UDC 004.832
DOI: 10.53297/18293336-2025.2-51

DOG EMOTION RECOGNITION IN IMAGES USING FINE-TUNED VISION TRANSFORMERS

E.A. Harutyunyan

National Polytechnic University of Armenia
Krisp

Recognizing canine emotions has a practical value for veterinary practice, welfare monitoring, and safer human–dog interaction. This work investigates transformer-based image classification for dog emotion recognition and presents a complete pipeline that fine-tunes a ViT-B/16 backbone on a four-class dataset (angry, happy, relaxed, sad). Images are standardized to 224×224 and normalized to ImageNet statistics, with stochastic augmentation (flips, rotations, color jitter, brightness/contrast, and small affine shifts) to improve robustness. A new 4-way classification head is trained with differential learning rates on top of a pretrained ImageNet-21k encoder and optimized with AdamW, class-weighted cross-entropy, warm-up, cosine annealing, early stopping, and checkpointing. Post-processing includes confidence thresholding and optional temporal smoothing for video scenarios. On the held-out test set, the fine-tuned ViT achieves 82,6% accuracy, outperforming a fine-tuned ResNet-50 (75,4%) and a ViT trained from scratch (68,9%). Per-class analysis shows the highest discrimination for “Happy,” while “Sad” and “Relaxed” are most frequently confused due to subtle visual overlap. These findings indicate that global self-attention in ViTs captures nuanced cues (e.g., ear position and mouth tension) better than convolutional baselines, and that transfer learning is critical under limited labeled data. The study highlights the remaining challenges in cross-breed generalization, viewpoint and lighting variation, and label subjectivity, and points toward multimodal extensions and temporally aware models for further gains.

Keywords: vision transformer, self-attention, fine-tuning, data augmentation.

Introduction. Recognizing animal emotions is a growing focus in veterinary care, animal welfare, and human–animal interaction research. Dogs, as highly social companion animals, express emotions through facial features, posture, and body language. Traditional assessment methods depend on expert observation, which is subjective, time-consuming, and difficult to scale. Advances in computer vision and deep learning provide opportunities for automated recognition systems that can support professionals and pet owners alike.

Recent progress in transformer-based models [1], particularly Vision Transformers (ViTs), has reshaped image classification. Unlike convolutional networks that emphasize local patterns, ViTs use self-attention to capture global dependencies, enabling them to detect subtle cues in complex visual data. Such characteristics make ViTs promising for interpreting nuanced canine expressions. However, challenges remain: morphological differences across breeds, variations in lighting

and perspective, and the absence of direct ground truth labels for non-verbal subjects complicate model training and evaluation.

Transfer learning through fine-tuning pre-trained models [2] offers an effective strategy to overcome data scarcity and improve generalization. By adapting representations learned from large-scale datasets, ViTs can be specialized for domains like animal emotion recognition. This work investigates fine-tuned ViTs for classifying four primary dog emotions—angry, happy, relaxed, and sad—and compares their performance against baseline methods, with attention to preprocessing, augmentation, and optimization strategies.

Literature Review. Prior work on automated dog emotion recognition has largely focused on optimizing classical neural networks or leveraging pose estimation.

In [3], one of the first studies applying deep learning to classify dog emotions from facial expressions is conducted. Working in a controlled experimental setup, they focus on differentiating between positive anticipation and negative frustration. The authors compare CNNs with ViTs under both supervised and self-supervised training. Notably, features from a DINO-pretrained ViT outperform other backbones, highlighting the value of global self-attention for subtle emotional cues such as ear position, eye shape, and mouth tension. Their findings demonstrate that self-supervised ViTs provide richer representations for fine-grained canine emotion recognition.

[4] develops a posture-based dog emotion classifier using DeepLabCut to detect 24 dog keypoints, then trains both a neural network and a decision tree on either raw coordinates or derived pose metrics. They have collected and annotated ~13,800 dog images to train the detector, then built a balanced dataset of 400 images for four emotion classes (anger, fear, happiness, relaxation). The neural network achieves ~67.5 % accuracy, while the decision tree reaches ~62.5 %. Their results suggest that the full-body pose, rather than facial features alone, is a viable input for the dog emotion recognition.

[5] presents a CNN-based framework for real-time dog detection and emotion recognition in surveillance video streams. The pipeline first tracks dogs across frames, crops their regions of interest, and then classifies their emotional state (e.g. calm, aggressive, neutral) using a CNN trained on annotated video sequences. The authors report that incorporating temporal continuity (i.e. frame-to-frame consistency) helps suppress false positives and smooth predictions. They demonstrate the system in real surveillance settings, showing robustness to occlusion, background clutter, and viewpoint changes.

[6] presents *DogChat*, a smart collar system integrating first-person visual

and auditory sensors with large language models (LLMs) to enable pets to "communicate" via WeChat. The system operates in three phases: Pet Profile Construction, Daily Experience Reconstruction, and Behavior Learning Integration, transforming the sensor data into descriptive emotional or behavioral outputs. Users can receive proactive updates about their pet's status or query the pet's "feelings" through chat. The prototype highlights potential in bridging nonverbal animal signals with natural language, though it faces challenges related to privacy, interpretability, and deployment in diverse real-world settings.

Research methodology. The proposed methodology outlines the complete workflow for fine-tuning ViT models to recognize emotional states in dogs from digital images.

Dataset description and characteristics. This study uses the Dog Emotions Prediction dataset [7] from Kaggle, which includes images labeled into four emotional states: angry, happy, relaxed, and sad (Fig. 1). The photographs come from diverse sources-personal collections, repositories, and online platforms-covering a wide range of breeds, sizes, ages, and contexts. Labels were assigned by annotators with canine behavior expertise, though subjectivity in emotion labeling remains a challenge. The dataset shows substantial variation in resolution, lighting, backgrounds, and framing, from close-up faces to full-body views, reflecting real-world conditions for recognition systems.

Fig. 1. Sample images of the four emotional states in the dataset: happy (left), angry, relaxed, and sad (right)

The analysis carried out assessed the class distribution, image quality, and potential biases were also examined. Brightness, contrast, and color properties to guide preprocessing. Manual inspection of samples confirmed the label quality and highlighted common cues were also examined for each class, such as ear position, mouth shape, or gaze direction.

Data preprocessing and augmentation. All images were resized to 224×224 pixels using bicubic interpolation and normalized with ImageNet statistics (mean = (0.485, 0.456, 0.406), std = (0.229, 0.224, 0.225)) to align with pretrained Vision Transformer (ViT) inputs. To improve robustness and mitigate

overfitting, we applied stochastic data augmentation during training: random horizontal flips (p = 0.5), rotations (±15°), brightness/contrast adjustments (0.8–1.2), color jitter, and small affine translations (≤10%). These augmentations simulate real-world variations without obscuring emotional cues.

The model architecture and transfer learning. We employ the ViT-B/16 model, which splits images into 16×16 patches, linearly embeds them, and applies multi-head self-attention across transformer layers (Fig. 2). The pretrained backbone (ImageNet-21k) provides rich feature representations. We replace the original classification head with a 4-class output layer (angry, happy, relaxed, sad) and apply dropout (p = 0.1).

For fine-tuning, we adopt differential learning rates:

- **Classification head:** lr = 1e-3 (newly initialized).
- **Transformer encoder:** lr = 1e-5 (pretrained layers, subtle updates).

Optimization is performed using AdamW with weight decay (0.01), β_1=0.9, β_2=0.999.

The training strategy and hyperparameter configuration. Training runs for 50 epochs with batch size 32. A warm-up phase (5 epochs) linearly increases the learning rate followed by cosine annealing to gradually reduce it. The loss function is weighted categorical cross-entropy [8] (Formula (1)):

$$L = \sum_{i=1}^{C} w_i y_i log(p_i), \quad (1)$$

where C is the number of classes, y_i - the true label, p_i - the predicted probability, and w_i - the class weight (inverse to class frequency).

To avoid overfitting, we use early stopping (patience = 10) and checkpointing. Validation (20% stratified split) is used only for hyperparameter tuning.

Postprocessing and prediction enhancement. Predictions are refined using temporal smoothing for video applications with an exponential moving average (EMA) [8] (Formula (2)):

$$p_i'(t) = ap_i(t) + (1 - a)p_i'(t - 1) \quad (2)$$

with the smoothing factor α=0.3.

For reliability, we apply a confidence threshold (0.6). Low-confidence outputs include the top-2 predicted classes. Prediction entropy [8] (Formula (3))

$$H = -\sum_{i=1}^{C} p_i log(p_i) \quad (3)$$

is used to flag uncertainty.

This workflow provides a systematic foundation for adapting transformer-based models to specialized animal emotion recognition tasks. The following section presents experimental results and performance comparisons that validate the effectiveness of the proposed approach.

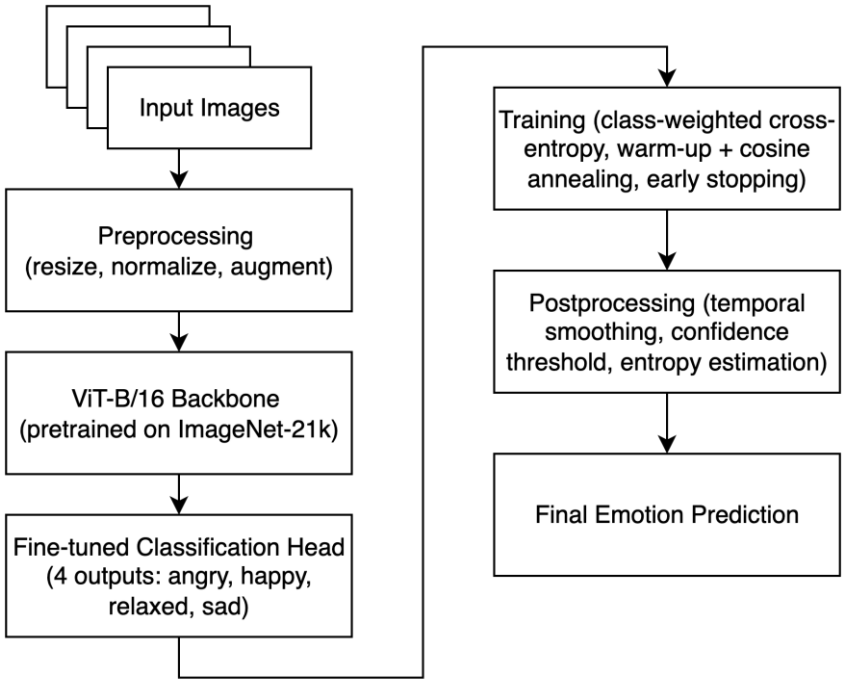

Fig. 2. End-to-end pipeline of dog emotion recognition using fine-tuned Vision Transformers

Results. The fine-tuned Vision Transformer (ViT-B/16) achieved an overall accuracy of 82,6% on the test set, outperforming both a fine-tuned ResNet-50 (75.4%) and a ViT trained from scratch (68.9%) (Table 1). This confirms the effectiveness of transfer learning for recognizing subtle emotional cues in dogs. Early stopping occurred at epoch 39, preventing overfitting while preserving optimal validation performance.

Table 1

Pre-trained vs Fine-tuned Model Performance

Model Architecture	**Accuracy, %**	**Precision (avg), %**	**Recall (avg), %**	**F1-Score (avg), %**
ViT-B/16 (Fine-tuned)	82.6	82.1	82.4	82.2
ResNet-50 (Fine-tuned)	75.4	74.0	74.6	74.3
ViT-B/16 (From scratch)	68.9	67.5	68.1	67.8
Simple CNN	63.2	62.1.	62.9	62.5

The ViT model most reliably detected happy emotions (distinct open-mouth expressions), while sad and relaxed states were occasionally confused due to overlapping visual patterns.

Per-class evaluation indicates consistent and balanced performance across all emotions, with slightly higher precision for active states like *Happy* and *Angry* (Table 2).

Table 2

Pre-trained vs Fine-tuned Model Performance

Emotion Class	**Precision, %**	**Recall, %**	**F1-Score, %**	**Support (samples)**
Happy	86.1	87.5	86.8	245
Relaxed	80.3	81.2	80.7	228
Angry	82.7	79.8	81.2	212
Sad	78.6	76.9	77.7	198

The *Happy* class achieved the highest AUC (0.931), followed by *Angry (0.902)*, *Relaxed (0.891)*, and *Sad (0.866)* - consistent with visual distinctiveness and data balance. Most misclassifications occurred between *Sad* and *Relaxed*, reflecting their subtle visual overlap.

The results demonstrate that fine-tuned ViT models effectively capture

emotional patterns in dog facial features, achieving robust performance across diverse conditions. Future work could enhance recognition of visually similar emotions through multimodal fusion (e.g., incorporating video or sound cues).

Conclusion. This study demonstrates that fine-tuned Vision Transformers provide an effective solution for recognizing dog emotions from images, delivering 82.6% test accuracy and surpassing both a fine-tuned ResNet-50 and a ViT trained from scratch. The pipeline—standardized preprocessing, targeted augmentation, differential learning rates, AdamW with weight decay, warm-up plus cosine annealing, and early stopping—yields balanced precision/recall across classes, with the clearest improvements on visually distinctive "Happy" and "Angry" states. The Remaining errors arise mainly between "Sad" and "Relaxed," consistent with their overlapping visual cues. Beyond static images, the proposed confidence thresholding and optional exponential-moving-average smoothing offer a pragmatic bridge to video scenarios. Looking ahead, three avenues appear most promising: (i) multimodal fusion with audio or sequential visual cues to disambiguate subtle states; (ii) domain-robust training that addresses the breed morphology, pose, lighting, and background shifts; and (iii) improved supervision, including expert-in-the-loop relabeling or soft labels to mitigate annotation subjectivity. With these extensions, transformer-based systems can evolve from accurate offline classifiers into dependable, deployable tools that support veterinarians, behaviorists, and pet-centric applications in real-world conditions.

References

1. Detection of depression severity in social media text using transformer-based models / **A. Qasim, G. Mehak, N. Hussain, et al** // Journal of Information. -2025, - P. 114.
2. Transfer learning for software vulnerability prediction using Transformer models / **I. Kalouptsoglou, M. Siavvas, A. Ampatzoglou, et al** // Journal of Systems and Software. -2025. - P. 112448.
3. Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration **T. Boneh-Shitrit, M. Feighelstein, A. Bremhorst, et al** // Scientific Reports. -2022. - P. 22611.
4. **Ferres K., Schloesser T., Gloor P.A.** Predicting dog emotions based on posture analysis using DeepLabCut. Future Internet. -2022. - P. 97.
5. **Chen H.-Y., Lin C.-H., Lai J.-W., Chan Y.-K.** Convolutional neural network-based automated system for dog tracking and emotion recognition in video surveillance // Applied Sciences. -2023. - P. 4596.
6. **Xue C., Zuo Z., Jiang X., Fu X.** DogChat: A pet-centered smart collar prototype based on large language models and WeChat // In Companion of the 2024

ACM International Joint Conference on Pervasive and Ubiquitous Computing. - 2024. - P. 162–166.
7. Dog Emotions Prediction Dataset. https://www.kaggle.com/datasets/devzohaib/dog-emotions-prediction, Accessed: 04/10/25.
8. **Yang X., Song Z., King I., Xu Z.** A survey on deep semi-supervised learning // IEEE Transactions on Knowledge and Data Engineering.-2022.-35(9). - P. 8934–8954.

Received on 07.11.2025.
Accepted for publication on 29.01.2026.

ՇՆԵՐԻ ՀՈՒՅԶԵՐԻ ՃԱՆԱՉՈՒՄԸ ՊԱՏԿԵՐՆԵՐՈՒՄ՝ ՕԳՏԱԳՈՐԾԵԼՈՎ ՎԵՐԱՎԱՐԺԵՑՎԱԾ ՏԵՍՈՂԱԿԱՆ ՏՐԱՆՍՖՈՐՄԵՐՆԵՐ

Է.Ա. Հարությունյան

Շների հույզերի ճանաչումը գործնական արժեք ունի անասնաբուժական պրակտիկայի, նրանց բարեկեցության մոնիտորինգի և մարդ-շուն ավելի անվտանգ փոխազդեցության համար: Հետազոտվում է տրանսֆորմերների վրա հիմնված պատկերների դասակարգումը շների հույզերի ճանաչման համար, և ներկայացվում է ամբողջական գործընթաց, որը կատարում է ViT-B/16 հիմնական ճարտարապետության լրացուցիչ վարժեցում չորս դասերի տվյալների բազայի վրա (բարկացած, հանգիստ, ուրախ, տխուր): Պատկերները ստանդարտացված են 224×224 չափով և նորմալացված են ImageNet վիճակագրության համաձայն՝ ստոխաստիկ ավելացման միջոցով (հայելային արտացոլումներ, պտույտներ, գունային ջիտտեր, պայծառություն/հակադրություն և փոքր աֆինային տեղաշարժեր)՝ կայունությունը բարելավելու համար: Նոր 4-ուղղություն դասակարգման գլուխը վարժեցվում է տարբերակված ուսուցման արագություններով ImageNet-21k-ի՝ նախապես վարժեցված կոդավորիչի վրա և օպտիմալացվում է AdamW-ի, դասերի կշռով խաչաձև էնտրոպիայի, տաքացման, կոսինուսային տեսակավորման, վաղ կանգի և ստուգակետերի պահպանման միջոցով: Հետմշակումը ներառում է վստահության շեմային զտում և ընտրովի ժամանակային հարթեցում վիդեո սցենարների համար: Առանձնացված թեստային բազայի վրա լրացուցիչ վարժեցված ViT-ը հասնում է 82,6% ճշգրտության՝ գերազանցելով լրացուցիչ վարժեցված ResNet-50-ին (75,4%) և զրոյից վարժեցված ViT-ին (68,9%): Դասային վերլուծությունը ցույց է տալիս առավելագույն տարբերակում «ուրախ» դասի համար, մինչդեռ «տխուր» և «հանգիստ» դասերը հաճախ շփոթվում են նրբին տեսողական համընկնման պատճառով: Այս արդյունքները ցույց են տալիս, որ ViT-ի գլոբալ ինքնուրույն ուշադրությունը ավելի լավ է որսում նուրբ հատկանիշները (օրինակ՝ ականջների դիրքը և բերանի լարվածությունը), քան կոնվոլյուցիոն բազային մոդելները, և որ տրանսֆերային ուսուցումը կրիտիկական նշանակություն ունի սահմանափակ

պիտակավորված տվյալների պայմաններում: Ուսումնասիրությունն ընդգծում է մնացած մարտահրավերները խաչաձև ցեղային ընդհանրացման, դիտանկյունի և լուսավորության տատանումների, ինչպես նաև պիտակների սուբյեկտիվության դեպքում, և մատնանշում է բազմամոդալ ընդլայնումներ և ժամանակային գիտակցությամբ մոդելներ հետագա բարելավումների համար:

Առանցքային բառեր. տեսողական տրանսֆորմեր, ինքնուշադրություն, վերավարժեցում, տվյալների ավելացում:

РАСПОЗНАВАНИЕ ЭМОЦИЙ СОБАК НА ИЗОБРАЖЕНИЯХ С ИСПОЛЬЗОВАНИЕМ ДООБУЧЕННЫХ ВИЗУАЛЬНЫХ ТРАНСФОРМЕРОВ

Э.А. Арутюнян

Распознавание эмоций собак имеет практическую ценность для ветеринарной практики, мониторинга благополучия животных и более безопасного взаимодействия человека с собакой. В данной работе исследуется классификация изображений на основе трансформеров для распознавания эмоций собак и представлен полный конвейер, который выполняет дообучение архитектуры ViT-B/16 на наборе данных из четырех классов (злой, счастливый, расслабленный, грустный). Изображения стандартизированы до размера 224×224 и нормализованы по статистике ImageNet со стохастической аугментацией (отражения, повороты, цветовой джиттер, яркость/контраст и небольшие аффинные сдвиги) для повышения устойчивости. Новая классификационная голова с 4 выходами обучается с дифференцированными скоростями обучения поверх предобученного на ImageNet-21k кодировщика и оптимизируется с помощью AdamW, взвешенной по классам кросс-энтропии, разогрева, косинусного отжига, ранней остановки и сохранения контрольных точек. Постобработка включает пороговую фильтрацию по уверенности и опциональное временное сглаживание для видеосценариев. На отложенном тестовом наборе дообученный ViT достигает точности 82,6%, превосходя дообученный ResNet-50 (75,4%) и ViT, обученный с нуля (68,9%). Поклассовый анализ показывает наивысшую различимость для класса «Счастливый», тогда как «Грустный» и «Расслабленный» чаще всего путаются из-за тонкого визуального перекрытия. Эти результаты указывают на то, что глобальное самовнимание в ViT лучше улавливает нюансированные признаки (например, положение ушей и напряжение рта), чем сверточные базовые модели, и что трансферное обучение критически важно при ограниченных размеченных данных. Исследование подчеркивает оставшиеся проблемы в кросс-породной генерализации, вариации угла обзора и освещения, а также субъективности разметки и указывает на мультимодальные расширения и темпорально осведомленные модели для дальнейших улучшений.

Ключевые слова: визуальный трансформер, самовнимание, дообучение, аугментация данных.