CAGE: Circumplex Affect Guided Expression Inference

Niklas Wagner[1,*], Felix Mätzler[1,*], Samed R. Vossberg[1,*], Helen Schneider[1*], Svetlana Pavlitska[2], J. Marius Zöllner[1,2]
[1] *Karlsruhe Institute of Technology (KIT), Germany*
[2] *FZI Research Center for Information Technology, Germany*
helen.schneider@kit.edu

Abstract

Understanding emotions and expressions is a task of interest across multiple disciplines, especially for improving user experiences. Contrary to the common perception, it has been shown that emotions are not discrete entities but instead exist along a continuum. People understand discrete emotions differently due to a variety of factors, including cultural background, individual experiences, and cognitive biases. Therefore, most approaches to expression understanding, particularly those relying on discrete categories, are inherently biased. In this paper, we present a comparative in-depth analysis of two common datasets (AffectNet and EMOTIC) equipped with the components of the circumplex model of affect. Further, we propose a model for the prediction of facial expressions tailored for lightweight applications. Using a small-scaled MaxViT-based model architecture, we evaluate the impact of discrete expression category labels in training with the continuous valence and arousal labels. We show that considering valence and arousal in addition to discrete category labels helps to significantly improve expression inference. The proposed model outperforms the current state-of-the-art models on AffectNet, establishing it as the best-performing model for inferring valence and arousal achieving a 7% lower RMSE. Training scripts and trained weights to reproduce our results can be found here: https://github.com/wagner-niklas/CAGE_expression_inference.

1. Introduction

The inference of emotions through expressions has been a topic of interest for the past years as it might give insights into a person's feelings towards other individuals or topics. Mehrabian and Wiener [35] suggest 55% of communication is perceived by expressions. Lapakko [28] argues, however, that these findings are limited to emotional states. Automation of analysis of expressions to get insights into user experience is one step towards live feedback without direct interaction with an individual.

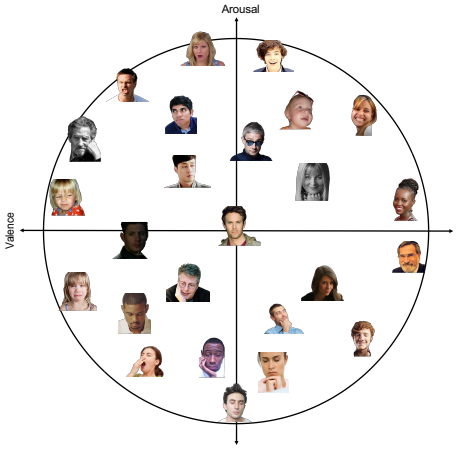


Figure 1. *Valence/arousal* for sample images from AffectNet [36].

A common approach is *expression inference*, i.e. classification of emotional expressions into discrete categories. However, a comprehensive meta-analysis of facial expressions research by Barrett et al. [2] has shown, that there is no consensus across cultures and intra-cultural over specific facial movements reliably depicting one category of emotion. They suggest that affective states can more reliably be inferred by a third-party individual. They emphasize that these states are inferred, not recognized. According to Russell [39], affects can be described as a set of dimensions with each dimension varying independently. These dimensions are called *valence* and *arousal*, representing the positivity/negativity and intensity-/activation of expressions respectively. Using *valence* and *arousal* of the circumplex model of affect [39] as additional dimensions rather than only discrete emotions for expression inference thus offers a more robust framework, as they provide a continuous spec-

*These authors contributed equally to this work

trum that captures the underlying affective states.

In this work, we compare training with *valence* and *arousal* labels merged with the commonly used discrete emotions to train with the two approaches separately. Our approach involves pinpointing the differences and similarities between two leading datasets that catalog images according to their explicit discrete and continuous emotional states: AffectNet [36] and EMOTIC [26]. We then develop a lightweight deep neural network tailored for computer vision tasks, aiming to accurately infer these discrete emotions as well as the continuous dimensions of *valence* and *arousal*, surpassing the performance of existing models. In particular, our model improves accuracy by reducing the root-mean-square error (RMSE) by 7.0% for *valence* and 6.8% for *arousal*. It also increases the concordance correlation coefficients (CCC) by 0.8% for *valence* and 2.0% for *arousal* when tested on the AffectNet dataset. These improvements are reflected in our final results, with CCC values of 0.716 for *valence* and 0.642 for *arousal*, and RMSE values of 0.331 for *valence* and 0.305 for *arousal*. Furthermore, we exceed the top-3 accuracy set by Khan *et al*. [13] on the EMOTIC dataset by 1.0%.

2. Related Work

In the field of affective computing, in particular expression inference, the integration of *valence/arousal* regression with discrete emotion classification has emerged as a promising approach to enhance the performance and applicability across diverse datasets. In the following, we discuss existing works in this domain.

2.1. Datasets for Expression Inference

In the domain of expression inference, several datasets exist. However, these datasets vary significantly in both the data they offer and their popularity. Among the most widely used datasets are FER2013 [9] and FERPlus [3], which provide annotated 48×48 pixel black-and-white facial images classified in seven (FER) or eight (FER+) discrete emotional states. While these datasets have been the foundation for numerous research contributions, they have been expanded in various ways over the past years. Notable examples in this context are the EMOTIC [26] and AffectNet [36] datasets, which both contain high-resolution RGB images. AffectNet is a large-scale database containing around 0.4 million facial images labeled by 12 annotators. Each image is annotated with categorical emotions, mirroring those used in the FER+ dataset, in addition to *valence* and *arousal* values. This approach offers a more refined representation of emotions compared to categorical labels only.

The EMOTIC (*Emotions in Context*) dataset provides a more nuanced perspective on affective states. Unlike earlier datasets focused solely on facial expressions, EMOTIC captures individuals in full-body shots within their surrounding context. EMOTIC features bounding boxes that encompass each individual's entire body, eliminating the need for a visible face. Furthermore, it categorizes emotions into 26 discrete categories, allowing for multiple labels per individual. In addition, the dataset expands these discrete values with continuous measures of *valence* and *arousal* as well as *dominance* that measures the level of control a person feels during a situation, ranging from submissive / non-control to dominant / in-control [27].

Method	Accuracy [%]	Date [mm-yy]
DDAMFN [45]	64.25	08-23
POSTER++ [34]	63.77	01-23
S2D [5]	63.06	12-22
MT EffNet-B2 [41]	63.03	07-22
MT-ArcRes [18]	63.00	09-19

Table 1. Top five models on AffectNet-8 benchmark [6].

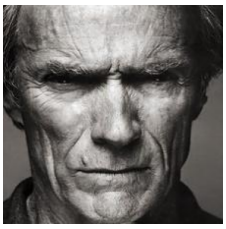
Method	Accuracy [%]	Date [mm-yy]
S2D [5]	67.62	12-22
POSTER++ [34]	67.49	01-23
DDAMFN [45]	67.03	08-23
EmoAffectNet [40]	66.49	12-22
Emotion-GCN [1]	66.46	07-21

Table 2. Top five models on AffectNet-7 benchmark [6].

While there are at least 28 datasets such as CK+ [33], RAF-DB [29] or Aff-Wild2 [14–25, 44] focusing specifically on *facial expression recognition/inference* featuring continuous and/or discrete measures, we chose to focus on the two mentioned above, since we are interested in both discrete emotion labeling on an individual basis as well as continuous measures of *valence* and *arousal*. AffectNet [36] as a state-of-the-art, is arguably the most represented dataset in the current research field. On the other hand, EMOTIC, although not being the most utilized dataset, offers the most refined representation of measures while still focusing on a combination of discrete and continuous variables to define individuals emotion.

2.2. Expression Inference Models

Expression inference on datasets like AffectNet has been addressed in numerous publications. According to Paperswithcode [6], 207 AffectNet-related papers have been published since 2020. Tables 1 and 2 show five best models in leaderboards for the AffectNet-8 and AffectNet-7 test benchmark as of 01.01.2024. As the initial FER dataset does not contain the emotion *Contempt*, there exists also an AffectNet-7 benchmark omitting this emotion. So far, the best-performing models for expression inference have been almost exclusively based on convolutional neural networks (CNNs), e.g. ResNet-18 [10]. Although CNNs are still



(a) AffectNet-8 (*anger*)

(b) EMOTIC (*disconnection*)

Figure 2. Example images from AffectNet-8 and EMOTIC.

Property	AffectNet-8	EMOTIC
Train Images	287,651	23,266
Validation Images	0	3,315
Test Images	3,999	7,203
Distinct Expressions	8	26
Valence	✓	✓
Arousal	✓	✓
Dominance	×	✓
Scale for valence, arousal, dominance	[-1, 1] (floats)	[1, 10] (integers)

Table 3. Comparison of AffectNet-8 and EMOTIC datasets.

competitive as shown by Savchenko *et al*. [41], more recent architectures like the POSTER++ [34] facilitate hybrid facial expression inference via networks that combine CNNs for feature extraction with vision transformer elements for efficient multi-scale feature integration and attention-based cross-fusion, achieving state-of-the-art performance with reduced computational cost. Because EMOTIC allows for multiple discrete labels for each individual, a general accuracy score is less applicable. Instead, Khan *et al*. [13] suggests the *top-k accuracy* can provide more insights. Utilizing a multi-modal approach leveraging region of interest heatmaps, a vision encoder, and a text encoder they achieve a top-3 accuracy of 13.73%. Khor Wen Hwooi *et al*. [12] suggested to extract features from CNNs and then apply model regression with a CultureNet [38] for the continuous prediction of affect from facial expression images within the *valence* and *arousal* space. The best results were achieved with DenseNet201 [11] for feature extraction. The work demonstrates superior performance in predicting *valence* and *arousal* levels, particularly on the AffectNet dataset.

3. Analysis of Datasets for Inference of Emotional Expressions

We assessed the predictive capabilities of AffectNet and EMOTIC (see Table 3), rating the dataset size, expression quantity, and the encoded dimension of the circumplex model. EMOTIC dataset has a much smaller data size whilst containing several more discrete expressions and offering the additional continuous value *dominance* in comparison to the AffectNet-8 dataset. In the following, we provide an in-depth analysis of the two datasets.

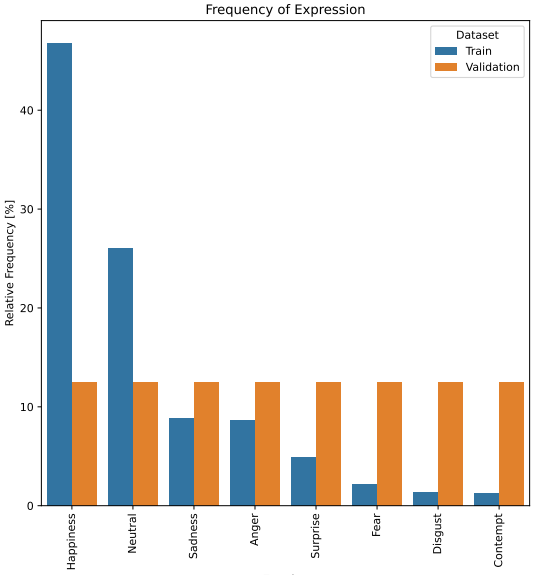


Figure 3. Frequency of expression of AffectNet.

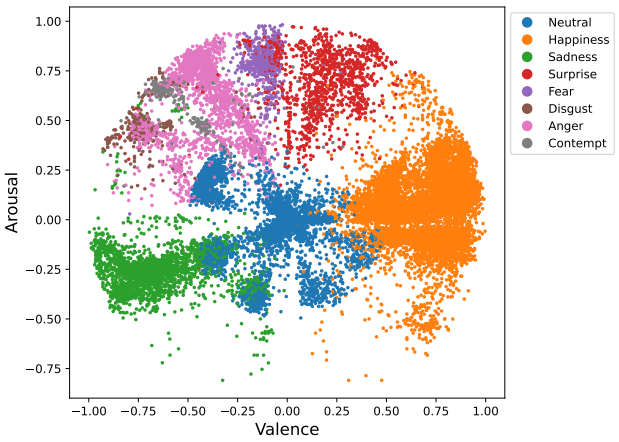


Figure 4. *Valence* and *Arousal* in a subset of the train dataset of AffectNet sorted by expression category.

3.1. AffectNet

Images in the AffectNet [36] dataset are labeled with (1) one out of eight possible discrete expression categories, (2) a *valence* value as a real number between -1 and 1, (3) an

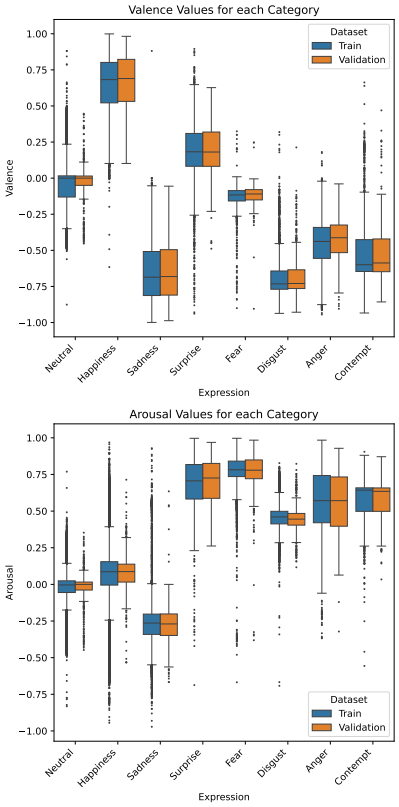


Figure 5. *Valence* and *Arousal* in AffectNet-8.

arousal value as a real number between -1 and 1, and (4) facial landmark points. The distribution of discrete categories (see Figure 3) is unbalanced. The sum of the probabilities would lead up to 70% only with two of the eight expression categories (*happiness* and *neutral*). On the other hand, validation data is evenly distributed across all labels.

To analyze the distribution of the continuous values *valence* and *arousal*, we displayed the values from the training set in the circumplex model of affect as originally proposed by Russell [39], with the values of *arousal* on the ordinate and *valence* on the abscissa (see Figure 4). As a result, the visualization clearly reveals that different expression categories can lead to an overlap in the *valence/arousal* values. Additionally, we analyze the distribution of *valence/arousal* per category as shown in Figure 5. For example, *neutral* and *happiness* expressions share a similar median in *arousal* dimension, whilst having a different median in the *valence* dimension. As expected, the *neutral* category is centered around zero for *valence* and *arousal*.

A comparison of the *valence* and *arousal* values across the train and validation datasets shows that imbalance is also present. In particular, more values from the training dataset are positive compared to the validation dataset with a higher portion of negative values (see Figure 6). This imbalance is

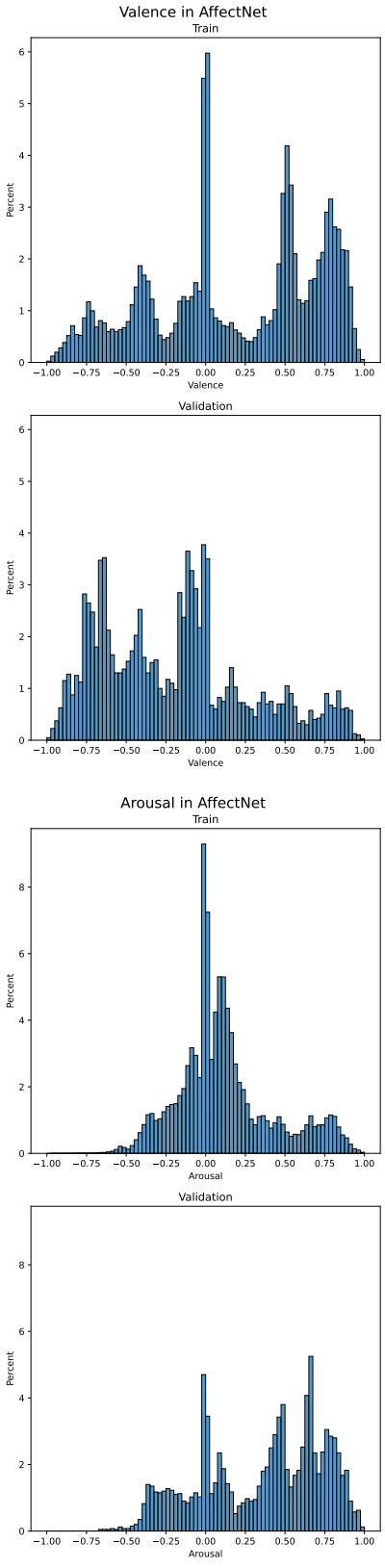


Figure 6. *Valence* and *Arousal* in AffectNet-8.

even more noticeable for arousal values. In the validation dataset, there are far more high-valued positive values compared to the training dataset.

3.2. EMOTIC

In EMOTIC [26], every image has a more complex labeling, targeting an overall context and a body focusing on the expression (see Figure 2). Each bounding box of a human is labeled with at least one expression, a *valence* value (integer between 1 and 10), an *arousal* value (integer between 1 and 10), a *dominance* value (integer between 1 and 10), gender (female/male) and age of a person (kid/teenager/adult).

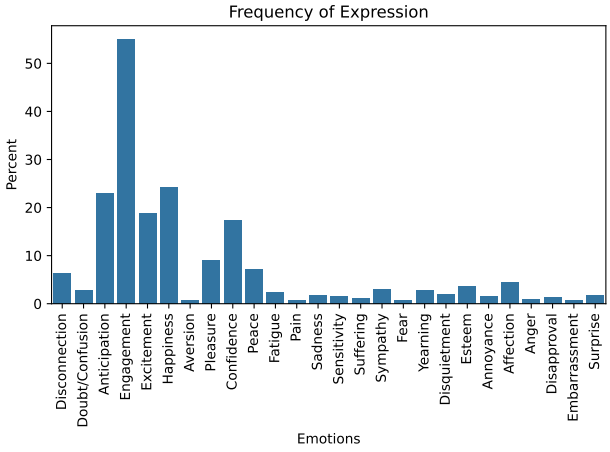


Figure 7. Distribution of expression frequency in EMOTIC train data.

Figure 7 shows the relative occurrence of each expression in the dataset. Due to the multi-labeling of the dataset, an image can have multiple labels. Also, the overall frequency of the expression *engagement* is dominating. Furthermore, all categories with a relative occurrence over 10% are "positive", corresponding to a positive *valence* value.

Analysis of the label frequency in subsets has shown, that the training dataset contains a lot of images with one, two, or three categories, consistently decreasing. On the other side, the validation and the training dataset have a lot of images labeled with four or more categories (see Figure 8).

In summary, the EMOTIC dataset leads to a much more challenging task to train a suitable computer vision model. A fairly small dataset size, multi-person context, multi-label encoding, inconsistent unbalanced datasets, and different expression frequencies have led us to much more severe effort in the choice of suitable model hyperparameters.

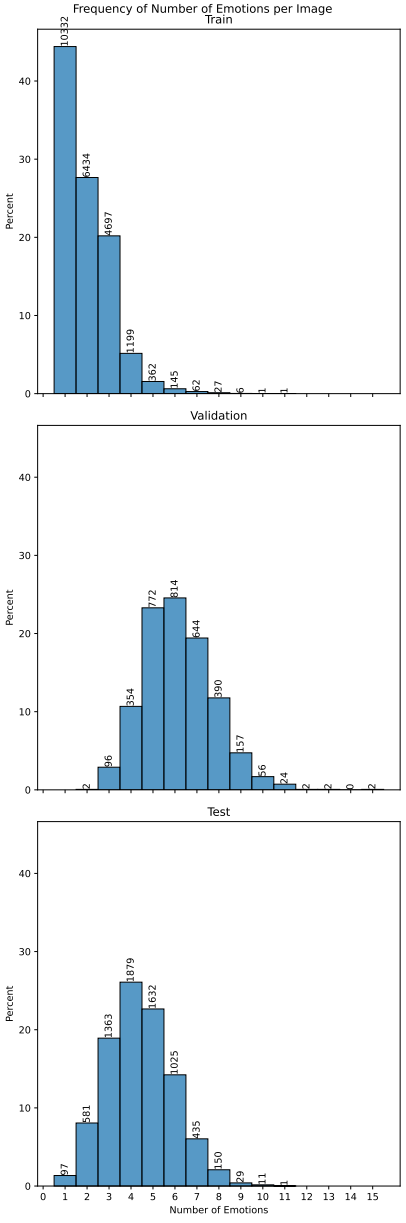


Figure 8. Instances of multiple true expressions per image in EMOTIC data.

4. Models for Discrete and Continuous Expression Inference

Starting with a state-of-the-art baseline model trained on the AffectNet data, we evaluated different approaches to check if combining the classification of discrete emotional expressions and regression for continuous *valence* and *arousal* values can lead to better results. After training and comparing our approaches on AffectNet, we use the architecture of the best model to train on the EMOTIC data.

4.1. Baseline Selection and Losses

To assess the impact of the usage of *valence/arousal* during training, we consider three model versions:

Discrete models with a cross-entropy loss. Due to an unbalanced class distribution, we used a weighted cross-entropy loss $L_{WeightedCE}$. The weights in the cross-entropy loss were set to the frequencies of expressions in the training dataset (see Table 4).

Combined models with an additional MSE loss for *valence* and *arousal*, weighted with a balance factor α:

$$L_{combined} = L_{WeightedCE} + \alpha \cdot L_{MSE} \quad (1)$$

Valence-arousal models with a CCC loss for the regression of continuous valence and arousal values:

$$L_{valence-arousal} = L_{CCC} + \beta \cdot L_{MSE} \quad (2)$$

Label	AffectNet-8	AffectNet-7
Neutral	0.015605	0.022600
Happiness	0.008709	0.012589
Sadness	0.046078	0.066464
Surprise	0.083078	0.120094
Fear	0.185434	0.265305
Disgust	0.305953	0.444943
Anger	0.046934	0.068006
Contempt	0.308210	/

Table 4. Weights for the cross-entropy loss.

4.2. Training Setup

The models proposed above were trained on the AffectNet data. Then, the best-performing model was selected for re-training on EMOTIC data. All training was performed using NVIDIA 4090 GPUs. Table 5 shows the hyperparameters.

Hyperparameter	Value
Batch Size	128
Learning rate	5e-5
Optimizer	AdamW [32]
Learning rate scheduler	Cosine annealing [31]
$L_{combined}$ factor	$\alpha = 5$
$L_{valence-arousal}$ factor	$\beta = 3$

Table 5. Hyperparameters for model training.

To train the proposed model architecture on AffectNet, we used the following data augmentation techniques:

- *RandomHorizontalFlip* with p=0.5,
- *RandomGrayscale* with p=0.01,
- *RandomRotation* with degree=10,
- *ColorJitter* with brightness=0.2, contrast=0.2, saturation=0.2 and hue=0.1,
- *RandomPerspective* with distortion=0.2 and p=0.5,
- *Normalize* with mean=[0.485, 0.456, 0.406] and std=[0.229, 0.224, 0.225],
- *RandomErasing* with p=0.5, scale=(0.02, 0.2), ratio=(0.3, 3.3) and value='random'.

Whilst most augmentation techniques target a more robust/stable model, we discovered that *RandomErasing* prevented model overfitting on the training dataset, which would otherwise occur due to the small dataset size. Based on the training behavior, we have chosen a comparably high batch size and a relatively small learning rate. We noticed, that even with this small learning rate, the proposed model achieved the best results in the fifth epoch. However, a change in the model architecture (more/less parameters, change in model architecture, different batch size, etc.) did not improve our results.

4.3. Evaluation Setup

Performance metrics for AffectNet: to address the dual nature of the proposed model, which integrates a classification- and/or a regression task, we evaluate its performance using state-of-the-art binary classification metrics as well as common regression metrics: precision P, recall R, F1 score F1, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE).

Performance metrics for EMOTIC: as mentioned above, we use the best model trained on AffectNet to retrain our model on the EMOTIC dataset. Because the EMOTIC dataset is a multi-label multi-classification dataset, a change of the loss is necessary. Hence, we changed the cross-entropy loss to a positive-weighted binary cross-entropy (BCE) loss, where the positive weights are defined as the inverse of the occurrence of each label.

$$\tilde{L}_{combined} = L_{WeightedBCE} + \alpha \cdot L_{MSE}$$

Cross-Validation of models: as both EMOTIC and AffectNet share the same dimension regarding *valence* and *arousal*, we test the proposed trained models on each test data. To achieve this, we transformed the dimension of *valence* and *arousal* to ensure its values are between -1 and 1, then evaluated the datasets/models' generalization on thoroughly unseen data samples.

5. Evaluation

In the following, we discuss and compare model performance on AffectNet and EMOTIC.

5.1. Model Architecture

For comparison, we have evaluated both CNN- and transformer-based architectures, while focusing on lightweight models: EfficientNetv2 [42] and MaxViT-Tiny [43] [37]. Furthermore, we have experimented with Swin

Dataset	Model	Precision ↑	Recall ↑	F1 ↑	MSE ↓	MAE ↓	RMSE ↓	CCC ↑
AffectNet-7	EfficientNetv2s$_{discrete}$	0.634	0.631	0.631	-	-	-	-
	MaxViT$_{discrete}$	0.640	0.639	0.638	-	-	-	-
	EfficientNetv2s$_{combined}$	0.650	0.646	0.646	0.0956	0.2298	0.3092	0.7636
	MaxViT$_{combined}$	**0.666**	**0.664**	**0.664**	0.0947	0.2251	0.3077	0.7640
	EfficientNetv2s$_{valence-arousal}$	-	-	-	**0.0833**	**0.2098**	**0.2887**	**0.8206**
	MaxViT$_{valence-arousal}$	-	-	-	0.0841	0.2121	0.2901	0.8196
AffectNet-8	EfficientNetv2s$_{discrete}$	0.605	0.599	0.599	-	-	-	-
	MaxViT$_{discrete}$	0.602	0.598	0.599	-	-	-	-
	EfficientNetv2s$_{combined}$	0.612	0.606	0.607	0.1420	0.2781	0.3769	0.6413
	MaxViT$_{combined}$	**0.623**	**0.621**	**0.621**	0.1370	0.2715	0.3701	0.6592
	EfficientNetv2s$_{valence-arousal}$	-	-	-	0.1028	0.2387	0.3206	0.7816
	MaxViT$_{valence-arousal}$	-	-	-	**0.1021**	**0.2351**	**0.3196**	**0.7840**

Table 6. Comparison of model performance on AffectNet. The best results for AffectNet-7 and AffectNet-8 are highlighted.

transformer [30] models. However, these have demonstrated worse results with precision below 0.35. PyTorch implementations of models, pre-trained on ImageNet [8] were used. The best results were achieved with the MaxViT models (see Table 6).

5.2. Impact of Training with Valence and Arousal on Discrete Expressions for AffectNet

A different model architecture and a combined training approach increased the baseline F1-score from 60% to 62% when using the AffectNet-8 dataset (see Table 6). With a reduced AffectNet-7 dataset, we also increased our model performance leading to an F1 score of 66%. The combined approach thus improved the classification results for both datasets by 2%.

Figure 9 displays the confusion matrix of the MaxViT$_{combined}$ for AffectNet-8. In accordance with the theory of the circumplex model of affect, the displayed transition of discrete emotional expressions is smooth. For example, *surprise* and *fear* have a similar median *arousal* value or *disgust* and *anger* around their corresponding *valence* value. A model using continuous values can thus potentially outperform the one with discrete values.

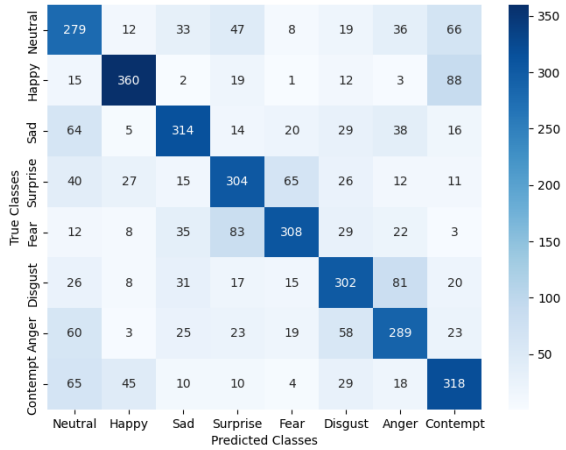


Figure 9. Confusion matrix for MaxViT$_{combined}$ on AffectNet-8.

5.3. Best AffectNet Model Regarding Valence and Arousal

In contrast to the proposed combined training methodology, the best regression results were gained with the MaxViT$_{valence-arousal}$ model. To reduce noticeable oscillating behavior during training, we reduced the training dataset by balancing according to the discrete expression labels. Furthermore, we added the CCC loss to the L$_{valence-arousal}$ loss function and used the pre-trained weights of the best model from the combined method (MaxViT$_{combined}$). With a duration of two minutes per epoch, the best results were achieved in epoch seven.

Metric	VGG-F [4]	Ours	Improvement
RMSE$_{valence}$ ↓	0.356	**0.331**	7,0%
RMSE$_{arousal}$ ↓	0.326	**0.305**	6,4%
CCC$_{valence}$ ↑	0.710	**0.716**	0,8%
CCC$_{arousal}$ ↑	0.629	**0.642**	2,0%

Table 7. Benchmark vs. MaxViT$_{combined}$ for AffectNet-8

Figure 10 shows the percentage of data points within the absolute error range. When focusing on the ordinate, 80% of the *valence* and *arousal* predictions differ only ±0.3 from their true value. The resulting model is thus more robust.

5.4. Performance of the EMOTIC Model

For the EMOTIC dataset, we calculated the positive weights for each label for the training and validation dataset. Similar to the weights of the cross-entropy loss, the positive weight for each class is the inverse of the overall occurrence of a label. We chose this method to compensate for the imbalance

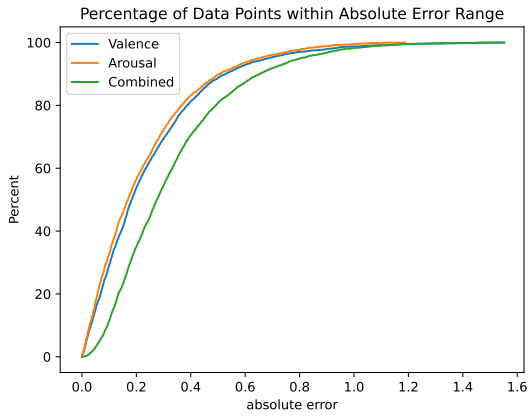


Figure 10. Absolute error for MaxViT$_{combined}$ trained on AffectNet-8.

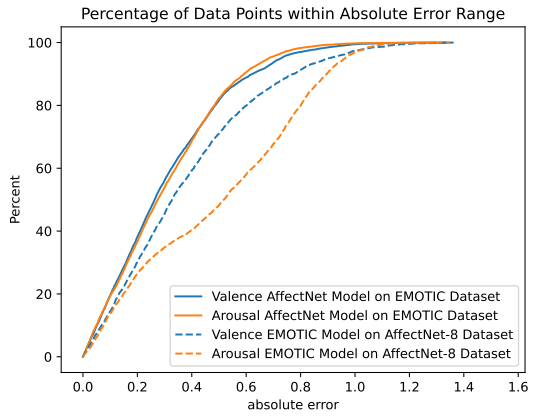


Figure 11. Absolute error for cross-validation of the MaxViT models.

of the frequency of expression (see Figure 7).

Table 8 shows the overall metrics of our best EMOTIC model. To compare the RMSE of *valence* and *arousal* with the AffectNet dataset, we added a scaled version. For this, integers between 1 and 10 from EMOTIC are scaled to real values between -1 and 1. By definition, the CCC is invariant to shifts and scale transformations. Additionally, our EMOTIC model outperforms the best model by Khan *et al*. [13] according to Top-3 accuracy, which reaches 13.73 % on the benchmark [7].

Metric	**Ours** (Original)	**Ours** (Scaled)
Top-3 Accuracy	14.73 %	/
RMSE$_{valence}$ ↓	1.150	**0.256**
RMSE$_{arousal}$ ↓	1.209	**0.269**
RMSE$_{dominance}$ ↓	1.169	**0.260**
CCC$_{valence}$ ↑	0.316	**0.316**
CCC$_{arousal}$ ↑	0.594	**0.595**
CCC$_{dominance}$ ↑	0.300	**0.301**

Table 8. Performance of MaxViT$_{combined}$ on EMOTIC.

5.5. Cross-Validation of the Models for Valence

We assess the performance of the proposed trained AffectNet/EMOTIC model on the respective test datasets. The analysis of cross-validation results revealed that AffectNet outperformed EMOTIC notably in terms of absolute error metrics when evaluated on the AffectNet dataset.

As Figure 11 shows, the absolute errors for *valence* and *arousal* are significantly higher. Thus, the substantial advantage of AffectNet over EMOTIC in absolute error rates during cross-validation stresses its generalization ability for expression inference tasks.

6. Conclusion & Outlook

In this paper, we assessed the capability of discrete classifier approaches with multi-task learning models when inferring emotional expressions. We used two prominent datasets tailored for discrete expressions and values based on the circumplex model of affect to train our models.

Firstly, we have performed in-depth analysis of the datasets. It was observed that while test datasets are often balanced concerning emotional expressions, the balance is not maintained for *valence* and *arousal*. Models trained solely on *valence* and *arousal* tend to minimize errors. Additionally, it is noteworthy to delve into the intricate distribution of the EMOTIC dataset, especially how it varies concerning the number of classes in the train and test sets.

Secondly, we proposed to use the MaxViT model architecture and described the training and evaluation protocol for both datasets. The proposed approach significantly improved model accuracy. Even in cases of misclassification, the predicted *valence* and *arousal* values often remained accurate. Establishing a threshold for correct prediction of *valence* and *arousal* poses an interesting challenge for future work, as it involves considering factors such as human error and the inherent complexity of emotional expression perception. Furthermore, our model based on AffectNet demonstrated robust performance in *valence* and *arousal* estimation via cross-validation. This suggests the potential for it to serve as a well-generalized model. Conversely, the performance of our EMOTIC-based approach was less conclusive, possibly due to insufficient data or other factors.

In conclusion, our research underscores the effectiveness of continuous value approaches within multi-task learning frameworks for emotional expression inference. Further exploration and refinement of these methodologies could yield even more accurate and robust models in the future.

References

[1] Panagiotis Antoniadis, Panagiotis Paraskevas Filntisis, and Petros Maragos. Exploiting emotional dependencies with graph convolutional networks for facial expression recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2021. 2

[2] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 2019. 1

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *International Conference on Multimodal Interaction*. Association for Computing Machinery, 2016. 2

[4] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. *CoRR*, 2022. 7

[5] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *CoRR*, 2023. 2

[6] Papers With Code. Facial expression recognition on affectnet. https://paperswithcode.com/sota/facial-expression-recognition-on-affectnet, 2024. 2

[7] Papers With Code. Emotion recognition on emotic. https://paperswithcode.com/sota/emotion-recognition-on-emotic, 2024. 8

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009. 7

[9] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference, on Neural Information Processing (ICONIP)*. Springer, 2013. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[12] Stephen Khor Wen Hwooi, Alice Othmani, and Aznul Qalid Md. Sabri. Deep Learning-Based Approach for Continuous Affect Prediction From Facial Expression Images in Valence-Arousal Space. *IEEE Access*, 2022. 3

[13] Muhammad Saif Ullah Khan, Muhammad Ferjad Naeem, Federico Tombari, Luc Van Gool, Didier Stricker, and Muhammad Zeshan Afzal. Focusclip: Multimodal subject-level guidance for zero-shot transfer in human-centric tasks. *CoRR*, 2024. 2, 3, 8

[14] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[15] Dimitrios Kollias. Abaw: learning from synthetic data & multi-task learning challenges. In *European Conference on Computer Vision (ECCV)*. Springer, 2023.

[16] Dimitrios Kollias. Multi-label compound expression recognition: C-expr database & network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *CoRR*, 2019.

[18] Dimitrios Kollias and Stefanos Zafeiriou. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. *CoRR*, 2019. 2

[19] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *CoRR*, 2021.

[20] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[21] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *CoRR*, 2019.

[22] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision (IJCV)*, 2019.

[23] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020.

[24] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *CoRR*, 2021.

[25] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[26] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. EMOTIC: Emotions in Context Dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 2, 5

[27] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[28] David Lapakko. Communication is 93% nonverbal: An urban legend proliferates. *Communication and Theater Association of Minnesota Journal*, 2015. 1

[29] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *CoRR*, 2016. 6

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *CoRR*, 2017. 6

[33] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2010. 2

[34] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster++: A simpler and stronger facial expression recognition network. *CoRR*, 2023. 2, 3

[35] Albert Mehrabian and Morton Wiener. Decoding of inconsistent communications. *Journal of personality and social psychology*, 1967. 1

[36] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A new database for facial expression, valence, and arousal computation in the wild. *IEEE Transactions on Affective Computing*, 2017. 1, 2, 3

[37] PyTorch. Models and pre-trained weights - maxvit: Multi-axis vision transformer. https://pytorch.org/vision/main/models/maxvit.html, 2024. 6

[38] Ognjen Rudovic, Yuria Utsumi, Jaeryoung Lee, Javier Hernandez, Eduardo Castelló Ferrer, Björn Schuller, and Rosalind W Picard. Culturenet: a deep learning approach for engagement intensity estimation from face images of children with autism. In *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018. 3

[39] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980. 1, 4

[40] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 2022. 2

[41] Andrey V. Savchenko, Lyudmila V. Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022. 2, 3

[42] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning (ICML)*. PMLR, 2021. 6

[43] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *CoRR*, 2022. 6

[44] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 2017. 2

[45] Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 2023. 2