the sum of all elements in the map. This conversion provides a more intuitive and interpretable representation of the maps.

Figure 2 provides a high-level overview of the suggested framework and the interplay between the above mentioned elements. The semantic parts of interest (such as ears or eyes) are extracted as segments, and their corresponding masks are applied to heatmaps extracted from the classifier. The framework has two outputs: explanations in the form of relative importance of each segment of interest, and quality of heatmaps assessing the extent to which the classifier relied on areas within the examined segments in comparison to regions outside of these segments. These two outputs enable a novel quantifiable, statistical explainability of the classification outcomes, providing insights into which semantic part was more ‘informative’ for the classifier.

An ‘explanation’ in our framework is a segment importance metric, quantifying the importance of the particular segment for the classification decision. This is determined by calculating the probability for each segment in every image within our test set. Since semantic segments may occupy different relative space (e.g., eyes are much smaller than ears), the probability is divided by the segment’s relative size to obtain a ‘normalized grade’. Images where the normalized grade of a segment exceeds 1, are considered informative with respect to the segment, indicating the segment’s contribution to the classifier’s decision. The overall quality of the segment is then computed by multiplying the percentage of images where the segment was informative by the average of the informative grades (those greater than one). This approach balances the importance of having a high proportion of images with strong grades while also favoring higher-grade segments.

Another important aspect of our framework is the heatmap quality metric. Researchers often face challenges in interpreting their classifiers, especially when comparing different backbones for classification. They seek to understand their classifier’s decisions in relation to human expert insights or to gain new perspectives. While visualization methods like CAM-based heatmaps are frequently used, they lack quantifiable metrics that help

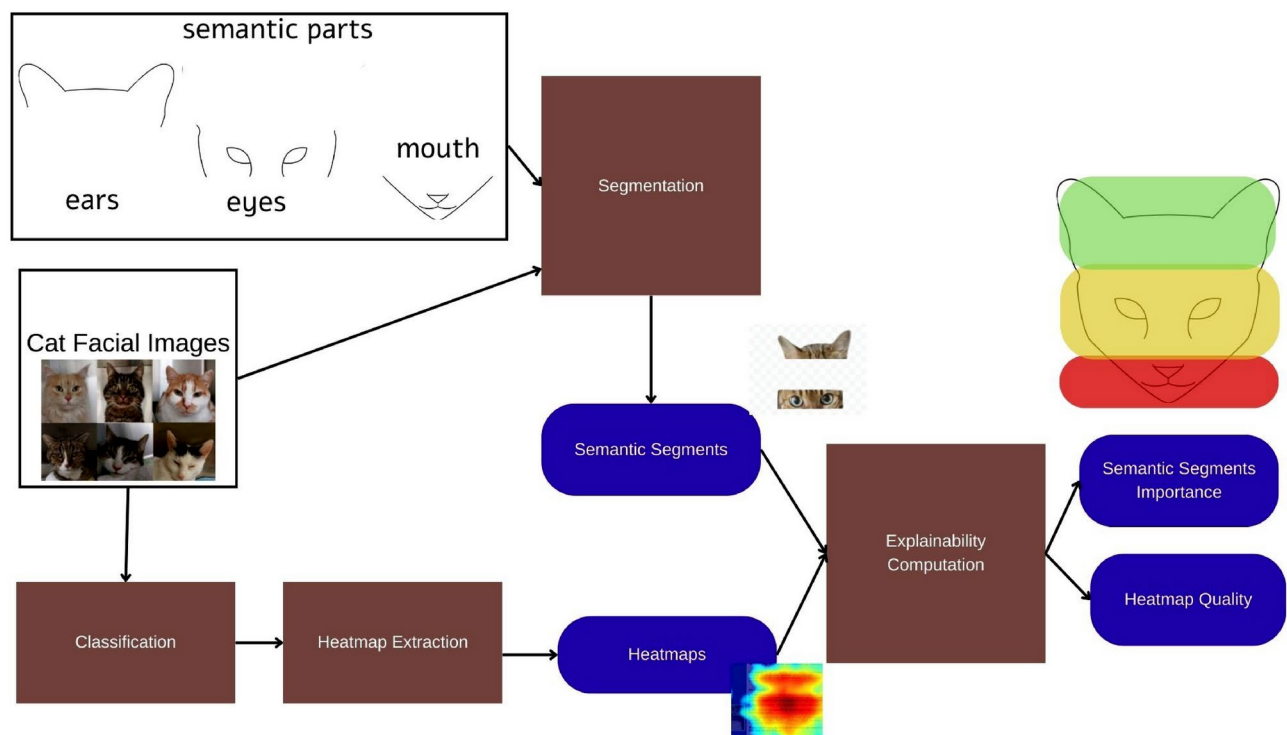


Fig. 2. Suggested Framework: a high-level overview.

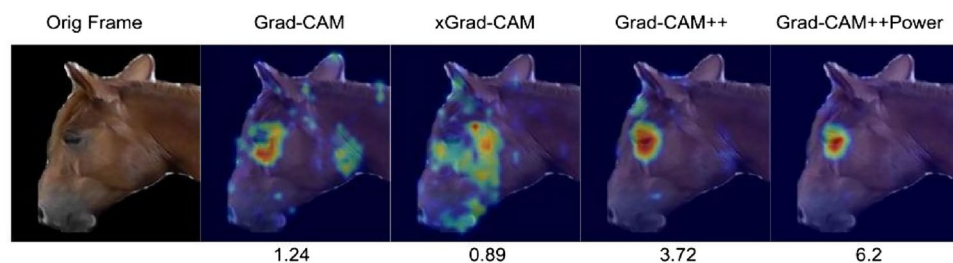


Fig. 3. Sample image with quality grades for different heatmaps options: a higher quality grade reflects a heatmap more focused on a meaningful area (eye).

researchers understand how their classifiers perform across datasets. We introduce a heatmap quality metric that enables a more systematic comparison of different architectures and saliency map techniques. We assign a quality grade to a specific heatmap type by averaging the quality scores of all its segments. Higher scores indicate that the heatmap type more effectively focuses on biologically relevant segments compared to other heatmap algorithms. We use the term ‘quality’ since different saliency map methods can generate heatmaps with varying focal points, we consider a method of higher quality if it aligns better with our tested segments. As researchers, we want to avoid mistakenly attributing poor performance to a classifier simply due to selecting an inappropriate heatmap generation method. The framework assists us in choosing the most suitable method for our task.

Figure 3 presents an example image with various heatmaps and their corresponding quality scores for the image in question. Among them, GRAD-CAM++²⁸ exhibits a more focused heatmap over biologically meaningful segments, particularly around the eyes, as compared to GRAD-CAM²⁷ and xGRAD-CAM⁴⁹, resulting in a higher score. Additionally, after applying a power transform (using a factor of 2) to the GRAD-CAM++²⁸ heatmap, the concentration on the eye region becomes even stronger, further improving the quality score.

Case studies

We demonstrate the proposed framework on three case studies related to facial expression analysis in cats, horses and dogs, using previously generated datasets and analyzing them in a new way. All three datasets underwent similar preprocessing stages, including semantic segmentation and background masking. We compare different classifiers and saliency maps techniques by extracting heatmap quality grades, and generate explanations in the form of segment significance for these various combinations.



Fig. 4. Example frames extracted from the three datasets.

Datasets description

The Cat Pain dataset we use was originally generated in Finka et al.⁴⁴. Frames were extracted from footage captured from healthy mixed-breed (domestic short hair) female cats, undergoing ovariohysterectomy. The dataset is balanced, containing 450 images obtained from 29 subjects, half of which are labeled with 'pain' state and the other half labeled with 'no pain'. Based on the previous studies with this dataset^{44,50}, we set facial parts of ears, eyes and mouth as the semantic parts of interest in our study.

The Horse Pain dataset was generated in Dalla Costa et al.⁴⁵. Frames were extracted from footage captured from healthy thirty-nine horses undergoing a routine castration procedure. The dataset is balanced, containing a total of 126 images of horses (63 pre-surgery and 63 post-surgery) labeled with 'pain'/'no pain'. Similarly to the previous case, we set ears, eyes and mouth (muzzle) as the semantic parts of interest.

The Dog Emotion dataset, developed by Bremhorst et al.⁴⁶, comprises recordings of 29 *Labrador Retriever* dogs, totaling 248 videos, each lasting approximately 3 seconds. These recordings were conducted in a controlled laboratory setting to induce two emotional states: positive (anticipation of a food reward) and negative (frustration due to the reward being inaccessible), with each video labeled accordingly. Approximately two-thirds of the videos were labeled as negative, while one-third were labeled as positive. For our study, frames were sampled from the videos, resulting in around 75 images per video.

Ethical statements

All experiments were performed in accordance with relevant guidelines and regulations.

The dog dataset was collected previously under the ethical approval of the University of Lincoln, (UID: CoSREC252).

The cat dataset was collected previously under the following ethical approvals of the Institutional Animal Research Ethical Committee of the FMVZ-UNESP-Botucatu (protocol number of 20/2008) and the University of Lincoln, (UID: CoSREC252).

The horse dataset was collected in a previous study registered as an animal experiment at the Brandenburg State Veterinary Authority (V3-2347-A-42-1-2012). Castration is a routinely conducted husbandry procedure that was carried out in compliance with the European Communities Council Directive of 24 November 1986 (No. 86/609/EEC). Horses involved in this study underwent routine veterinary procedures for health or husbandry purposes at the request of their owner on a voluntary basis. Consequently, no animals underwent anaesthesia or surgery or were directly used in order to record data for the purposes of this study. Verbal informed consent was gained from each participant prior to taking part in this research. Written consent was deemed unnecessary as no personal details of the participants were recorded. No animals received less than the standard analgesic regimen for the purposes of the study. The study employed a strict "rescue" analgesia policy: if any animal was deemed to be in greater than mild pain (assessed live by an independent veterinarian), then additional, pain relieving medication would immediately be administered and the animal removed from the study. The choice of medication and dosage would be based on the severity of pain identified thorough the clinical examination of the individual horse.

The current protocol using these datasets was further reviewed by the Ethical Committee of the University of Haifa and no further approval was required.

Experimental results

Table 1 presents performance metrics for the different classifiers developed for each case study. In two out of the three tasks (cat and horse pain recognition), the Vision Transformer (ViT) initialized with DINO weights⁴⁸ was the top-performing classifier, with Google's NesT-tiny⁵¹ model ranking second. For dog emotion recognition, however, NesT-tiny outperformed ViT-DINO as a classifier. It should be noted that these were 'vanilla' classifiers, as improving classifier performance was not the focus in this work. We are certain that domain-specific and species-specific improvements can be performed and leave that for future work.

Nevertheless, our best-performing model achieved higher accuracy compared to previous works on the cats and dog datasets and achieved comparable results on the horses dataset. For example, in the cat pain dataset, previously studied by⁵⁰, the authors employed a ResNet50⁵² with an additional subnetwork replacing its head for classification. The images were manually annotated with 48 landmarks per image and underwent an alignment preprocessing stage, with the best reported accuracy approximately 0.73. In comparison, we obtained an accuracy of 0.86 using a ViT pretrained with DINO weights⁴⁸. Similarly, for the dog emotional states dataset, analyzed in³⁷, an accuracy of 0.85 was reported using a ViT pretrained with DINO weights⁴⁸, which aligns with the results we achieved with this model. However, using Google's NesT-tiny model⁵¹, we surpassed this, reaching an accuracy of 0.89. It is important to highlight that our data handling differed from that in³⁷. The

Dataset	Backbone	Acc	Prec	Rec	F1
Cat pain	ResNet50	0.69	0.68	0.72	0.7
Cat pain	ViT-supervised	0.5	0.49	0.12	0.19
Cat pain	ViT-Dino	0.86	0.85	0.87	0.86
Cat pain	NesT-tiny	0.74	0.72	0.74	0.74
Horse pain	ResNet50	0.54	0.54	0.49	0.52
Horse pain	ViT-supervised	0.42	0.41	0.33	0.37
Horse pain	ViT-Dino	0.71	0.72	0.68	0.7
Horse pain	NesT-tiny	0.7	0.76	0.6	0.67
Dog emotion	ResNet50	0.85	0.83	0.96	0.89
Dog emotions	ViT-supervised	0.81	0.86	83	0.85
Dog emotions	ViT-Dino	0.85	0.83	0.96	0.83
Dog emotions	NesT-tiny	0.89	0.85	0.95	0.9

Table 1. Comparison of performance of different backbones on the three classification tasks.

previous study did not mask the dog’s background, excluded certain videos to create a balanced dataset, and employed a Leave-One-Animal-Out training approach. In contrast, our study used masked images of the dogs’ faces and utilized all available videos, assigning videos from 6 dogs for validation and using the remaining 23 for training. This methodological difference may account for discrepancies observed in the performance of other models compared across both studies. For example, our study recorded accuracies of 0.85 for ResNet50⁵² and 0.81 for a supervised ViT⁴⁷, whereas³⁷ reported 0.81 for ResNet50 and 0.82 for the supervised ViT. The relatively larger difference observed in ResNet50 performance suggests that it may be less stable and more sensitive to data variations compared to the ViT architecture. While the work on the horse pain dataset is yet unpublished, it reports an accuracy of 0.73 using Dino-v2⁵³ embeddings combined with an NU-SVM⁵⁴, which is similar to our 0.71 achieved with the Dino-ViT⁴⁸ model. Additionally, the authors developed a model that regresses embeddings to Facial Action Unit (FAU) scores, achieving 0.79 accuracy. However, this approach introduces an additional layer of complexity, as it requires an FAU decoding step and verification that the FAUs are correctly classified.

Heatmap quality scores for different combinations of classifiers and heatmap types is shown in Fig. 5. Across all datasets, the best performance is observed with the ViT⁴⁷ pre-trained using DINO weights⁴⁸ combined with Grad-CAM++²⁸, with further improvement when a power transform is applied. In most cases, the second-best performer is Google’s NesT-tiny⁵¹, which delivers consistent quality across all heatmap types, also benefiting from the power transform.

As depicted in Fig. 6, the eyes consistently emerge as the most significant feature across all three datasets when assessing segment importance. For cats and dogs, the mouth and ears follow in importance, while in the horse pain dataset, the ears rank second, followed by the mouth. However, the ratings for the mouth and ears are relatively close, unlike the eyes, which are clearly the dominant feature. Additionally, the heatmap for the horse dataset emphasizes the eyes more than the other datasets, while the mouth receives a noticeably lower rating, indicating it holds less significance in this context.

Figure 7 presents the saliency maps generated by various CAM based algorithms for the ViT-DINO classifier, which achieved the highest quality score. Grad-CAM++²⁸ demonstrates superior localization of the relevant facial parts compared to Grad-CAM²⁷ and xGrad-CAM⁴⁹. Additionally, applying a power transform to the Grad-CAM++²⁸ map enhances its clarity, making the highlighted features more distinct.

Discussion

We have presented a framework for explainability that generates explanations by highlighting the importance of meaningful semantic elements for classification. To evaluate this framework, we trained classifiers using different backbones across three classification tasks: cat pain recognition, horse pain recognition, and dog emotion recognition. It’s important to note that these were “vanilla” classifiers, as improving classifier performance was not the primary focus of this work, and we anticipate that further improvements could be achieved with additional refinements.

Across the three case studies, the Vision Transformer (ViT)⁴⁷ initialized with DINO weights⁴⁸ consistently delivered the best performance. In two out of the three tasks (cat and horse pain recognition), it was the top-performing classifier, with Google’s NesT-tiny⁵¹ model ranking second. For dog emotion recognition, however, NesT-tiny⁵¹ outperformed ViT-DINO⁴⁸ as a classifier, although the ViT-DINO⁴⁸ heatmaps demonstrated superior quality. The Google’s NesT-tiny⁵¹ model produced heatmaps with consistent quality across different methods, with a notable improvement observed when a power transform was applied, making it a reliable option. Among the various heatmap generation methods we tested, Grad-CAM++²⁸ consistently yielded the best results in all scenarios. Its key strengths include improved localization and more precise attribution of class predictions to specific image regions, which are valuable for our tasks. Although previous comparisons by the authors of xGrad-CAM⁴⁹ have shown that xGrad-CAM⁴⁹ outperforms both Grad-CAM²⁷ and Grad-CAM++²⁸ in terms of visualization quality, our experiments revealed that xGrad-CAM⁴⁹ produced lower-quality results in

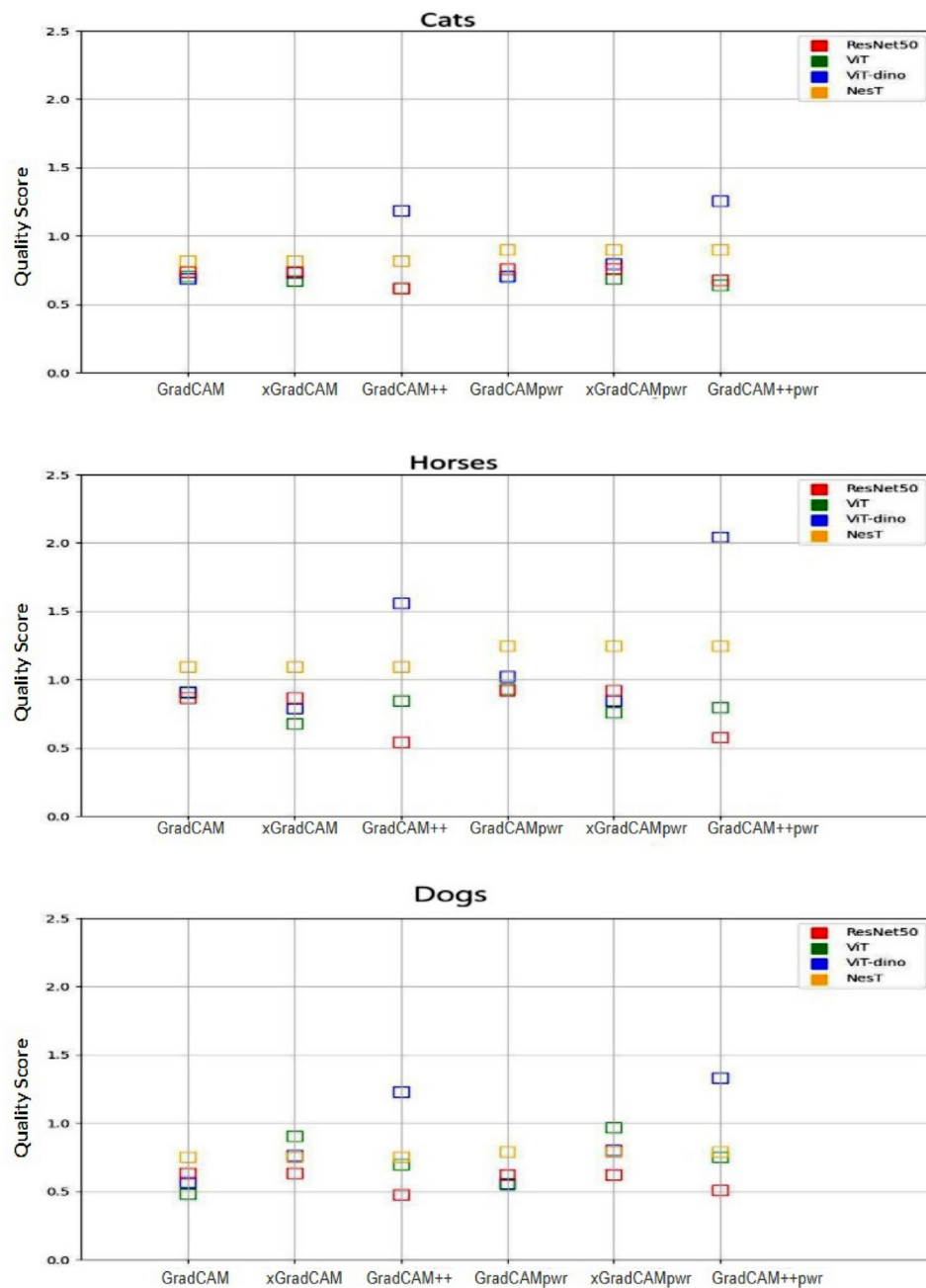


Fig. 5. Quality grades of maps and segments.

our benchmarking. These findings underscore the importance of selecting the most appropriate visualization technique based on the specific task.

In terms of explanations, the eye area consistently emerged as the most important across all datasets. The second and third most significant areas varied between datasets. For cats and dogs, the mouth ranks second, followed by the ears, while for horses, the order is reversed, with the ears being more important than the mouth. In the case of the ViT-DINO⁴⁸ Grad-CAM++²⁸ combination, the eyes are not only the most important feature, but the gap between the eyes' quality score and other areas is more pronounced compared to other classifiers. This difference becomes even more evident after applying a power transform to this combination heatmaps. In other classifiers and heatmap techniques, the power transform also increases the distinction between segment grades, but not to the same extent as it does for ViT-DINO⁴⁸ with Grad-CAM++²⁸. Despite these differences, the order of significance among the segments remains consistent across various heatmap methods used in our study. It is important to note the findings of³⁹, which addressed the explainability of the cat pain dataset by averaging Grad-CAM²⁷ heatmaps over facial landmarks using a ResNet50⁵² classifier. In their study, the mouth was identified as the most significant area, followed by the eyes and ears. This discrepancy in results can be attributed to differences in training processes (such as image preprocessing and parameters), leading the

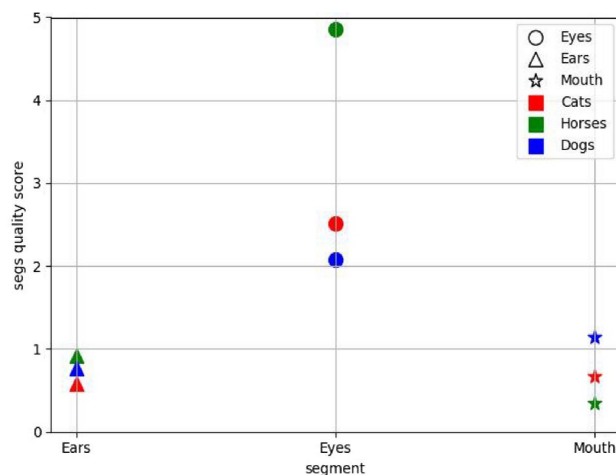


Fig. 6. Quality Grade of segments over the datasets for the top combination.

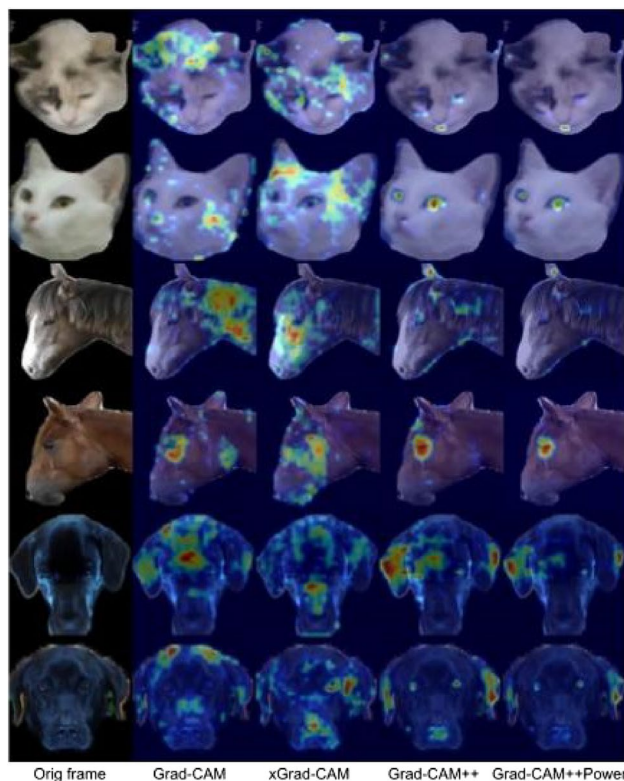


Fig. 7. Grade maps generated by different CAM algorithms for ViT-Dino classifier Grad-CAM++²⁸ shows better localization on relevant facial parts compared to Grad-CAM²⁷ and xGrad-CAM⁴⁹. Implying power transform on Grad-CAM++²⁸ map (using factor of 2), makes it more distinct.

classifier to focus on different facial regions. Additionally, the difference in approaches to explainability plays a role: while³⁹ focused on specific landmarks, this work analyzes entire segments. It is possible that the classifier focused on a portion of the ear that does not necessarily correspond to a landmark. We opted to focus on biologically significant segments as a whole, leaving the exploration of specific regions within those segments for future research. Further investigation into selecting the best combination of deep neural networks and heatmaps could benefit from collaboration with animal behavior experts. Such cooperation could help quantify the desired significance of each segment in an image, which could drive improvements in both classifiers and visualization techniques, ultimately allowing users to trust and interpret the system's output more effectively.

In tasks where well-established biological concepts inform the classification of animal affective states, aligning model behavior with expert knowledge is particularly important. For instance, the validated Feline Grimace

Scale⁴³ identifies the eyes, ears, and mouth as key indicators of pain and emotional states. Demonstrating that a deep learning model leverages closely related biological concepts can help increase confidence among end-users, such as veterinarians and animal behavior specialists. Importantly, alignment does not imply that the model must replicate human attention patterns-experts may prioritize the ears over the eyes, for example-but rather that it should rely on the same underlying biological cues. In such cases, we expect the model's heatmaps to emphasize these meaningful regions. Conversely, when the objective is to discover novel indicators of affective states, the proposed framework can facilitate exploration. By testing candidate regions and quantifying their significance, the framework allows researchers to assess how strongly the model depends on these regions for its predictions. This can provide valuable insights into previously unrecognized behavioral markers. Overall, the proposed framework supports both the validation of established biological concepts and the discovery of new behavioral indicators. Furthermore, it provides a systematic approach to comparing different model architectures and saliency map generation methods. By bridging deep learning techniques with domain expertise, this approach has the potential to advance explainability, enhance model interpretability, and foster trust in automated methods for animal affective computing.

Methods

The explainability framework

The following sections describe our approach to gathering the necessary data and illustrate how our framework supports the comparison of various classifier and heatmap combinations. This approach aids in identifying the most informative pairing of model and heatmap for our objectives. Each analyzed image is assumed to depict an animal with k semantic parts which are relevant for the classification task. We consider a pre-trained classifier with satisfactory performance on the task and aim to explore its explainability. In addition, for each image I , we assume the ability to obtain segments $\{S_1^I, \dots, S_k^I\}$ representing the semantic parts, as well as S_{full}^I , which corresponds to the entire animal or a specific whole body region (e.g., the face without background). Each segment is assumed to have an associated mask $mask(S_i^I)$, marking the pixel locations of the segment. Lastly, we suppose a technique that is chosen for producing a heatmap $H(I)$ for each image I , where each pixel of the image I is assigned a value indicating its significance to the classifier decision. It should be noted that saliency maps are assumed to contain only non-negative values, which can be ensured through rescaling or the application of the ReLU function. The heat maps are then converted into probability maps, each element is divided by the sum of all elements in the map. This conversion facilitates a more intuitive and interpretable representation of the maps.

Algorithm 1 provides the pseudo-code for the outlined calculations.

Normalized score

Semantic segments can vary greatly in relative size (e.g., an eye is much smaller than a tail), yet our focus is on their relative importance to classification rather than their absolute size. To ensure scale-invariant contributions, we expect that a meaningful segment with high importance will carry a probability greater than what would be assigned under a uniform distribution. To assess the relative importance of different areas, we compute a normalized score for each segment by dividing its total probability by its relative area within the face region of the image. This normalization allows us to evaluate each segment's relevance independently of its size. A score greater than one indicates that the segment provides valuable information for the classifier, demonstrating that the heatmap effectively highlights it. Conversely, segments with a normalized score below one contribute less than or equal to a uniform distribution and therefore cannot be considered significant. The issue of large regions being disproportionately emphasized in explanations has been addressed in prior work, such as⁵⁵, where the authors demonstrate that naive backpropagation-based explanations tend to highlight larger areas due to activation summation.

Semantic explanations

An 'explanation' in our framework is a segment importance metric, intuitively quantifying the importance of the particular segment for the classification decision. This is determined by calculating the normalized score for each segment in every image within our test set. Images where the normalized grade of a segment exceeds one are considered informative with respect to the segment, indicating the segment's contribution to the classifier's decision. The overall quality of the segment is then computed by multiplying the percentage of images where the segment was informative by the average of the informative grades (those greater than one). The first term represents the expected importance of the segment given that it was relevant. The second term represents the empirical probability that the segment is important across different images. Multiplying these two components ensures that both frequency and intensity contribute proportionally: a segment that is highly important but rarely active receives a low total grade, as does a segment that is frequently active but only weakly important. Conversely, a segment that is both frequently active and strongly important attains the highest grade. This prevents overemphasis on rare but extreme scores (which would happen in a simple mean) and avoids giving too much weight to common but weakly relevant segments.

Measuring quality of heatmaps

We also introduce a heatmap quality metric that enables systematic comparison of different architectures and saliency map techniques. Using our Semantic Explanations, we assign a quality grade to a specific heatmap type by averaging the quality scores of all its segments. Higher scores indicate that the heatmap type more effectively focuses on biologically relevant segments compared to other heatmap algorithms.

```

for each image  $I_i$  in images do
   $segsGrades \leftarrow []$ 
   $H(S_{full}^I) \leftarrow H(I) \times mask(S_{full}^I)$ 
   $probMap[I] \leftarrow \frac{H(S_{full}^I)}{\sum(H(S_{full}^I))}$ 
  for ( $mask(S_j^I)$  In masks) do
     $prob[S_j^I] \leftarrow \sum(probMap \times mask(S_j^I))$ 
     $normScore[S_j^I] \leftarrow \frac{prob(S_j^I)}{\sum mask(S_j^I)}$ 
     $append(segsGrades[S_j], normProb[S_j^I])$ 
  end for
end for
 $qualities \leftarrow []$ 
for each list of segment's grades  $SL_i$  in segsGrades do
   $highGrades \leftarrow SL_i[SL_i \geq 1]$ 
   $segQuality \leftarrow mean(highGrades) \times \frac{numel(highGrades)}{numel(SL_i)}$ 
   $append(qualities, segQuality)$ 
end for
 $sort(qualities)$ 
 $totalQuality = mean(qualities)$ 

```

▷ Elements of *segsGrades* are lists of *normProb* for each segment
 ▷ Convert the heatmap to probability map
 ▷ Go over all segments
 ▷ Accumulate probability of segment
 ▷ Calculate normalized score
 ▷ If $normProb[S_j^I] \geq 1$ then S_j^I is significant
 ▷ Get grades g.t 1
 ▷ Calculate segment quality grade
 ▷ Obtain the segments quality ranked by significance.

Algorithm 1. Segment quality and map type quality calculation

To assess the contribution of a specific image to the overall quality, we calculate the normalized probabilities of the segments of interest within that image and average their contributions. If a segment's normalized score is below 1, it does not contribute to the quality.

Experiments

We performed an ablation study comparing the explainability quality of different classifiers for the datasets, creating distinct grade maps for each classifier. We employed transfer learning to evaluate performance using ResNet50⁵², ViT⁴⁷, ViT with pretrained DINO⁴⁸ weights, and Google's NesT-tiny⁵¹ as backbones. We generated grade maps and calculated quality as described in Algorithm 1, based on Grad-CAM²⁷, Axiom-based Grad-CAM (XGrad-CAM)⁴⁹, and Grad-CAM++²⁸. These heatmaps were then compared after applying a power transform (using a factor of 2), which enhances the distinction between contributing and non-contributing pixels by amplifying high values and reducing low values. Our findings indicate that the highest quality is generally achieved with the combination of ViT with initial DINO⁴⁸ weights and Grad-CAM++²⁸ grade maps that underwent power transformation. Additionally, we found that the Google's NesT-tiny⁵¹ model provided consistent quality across all types of CAM algorithms, making it a robust choice for classification with a wide variety of saliency maps.

As shown by Fig. 2, every dataset underwent the following stages: (1) training the classifier, (2) classification, (3) segmentation, and (4) GradeMaps creation and quality calculation. For each task, we decided on the relevant semantic parts (e.g., ears, eyes and mouth in all of our case studies). The images then underwent segmentation, cutting the semantic parts out. Given a sufficiently well performing classifier, we extracted heatmaps from it. We used the heatmaps and the semantic parts to produce explanations (measuring the importance of each semantic part for the classification), and a quality metric of the heatmap.

Segmentation of the faces and facial parts in all of the three case studies was done by finetuned YOLOv8⁵⁶. For each dataset, YOLOv8⁵⁶ was trained to segment the face, and then the relevant semantic facial parts (eyes, ears and mouth in all the three cases).

We explored four different architectures: ResNet50⁵², ViT⁴⁷, ViT with pretrained DINO⁴⁸ weights, and Google's NesT-tiny⁵¹ as backbones. All classifiers were trained using facial images with masked background (masking was done automatically using the YOLO8 segmentation). We used different data augmentations such as color space and geometric transformations. The networks head was replaced to a 2 classes output head ('Pain/No pain' for cats and horses, 'positive'/'negative' for dogs). All backbones were trained using leave-one-out cross-validation⁵⁷, with no subject overlap for both the cat pain and horse pain datasets. The dog emotional state dataset was split into a training set, which included images sampled from videos of 23 dogs, and a validation set, containing images from videos of 6 dogs.

Dataset	Segment	Dataset size for YOLOv8	Dataset split - training	Input img size	#epochs
Cat pain	Face	54	0.85	800x800	10
Cat pain	Ears	54	0.85	224x224	25
Cat pain	Eyes	54	0.85	224x224	25
Cat pain	Mouth	54	0.85	224x224	25
Horse pain	Face	39	0.82	224x224	20
Horse pain	Ears	39	0.82	224x224	35
Horse pain	Eyes	39	0.82	224x224	35
Horse pain	Mouth	39	0.82	224x224	20
Dog emotion	Face	135	0.87	640x640	30
Dog emotions	Ears	145	0.87	224x224	10
Dog emotions	Eyes	145	0.87	224x224	10
Dog emotions	Mouth	145	0.87	224x224	20

Table 2. Segmentation details for YOLOv8 training, including dataset size, training split percentage, and number of epochs for each segmented facial part across the three datasets.



Fig. 8. Example of the YOLO8 segmentation results.

We calculated the heatmaps quality metric for each backbone architecture, using Grad-CAM²⁷, Axiom-based Grad-CAM (xGrad-CAM)⁴⁹, and Grad-CAM++²⁸. In addition we also used power transform on the grades maps for clearer separation between more and less contributing pixels.

We then proceeded to calculating ‘explanations’ for each configuration, i.e., measuring the segments’ importance in each case.

Segmentation

Segmentation was performed by fine-tuning yolov8s-seg, a version of YOLOv8⁵⁶ specifically trained for segmentation. First, we trained YOLOv8 to segment the animal’s face. In the cats and dogs dataset, most images contain the entire animal body, whereas in the horses dataset, images primarily show the face. However, since we needed to separate the face from the background, we trained YOLOv8⁵⁶ to segment the face in this dataset as well. After this initial step, we cropped the detected face regions and retrained YOLOv8 to segment specific facial parts, creating a separate model for each part. The use of a separate model for each facial part enabled an optimization for the different segmented parts. Due to differences in image structure across datasets, the segmentation approach varied:

- **Cats and Dogs:** We trained separate models to segment a single ear, a single eye, and the mouth. During evaluation, the model was configured to detect up to two ears and two eyes per image.
- **Horses:** Since all horses in the dataset faced left, the images had a consistent structure. This allowed us to train one model to detect either one or both ears together, another model to segment the visible eye (only one eye is visible due to the positioning), and a third model to segment the muzzle.

Various augmentation techniques were applied during training to expand the dataset, such as adding noise, blurring, adjusting exposure and brightness, and applying rotation and shear transformations.

Table 2 summarizes the details for YOLOv8 training, including dataset size, training split percentage, and number of epochs.

Figure 8 presents an example of the segmentation results produced by YOLOv8.

The trained YOLOv8 models were utilized to preprocess all images in the datasets, generating cropped face images with masked backgrounds. These processed images served as input for the classifiers during both training and evaluation. For classifier evaluation within our framework, the models were employed to locate and segment specific facial parts in each image.

Training details for the classifiers

The original datasets were processed using a fine tuned YOLOv8 segmentation model to create datasets containing only cropped face images with masked backgrounds. All images were resized to 224×224 pixels.

Given the relatively small size of the cats and horses datasets (450 and 120 images, respectively), we employed leave-one-out cross-validation (LOO-CV) with no subject overlap. LOO-CV is a specialized form of cross-validation where the number of folds equals the number of instances in the dataset. This method involves training the model on all instances except one, which is used as the test set, and repeating this process for each instance⁵⁸. In our case, each individual cat or horse was treated as a separate test set. During training, each epoch consisted of multiple stages: at stage i , the model was trained on images from all individuals except subject i , and validation was performed on subject i . The final accuracy and loss for the epoch were averaged across all subjects. This approach is particularly recommended for datasets where each individual has multiple associated samples⁵⁹. For the dogs dataset, which contained significantly more images (approximately 75 images per video across 248 videos from 29 different dogs), we adopted a different strategy. Instead of LOO-CV, we partitioned the dataset, assigning images from 23 dogs to the training set and images from the remaining 6 dogs to the test set, ensuring no subject overlap between training and testing.

Heat maps generation

During our experiments, we utilized several CAM-based algorithms to generate heatmaps, including Grad-CAM²⁷, Axiom-based Grad-CAM (xGrad-CAM)⁴⁹, and Grad-CAM++²⁸. Class Activation Mapping (CAM) algorithms are a widely used approach for generating heatmaps that illustrate the importance of different regions in an image for the final output. Originally developed for convolutional neural networks (CNNs), these algorithms have since been adapted for other architectures, such as Vision Transformers (ViTs). CAM-based methods work by assigning weights to each feature map produced by the convolutional layers of a neural network, determining the significance of each feature map in the final classification decision. As feature maps are smaller than the input image, these maps are later up-sampled into the original image size. Visualization is done by converting the heatmap value into rgb values and overlaying the result over the input image. In this paper we use the raw grades given to the pixels, before conversion to rgb, and convert the map into a probability map, by dividing the pixels by the map's sum. It should be noticed that the map values are assumed to contain only non-negative values. This can be ensured through rescaling or the application of the ReLU function. The following paragraphs provide a detailed explanation of the basic CAM²⁶ algorithm and the methods applied in this work, along with the formulation of the power transform used as a post-processing phase over the heatmaps.

CAM

The Class Activation Mapping (CAM) algorithm, introduced by Zhou et al.²⁶, leverages the global average pooling (GAP) layer to replace the fully connected layers in a convolutional neural network (CNN). This technique visualizes the regions of an input image that are important for the CNN's classification decision.

First, the GAP layer computes the average of each feature map f_k from the last convolutional layer:

$$F_k = \frac{1}{Z} \sum_i \sum_j f_k(i, j)$$

where Z is the total number of pixels in the feature map. These averaged values F_k are then multiplied by the corresponding class weights w_k^c from the final classification layer for class c :

$$S_c = \sum_k w_k^c F_k$$

Next, the class activation map M_c is generated by summing the weighted feature maps:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y)$$

This heatmap M_c highlights the discriminative regions of the image for the predicted class. Finally, the heatmap is upsampled to match the input image size, providing a visual explanation of the model's decision-making process.

Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM)²⁷, is an extension of the original CAM²⁶ algorithm. While CAM requires a specific architecture with a global average pooling (GAP) layer, Grad-CAM can be applied to any convolutional neural network (CNN) architecture without modifications, making it more versatile.

The mathematical formulation of Grad-CAM involves computing the gradient of the class score y^c with respect to the feature maps A^k of the last convolutional layer. These gradients are then averaged to obtain the weights α_k^c :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where Z is the total number of pixels in the feature map.

The class activation map L^c is then calculated as a weighted sum of the feature maps:

$$L^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right)$$

Grad-CAM has gained significant popularity due to its ability to provide clear and intuitive visualizations, making it a widely used tool.

Grad-CAM++

Grad-CAM++²⁸ is an advanced version of the Grad-CAM algorithm designed to provide more precise and detailed visual explanations for convolutional neural network (CNN) predictions. This is achieved by using a weighted combination of the positive partial derivatives of the class score with respect to the feature maps, which allows for more accurate identification of the important regions in the image.

To begin, the gradients of the class score y^c with respect to the feature maps A^k of the last convolutional layer are computed:

$$\frac{\partial y^c}{\partial A_{ij}^k}$$

Next, the positive partial derivatives are used to obtain the weights α_{ij}^k . These weights are calculated as a combination of the second and third partial derivatives of the class score:

$$\alpha_{ij}^k = \frac{\partial^2 y^c}{\partial (A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \frac{\partial^3 y^c}{\partial (A_{ij}^k)^3}$$

These weights are then averaged over all spatial locations to get the final weights α_k^c :

$$\alpha_k^c = \sum_i \sum_j \alpha_{ij}^k$$

Finally, the class activation map L^c is generated as a weighted sum of the feature maps:

$$L^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right)$$

The authors of Grad-CAM++ evaluated their algorithm through a series of experiments on various datasets. These evaluations included both subjective and objective tests to assess the effectiveness of the visual explanations. The evaluations involved measuring localization accuracy and user studies to gather subjective feedback on the interpretability and usefulness of the visual explanations provided by Grad-CAM++.

xGrad-CAM

Axiom-based Grad-CAM (xGrad-CAM)⁴⁹ is an enhanced version of the traditional Grad-CAM method used for visualizing and interpreting Convolutional Neural Networks (CNNs). It integrates two novel key aspects: sensitivity and conservation, to improve the accuracy and reliability of the visualizations.

- Sensitivity ensures that if a feature map has a significant impact on the output, its corresponding gradient should also be significant.
- Conservation ensures that the sum of the importance scores of all feature maps should be conserved, meaning the total importance remains constant.

xGrad-CAM modifies the computation of α_k^c (the weight for the k -th feature map with respect to class c) to better satisfy the Sensitivity and Conservation axioms. The modified weight α_k^c is computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \left(\frac{\partial y^c}{\partial A_{ij}^k} \cdot A_{ij}^k \right)$$

This modification ensures that the importance of each feature map is weighted by both its gradient and its activation, aligning with the Sensitivity and Conservation axioms. xGrad-CAM was tested on various datasets, including image classification and object detection tasks. The evaluations involved measuring localization accuracy and user studies.

Power transform

The power transformation applies a power function to each data point, which results in either compressing higher values or expanding lower ones, depending on the chosen exponent. In this work, a power transform with an exponent of 2.0 was applied to the heatmaps, after adjusting the values to be non-negative and before converting them to probabilities.

By using an exponent greater than 1 (such as squaring the data), the transformation amplifies high values, making them more pronounced. On the other hand, values close to zero are further compressed.

Applying a power transform as a post-processing step for heatmaps in deep learning enhances the contrast between areas of high and low importance, thereby improving the heatmap's interpretability and making it more useful for analysis. The last two columns of Fig. 7 visually illustrate the effect of the power transform on the tested datasets - the maps after applying the power transform are more distinct.

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.⁶⁰.

For data citations of datasets uploaded to e.g. *figshare*, please use the [SPSVERBc1SPS] option in the bib entry to specify the platform and the link, as in the [SPSVERBc2SPS] example in the sample bibliography file.

Data availability

The algorithms used in this paper are available at [GitHub repository](#). Data is available upon reasonable request from the corresponding author.

Received: 2 January 2025; Accepted: 31 March 2025

Published online: 21 April 2025

References

- Hogg, M. A. & Abrams, D. Social cognition and attitudes. In *Psychology* 3rd edn (eds Martin, G. N. et al.) 684–721 (Pearson Education Limited, 2007).
- Picard, R. W. *Affective computing* (MIT press, 2000).
- Ho, M.-T., Mantello, P., Nguyen, H.-K.T. & Vuong, Q.-H. Affective computing scholarship and the rise of china: a view from 25 years of bibliometric data. *Humanities Soc. Sci. Commun.* **8**, 1–14 (2021).
- Tao, J. & Tan, T. Affective computing: A review. In *Affective Computing and Intelligent Interaction* (eds Tao, J. et al.) 981–995 (Springer, 2005).
- Sharma, G. & Dhall, A. A survey on automatic multimodal emotion recognition in the wild. In *Advances in data science: Methodologies and applications* (eds Phillips-Wren, G. et al.) 35–64 (Springer International Publishing, 2021). <https://doi.org/10.1007/978-3-030-51870-73>.
- Zhu, X. et al. A review of key technologies for emotion analysis using multimodal information. *Cogn. Comput.* **16**, 1504–1530 (2024).
- Zhu, X., Huang, Y., Wang, X. & Wang, R. Emotion recognition based on brain-like multimodal hierarchical perception. *Multimed. Tools Appl.* **83**, 56039–56057 (2024).
- Wang, R. et al. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. *Int. J. Multimed. Inform. Retrieval* **13**, 39 (2024).
- Zhu, X. et al. A client-server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods. *Comput. Methods Programs Biomed.* **260**, 108564 (2025).
- Diogo, R., Abdala, V., Lonergan, N. & Wood, B. From fish to modern humans-comparative anatomy, homologies and evolution of the head and neck musculature. *J. Anat.* **213**, 391–424 (2008).
- Mathis, A. et al. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281 (2018).
- Pennington, Z. T. et al. ezTrack: An open-source video analysis pipeline for the investigation of animal behavior. *Sci. Rep.* **9**, 1–11 (2019).
- Amir, S., Zamansky, A., van der Linden, D. K9-blyzer-towards video-based automatic analysis of canine behavior. In: *Proceedings of Animal-Computer Interaction 2017* (2017).
- Pereira, T. D. et al. Fast animal pose estimation using deep neural networks. *Nat. Methods* **16**, 117–125 (2019).
- Graving, J. M. et al. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife* **8**, e47994 (2019).
- Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J. & de Polavieja, G. G. Idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nature Methods* **16**, 179–182 (2019).
- Broomé, S. et al. Going deeper than tracking: a survey of computer-vision based recognition of animal pain and affective states. *arXiv preprint arXiv:2206.08405* (2022).
- Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning* Vol. 1 (MIT press Cambridge, 2016).
- Wang, Y. et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* (2022).
- Arrieta, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inform. Fusion* **58**, 82–115 (2020).
- Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. & Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **109**, 247–278. <https://doi.org/10.1109/JPROC.2021.3060483> (2021).
- Räuker, T., Ho, A., Casper, S., Hadfield-Menell, D. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In: *2023 IEEE conference on secure and trustworthy machine learning (satml)*, 464–483 (IEEE, 2023).
- Ras, G., Xie, N., Van Gerven, M. & Doran, D. Explainable deep learning: A field guide for the uninitiated. *J. Artificial Intell. Res.* **73**, 329–396 (2022).
- La Rosa, B. et al. State of the art of visual analytics for explainable deep learning. In: *Computer Graphics Forum*, vol. 42, 319–355 (Wiley Online Library, 2023).
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **116**, 22071–22080 (2019).
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929 (2016).
- Selvaraju, R.R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *ICCV*, 618–626 (IEEE Computer Society, 2017).
- Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRRabs/1710.11063* (2017). 1710.11063.
- Kindermans, P.-J. et al. The (un) reliability of saliency methods. In: *Explainable AI: Interpreting, explaining and visualizing deep learning* 267–280 (2019).
- Ellis, G. *Cognitive biases in visualizations* (Springer, 2018).
- Blumenthal-Barby, J. S. & Krieger, H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. *Med. Decis. Making* **35**, 539–557 (2015).

32. Adebayo, J. et al. Sanity checks for saliency maps. *Advances in neural information processing systems* **31** (2018).
33. Hooker, S., Erhan, D., Kindermans, P.-J., Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems* **32** (2019).
34. Kim, S. S., Meister, N., Ramaswamy, V. V., Fong, R., Russakovsky, O. Hive: Evaluating the human interpretability of visual explanations. In: *European Conference on Computer Vision*, 280–298 (Springer, 2022).
35. Tjoa, E. & Guan, C. Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset. *IEEE Trans. Artif. Intell.* **4**, 858–870. <https://doi.org/10.1109/TAI.2022.3228834> (2023).
36. Li, X.-H. et al. An experimental study of quantitative evaluations on saliency methods. In: *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 3200–3208 (2021).
37. Boneh-Shitrit, T. et al. Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration. *Sci. Rep.* **12**, 22611 (2022).
38. Broomé, S., Gleerup, K. B., Andersen, P. H., Kjellstrom, H. Dynamics are important for the recognition of equine pain in video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12667–12676 (2019).
39. Feighelstein, M. et al. Explainable automated pain recognition in cats. *Sci. Rep.* **13**, 8973 (2023).
40. Correia-Caeiro, C., Burrows, A., Wilson, D. A., Abdelrahman, A. & Miyabe-Nishiwaki, T. Callifacs: The common marmoset facial action coding system. *PLoS One* **17**, e0266442 (2022).
41. Caeiro, C., Waller, B., Zimmerman, E., Burrows, A. & Davila Ross, M. Orangfacs: A muscle-based movement coding system for facial communication in orangutans. *Int. J. Primatol.* **34**, 115–129 (2013).
42. Waller, B., Correia Caeiro, C., Peirce, K., Burrows, A., Kaminski, J. Dogfacs: the dog facial action coding system (2013).
43. Evangelista, M. C. et al. Facial expressions of pain in cats: the development and validation of a feline grimace scale. *Sci. Rep.* **9**, 1–11 (2019).
44. Finka, L. R. et al. Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. *Sci. Rep.* **9**, 1–12 (2019).
45. Dalla Costa, E. et al. Development of the horse grimace scale (hgs) as a pain assessment tool in horses undergoing routine castration. *PLoS One* **9**, e92281 (2014).
46. Bremhorst, A., Sutter, N. A., Würbel, H., Mills, D. S. & Riemer, S. Differences in facial expressions during positive anticipation and frustration in dogs awaiting a reward. *Sci. Rep.* **9**, 1–13 (2019).
47. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).
48. Caron, M. et al. Emerging properties in self-supervised vision transformers. *ICCV* (2021).
49. Fu, R. et al. *Axiom-based grad-cam: Towards accurate visualization and explanation of cnns* **2008**, 02312 (2020).
50. Feighelstein, M. et al. Automated recognition of pain in cats. *Sci. Rep.* **12**, 9575 (2022).
51. Zhang, Z. et al. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In: *AAAI Conference on Artificial Intelligence (AAAI)* (2022).
52. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. *CVPR* (2016).
53. Oquab, M. et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
54. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245 (2000).
55. Kindermans, P.-J. et al. Learning how to explain neural networks: Patternnet and patternattribution. 10.48550/arXiv.1705.05598 (2017).
56. Jocher, G., Chaurasia, A., Qiu, J. Ultralytics yolov8 (2023).
57. Sammut, C. & Webb, G. I. (eds) *Leave-One-Out cross-validation* 600–601 (Springer, 2010).
58. Sammut, C., Webb, G. I. Leave-one-out cross-validation. *Encyclopedia of machine learning* 600–601 (2010).
59. Broome, S. et al. Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions. *Int. J. Comput. Vision* **131**, 572–590 (2023).
60. Hao, Z., AghaKouchak, A., Nakhjiri, N., Farahmand, A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. *figshare* <https://doi.org/10.6084/m9.figshare.853801> (2014).

Acknowledgements

The research was partially supported by the SNSF-ISF binational project Switzerland - Israel (grant number 1050/24). We thank Nareed Farhat, Ephantus Kanyugi and Yaron Jossef for data management assistance.

Author contributions

A.B., L.F., D.M., S.L., E.D.C. have collected and annotated the data. T.S. and A.Z. conceived the experiment(s), T.S. conducted the experiment(s), T.S. and A.Z. analysed the results. All authors participated in writing and reviewing the manuscript.

Additional information

Correspondence and requests for materials should be addressed to A.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025