FindingEmo: An Image Dataset for Emotion Recognition in the Wild

Laurent Mertens[1,2] **Elahe' Yargholi**[3] **Hans Op de Beeck**[3]

Jan Van den Stock[4] **Joost Vennekens**[1,2,5]

[1]KU Leuven, De Nayer Campus, Dept. of Computer Science
J.-P. De Nayerlaan 5, 2860 Sint-Katelijne-Waver, Belgium
[2] Leuven.AI - KU Leuven Institute for AI, 3000 Leuven, Belgium
[3]Department of Brain and Cognition, Leuven Brain Institute,
Faculty of Psychology & Educational Sciences
KU Leuven, 3000 Leuven, Belgium
[4]Neuropsychiatry, Leuven Brain Institute
KU Leuven, 3000 Leuven, Belgium
[5]Flanders Make@KU Leuven, 3000 Leuven, Belgium
laurent.mertens@kuleuven.be

Abstract

We introduce FindingEmo, a new image dataset containing annotations for 25k images, specifically tailored to Emotion Recognition. Contrary to existing datasets, it focuses on complex scenes depicting multiple people in various naturalistic, social settings, with images being annotated as a whole, thereby going beyond the traditional focus on faces or single individuals. Annotated dimensions include Valence, Arousal and Emotion label, with annotations gathered using Prolific. Together with the annotations, we release the list of URLs pointing to the original images, as well as all associated source code.

1 Introduction

Computer vision has known an explosive growth over the past decade, most notably due to the resurgence of Artificial Neural Networks (ANNs). For many vision-related tasks, computer models have been developed that match or exceed human performance, e.g., image classification [1] and mammographic screening [2]. Many of these tasks, however, are relatively simplistic in nature: detecting the absence or presence of an object, or naming an item in the picture. When it comes to more complex tasks, Artificial Intelligence (AI) still has a long way to go. Affective Computing [3], a field that combines disciplines such as computer science and cognitive psychology to study human affect and attempt to make computers understand emotions, is an example of such a complex problem. This paper is concerned in particular with the subtask of Emotion Recognition, i.e., building AI models to recognize the emotional state of individuals, in our case from pictures. This problem has many applications, ranging from psychology [4], to human-computer interaction [5], to robotics [6]. It is, however, complex: in the field of psychology, the concept of what an emotion *is* exactly is heavily debated [7, 8, 9], resulting in several ways of describing emotions, either by means of continuous dimensions [10, 11], or by means of labels, with different competing label classification schemes existing [12, 13, 14].

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

The application of computer vision techniques toward Emotion Recognition has historically largely focused on detecting emotions from human facial expressions, with the problem still being actively investigated [15, 16, 17, 18, 19, 20, 21]. However, the importance of *context* in emotion recognition is increasingly being acknowledged in psychology [22, 23]. This led to the release of the computer vision dataset EMOTIC [24], presenting photos of people in natural settings, rather than face-focused close-ups, and leading the way to more complex ANN systems that attempt to combine multiple information streams extracted from these images [25, 26, 27].

Figure 1: Photo courtesy The Kitcheners (`https://thekitcheners.co.uk/`).

Nevertheless, even these more recent efforts focus on the emotional state of one particular individual within the picture. In this paper, we present the FindingEmo dataset, which is the first to target higher-order social cognition. The dataset was developed as part of an interdisciplinary project in which researchers from the fields of Psychology, Psychiatry and Computer Science investigate the use of ANNs to simulate Social Cognition, as a way to better understand the corresponding mechanisms in the human brain, and how these mechanisms are affected by conditions that correlate with atypical social behavior, in particular ASD and FTD [28]. Each image in the dataset depicts multiple people in a specific social setting, and has been annotated for the *overall* emotional content of the *entire* scene, instead of focusing on a single individual. We hope this data can be used by AI practitioners and psychologists alike to further the understanding of Emotion Recognition, and more broadly, Social Cognition. This is a complex process, consisting of many layers. Consider, e.g., the photograph depicted in Figure 1. Looking only at the bride's face, one could easily assume she is very sad, or even distressed. Taking also her wedding gown into account, a positive setting is suddenly suggested; perhaps her tears are tears of joy? Only when looking at the full picture does it become clear that the bride is overcome with emotion in a positive way, as conveyed by the setting, the groom reading a prepared text and the clearly supportive bystanders. Thus, full understanding of the bride's emotional state requires the full scene, including the groom and the solemnly smiling bystanders. This example illustrates how Social Cognition involves detection of relevant elements, extracting relations among these and attributing meaning to construct a coherent whole.

The source code for the scraper and annotation interface used to create the dataset are available from our dedicated repository[1], together with the URLs of the annotated images and their corresponding annotations. To mitigate the issue of broken URLs, we provide multiple URLs for a same image whenever possible, and are continuously expanding the set of images for which multiple URLs are provided (about 10k so far). For copyright reasons, we do not share the images themselves. More information with regard to legal compliance can be found in §A.2.

The data collection process was approved by the KU Leuven Ethics Committee.

The remainder of the paper is structured as follows. In Section 2 the data collection process and dataset are described in detail. Next, baseline results for emotion classification and valence and arousal regression problems based on popular ImageNet ANN architectures, as well as Visual Transformers CLIP and DINOv2, are presented in Section 3. We build upon this by investigating the effect of merging the features and predictions of several models in Section 4. Finally, we conclude with a discussion in Section 5.

[1] `https://gitlab.com/EAVISE/lme/findingemo`

2 Dataset Description

The dataset is split into a publicly released set of annotations for 25,869 unique images, and a privately kept set of 1,525 images.[2] Each image depicts multiple people in various, naturalistic, social settings. We follow Emotic [24] in creating a training (=our public) set with one annotation per image, and a test (=our private) set with multiple annotations per image. In total, 655 participants—a short description of whom can be found in §A.8—contributed annotations. In what follows, we list the most important annotation dimensions; for a full list, see §A.3.

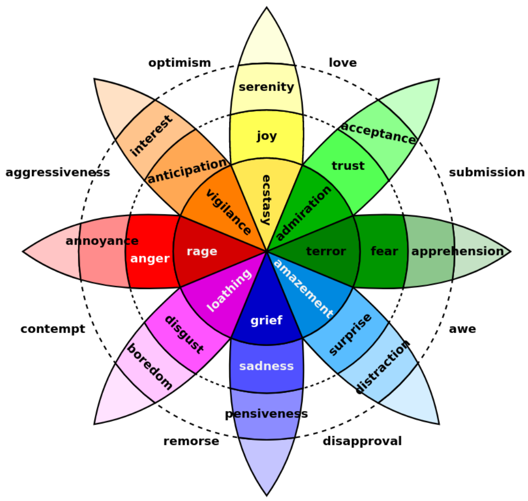

Figure 2: Plutchik's Wheel of Emotions.

Valence and Arousal We used Russell's continuous Valence and Arousal dimensions [10], with integer scales $[-3, -2, \ldots, 3]$ for Valence and $[0, 1, \ldots, 6]$ for Arousal. Arousal was named "Intensity" in our annotation interface, as we felt "Arousal" might carry a sexual connotation for some users.

Emotion Users had to pick an emotion from Plutchik's discrete Wheel of Emotions (PWoE) [13], shown in Figure 2. We opted for this particular emotion classification scheme as it strikes a balance between the more limited and sometimes contested Ekman's 6 [12], and the more expansive, and potentially more confusing, Geneva Emotion Wheel [14]. It defines 24 primary emotions, grouped into 8 groups of 3, with emotions within a group differing in intensity. It is depicted as a flower with the 24 emotions organized in 8 leaves and 3 concentric rings. Each leaf represents a group of 3, with opposite leaves representing opposite emotions. The rings represent the intensity levels, from most intense at the center to least intense at the outside. An additional advantage of PWoE is that one can easily opt to use all 24 emotions, or instead limit oneself to the 8 groups, allowing some granularity control. We refer to these choices as "Emo24" and "Emo8" respectively, and refer to the groups as "emotion leaves".

2.1 Positioning Versus Existing Datasets

Although research in automated Emotion Recognition has been gaining in popularity over the years, progress is still hampered by a lack of data. Earlier work tended to focus solely on recognizing emotions from faces. In their recent review paper, Khare et al. [29] list no less than 21 publicly available datasets of facial images for this purpose, typically annotated with Ekman's 6, potentially extended with a "neutral" category, or custom defined emotion categories. Some of the more popular such datasets, like JAFFE [30] and CK+ [31], make use of a limited number of actors (resp. 10 and 123) who were instructed to act out a certain emotion, resulting in caricatural emotional expressions.

Publicly available datasets going beyond the face are few in number. First, there is EMOTIC [24], a 23,571 image dataset depicting people in the wild, and with natural expressiveness. An explicit goal of EMOTIC is to take context into account when assessing a person's emotional state. One or more individual subjects are delineated by a bounding box in each picture for a total of 34,320 subjects, each annotated for Valence, Arousal, Dominance and one of 26 custom defined emotion categories.

CAER-S is a dataset of 70,000 stills taken from 79 TV shows. The stills were extracted from 13,201 video clips that were annotated for Ekman's 6 + neutral. Each still contains at least one visible face. The aim of the dataset is to allow augmenting facial emotion recognition with contextual features.

Similar to EMOTIC, there is HECO, a dataset of 9,385 images taken from previously released Human-Object Interaction datasets, films and the internet. Like EMOTIC, 19,781 individual subjects were

[2]This set is kept private to allow us to use it as a test set for dedicated workshops organized at a later date.

annotated in the pictures for Valence, Arousal, Dominance, 8 discrete emotion categories comprised of Ekman's 6 + Excitement and Peace, and two novel dimensions, Self-assurance and Catharsis.

Table 1 groups these dataset descriptions, together with ours, for easy comparison.

Table 1: Comparison of relevant datasets. "V/A/D" indicates which of the Valence, Arousal and Dominance dimensions were annotated.

Name	Nb. images	Image source	Annotation target	V/A/D	Emotions scheme	Reference
EMOTIC	23,571	COCO + Ade20k + internet	Single person	V/A/D	26 custom emotion categories	[24]
CAER-S	70,000	TV Shows	Single person (face visible)	–	Ekman's 6 + neutral	[32]
HECO	9,385	HICO-DET + V-COCO + film + internet	Single person	V/A/D	Ekman's 6 + Excitement and Peace	[27]
FindingEmo	25,869	Internet	Whole image	V/A	Plutchik's Wheel of Emotions	This paper

2.2 Dataset Creation Process

The creation of the dataset was split into two phases. The first phase focused on gathering a large set of images, *prioritizing quantity over quality*. The second phase consisted of collecting the annotations. We present a brief summary of both phases here, and refer to §A.4 for more details.

Phase 1 Images were gathered using a custom built image scraper that generates random search queries, each consisting of three terms selected from predefined lists of, respectively, emotions, groups of people (e.g., "adults", "seniors", etc.) and social settings/environments. For each query, the first N results were retrieved, filtered and downloaded. As obviously not all downloaded images satisfied our criterion of depicting multiple people in a natural setting, one particular filtering step involved labeling and classifying images as being either "keep" (useful) or "reject" (no use). In total 1,041,105 images were collected.

Phase 2 Annotations were gathered using a custom web interface (see §A.5 for a screenshot). Annotators were recruited through the Prolific[3] platform, and first required to agree to an Informed Consent clause, followed by detailed instructions (see §A.6 for a copy). To monitor the process closely, we performed many (51, to be exact) runs, each with a limited number (around 10 to 15) of participants. For each run, the Prolific user selection criteria were the same: fluent English speaker, (self-reported) neurotypical[4], and a 50/50 split male/female. Candidates were informed of a total expected task duration of 1h, and offered a $\pounds 10$ reward. Analysis of the durations (see §A.7) show our time estimation to be fair. In total, data collection costs were $\pounds 10$k, including fees and taxes.

2.3 Annotator Grading and Annotator Overlap

To assess the reliability of annotators, we used a set of 5 fixed images, referred to as "fixed overlap images", chosen specifically for being unambiguous.[5] For each image, a default annotation was defined consisting of the "keep/reject" choice (4 keeps, 1 reject), Valence (value range), Arousal (value range) and Emotion (emotion leaf). This results in 4 datapoints per image, or 20 datapoints in total. Annotators' submissions for these images were compared to the reference, earning 1/20 point per matching datapoint, resulting in a final "overlap score" $s \in [0, 1]$. Users with $s >= 0.8$ were automatically accepted. An alternative score s_{alt} was computed which ignored those overlap images whose reference value was "keep", but were annotated as "reject". The reason for this is that it quickly became clear that despite the system providing a "Skip" option in case users rather not annotate a certain image, some chose to "reject" these images instead. Also, one of the "keep" images shows a bit of text, which users were instructed to reject. Some users were more strict than others in applying this rule.

We defined a system parameter p_R that controls when overlap images (i.e., images already annotated by others) are shown to users. For each new image request, an overlap image

[3] `https://www.prolific.com/`

[4] Prolific users specify themselves whether or not they are neurotypical in their profile; we did not perform any screening ourselves.

[5] This amounts to an average of 10% of the shown images, similar to Emotic [24] who "randomly [inserted] 2 control images in every annotation batch of 20 images".

is served with probability p_R, starting with the 5 fixed overlap images, in a fixed sequence. Once these are annotated, the system serves other, non-fixed, already annotated images. At first, these were randomly chosen from all annotated images, but this resulted in too many images with only 2 annotations. Hence, we created a process that limits the pool of images to choose from, and attempts to strive for 5 annotations per (non-fixed) overlap image. Using this system, we obtained a dataset with 80.9/19.1 split single label/multi-label annotations. These multi-label images make up the private set. Detailed inter-annotator statistics on this private set are reported in §A.9, indicating that for 26.2% of the images, all annotators agreed on the emotion leaf, while for 46.6% of the images two labels were given. Out of these two-label annotations, 42.8% refer to adjacent emotion leafs. Annotators agree less on Arousal (average min-max difference of 2.7 $\pm$ 1.4) than on Valence (average min-max difference of 1.8 $\pm$ 1.2). Importantly, average Valence disagreement plateaus close to 2 with increasing number of annotations per image, while a linearly increasing trend is apparent for Arousal.

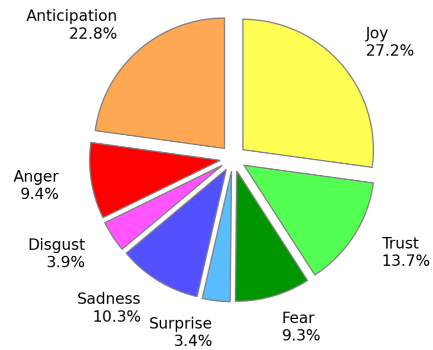

Figure 3: Distribution of Emotion annotations for the public set per Plutchik emotion leaf.

2.4 Statistics and Observations

This section presents statistics for the 8 leaves of PWoE. For the full 24 emotions, see §A.10.

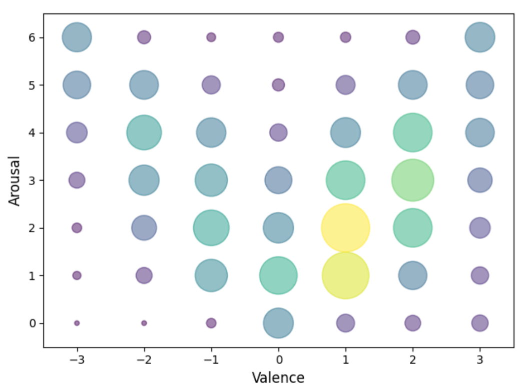

Figure 4: Association between Valence and Arousal values. The bigger the disc, the more often the (Valence, Arousal)-pair appears in the dataset.

Figure 3 shows the distribution of annotations per emotion leaf. An imbalance is obvious, with in particular "joy" and "anticipation" being overrepresented, and "surprise" and "disgust" heavily underrepresented, despite an added balancing mechanism (see §A.4.2). A similar imbalance is found in popular facial expression datasets, such as FER2013 [33] (only 600 "disgust" images versus nearly 5,000 for other Ekman's 6 labels) and AffectNet [34] (134,915 "happy" faces, 25,959 "sad" faces, 14,590 "surprise" faces, 4,303 "disgust" faces). Although EMOTIC [24] uses custom emotion labels, making a one-to-one comparison more difficult, it is also heavily skewed towards positive labels (top 3: "engagement", "happiness" and "anticipation"; bottom 3: "aversion", "pain" and "embarassement"). Compared to these other datasets, ours exhibits less imbalance.

In Table 2, we group average annotation values for Arousal, Valence and Ambiguity per emotion leaf. Figure 20 in §A.10 plots the distribution of Arousal and Valence annotations per emotion leaf, showing clear tendencies toward normal distributions, validating the use of averages and standard deviations. As expected, perceived "negative" emotions ("fear", "sadness", "disgust" and "anger") have a negative average Valence, with the inverse being true for "positive" emotions ("joy", "trust"). Somewhat undecided are "surprise" and "anticipation", which can go either way. The highest Arousal values are reserved for "anger, "sadness" and "fear". We hypothesize the unexpectedly high Arousal value for "sadness" might be due to naming this dimension "Intensity" in our interface; although a grieving person is generally considered to have low arousal, the emotion of sadness itself is felt intensely. Further analysis on the full emotion set reported in §A.10 verifies that also at this more fine-grained level, annotations conform to expectations, with Arousal levels increasing along with the intensity level of the PWoE ring, and Valence levels analogously increasing for "positive" and decreasing for "negative" emotions.

Table 2: Average Arousal, Valence and Ambiguity annotation values for the public set, per emotion leaf. Format x^y: x = average, y = standard deviation.

	Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
Nb.	7026	3549	2401	888	2665	1000	2439	5901
Arousal	$2.96^{0.96}$	$2.57^{1.09}$	$3.24^{1.24}$	$2.57^{1.41}$	$3.42^{1.29}$	$2.44^{1.23}$	$3.59^{1.17}$	$2.46^{1.21}$
Valence	$1.90^{0.96}$	$1.41^{1.09}$	$-1.34^{1.24}$	$0.48^{1.41}$	$-1.57^{1.29}$	$-0.88^{1.23}$	$-1.58^{1.17}$	$0.56^{1.21}$
Ambiguity	$1.58^{1.66}$	$1.88^{1.64}$	$2.09^{1.61}$	$2.39^{1.68}$	$1.84^{1.66}$	$2.22^{1.65}$	$1.99^{1.63}$	$2.15^{1.61}$

Figure 4 shows the association between Arousal and Valence annotations, indicating as expected a collinearity between higher Arousal values and the extremes of the Valence range. Scatterplots per opposite Emo8 pairs are grouped in Figure 21 in §A.10.

2.5 Cross-cultural Analysis

Most Prolific users participating in our task shared their country of birth and ethnicity with us (see A.8). To verify to what extent annotations are consistent accross people of different backgrounds, we performed the following two experiments.

To check the consistency between geographic regions, we first mapped countries of birth to the geographic regions they are embedded in. For pairs of regions with at least 100 common images (i.e., images annotated by members of both regions), we analysed the distribution of, and agreement between Arousal, Valence and Emo8 annotations. Seven pairs were left: Central Europa (C.Eur.)–Southern Africa (S.Afr.), C.Eur–Western Europe (W.Eur.), Eastern Europe (E.Eur.)–S.Afr, E.Eur–W.Eur, North America (N.Am.)–S.Afr, N.Am–W.Eur. and S.Afr.–W.Eur. We computed the similarity between the distributions, as well as inter-annotator agreement between annotation vectors for each region in the pair, where the average annotation value was used in case of multiple annotations from the same region, with results grouped in Figure 17 and Table 4, §A.9. For all pairs and all annotation dimensions, the Jensen-Shannon (JS) distance between the distributions stayed within the range $[0.040, 0.229]$, and all passed the two-sample Kolmogorov-Smirnov (KS) test ($p > 0.95$) except for Arousal in C.Eur.–S.Afr. and N.Am.–S.Afr., and Valence in N.Am.–S.Afr. Spearman's R between all pairs and dimensions was significant ($p \ll 0.05$) and varied in the range $[0.170, 0.617]$, except for Arousal in C.Eur.–W.Eur. (0.102, $p = 0.203$) and N.Am.–S.Afr. (0.116, $p = 0.168$). Highest values were observed for Valence, lowest for Arousal. Overall, although there are differences between regions, tendencies are clearly similar.

We performed the same experiment based on users' ethnicities, resulting in 5 ethnicity pairs: Black–Mixed, Black–Other, Black–White, Mixed–White and Other–White. The resulting plots and metrics are grouped in Figure 18 and Table 5 respectively, both in §A.9. Interestingly, for Arousal 3 out of 5 pairs fail the KS test, namely Black–Mixed, Black–Other and Other–White. For Emo8, all pairs pass the test, while for Valence only Black–Other fails it. For both Valence and Emo8 all pairs have a significant Spearman's R ($p < 0.001$) in the range $[0.433, 0.590]$ for Valence and $[0.255, 0.346]$ for Emo8, while for Arousal Spearman's R is not significant for Black–Mixed (-0.061, $p = 0.476$) and Black-Other (0.112, $p = 0.220$). In short, Arousal annotations appear more consistent among geographic regions than among ethnicities, although it is important to note that the low number of datapoints does not allow for strong conclusions.

3 Baseline Model Results

Baseline results are obtained by applying transfer learning to popular ImageNet-based ANN architectures AlexNet [35], VGG16 [36], ResNet 18, 50 and 101 [37] and DenseNet 161 [38].[6] For each, we use the default PyTorch implementations and weights, and replace the last layer with a new output layer that matches the chosen task (see below). Only this last layer is trained. We do the same experiment for some of these same architectures trained from scratch on the Places365 dataset [39], using the official PyTorch models. We also consider EmoNet [40], a model for labeling images with one out of 20 custom emotion labels reflecting the emotion elicited in the observer, obtained by applying transfer learning to AlexNet and trained on a private database. In this case, we first process

[6] Our GitLab repository contains all logs used to generate all reported results, and includes additional results for models like VGG19, ResNet34 and DenseNet121, that were in line with other same-architecture models.

the image with EmoNet, and then send the resulting 20-feature vector through a new linear layer. We use the EmoNet PyTorch port by the main author[7]. Lastly, we also use Visual Transformer models CLIP [41] (ViT-B/32) and DINOv2 [42] (ViT-B/14 distilled with registers)[8], using both models to obtain embeddings for input images, and like with EmoNet, use these as input to a single linear layer.

We distinguish three tasks: *Emo8 classification*, where we predict one of the 8 primary emotions defined by the emotion leaves of PWoE; *Arousal regression*, where we predict the numerical arousal value; *Valence regression*, where we predict the numerical valence value.

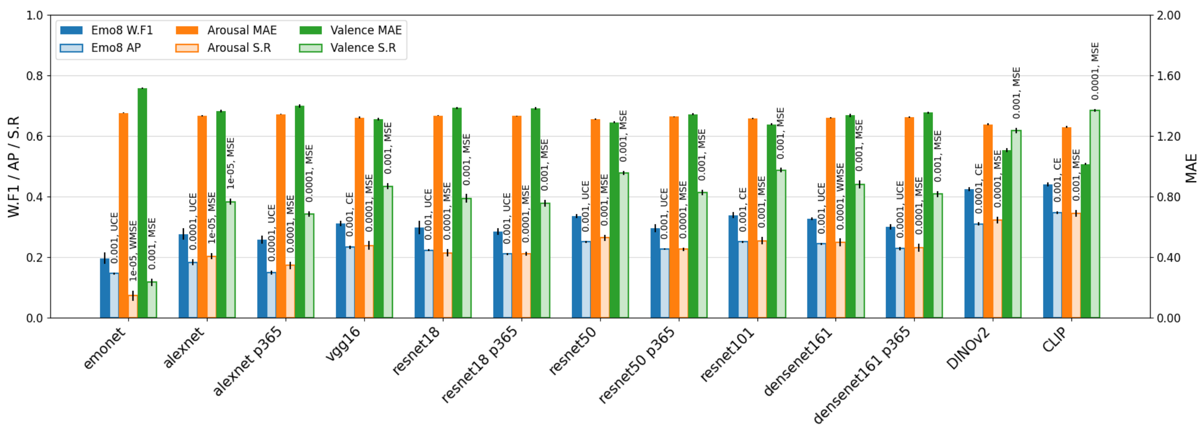

Figure 5: Test data baseline performance on the Emo8 classification and Arousal and Valence regression tasks. Metrics are: Weighted F1 (W.F1) and Average Precision (AP) for classification, and Mean Absolute Error (MAE) and Spearman R correlation coefficient (S.R) for regression. The starting learning rate and loss corresponding to each model are displayed above the training bars. (U)CE = (Unbalanced)CrossEntropyLoss, (W)MSE = (Weighted)MeanSquaredError loss, p365 = original model trained on Places365 dataset.

For classification, we apply a softmax to the output of the final layer. Target values for regression problems are reduced to the range $[0, 1]$ using an appropriate linear rescaling. Hence, we apply a sigmoid function to the model output. Network outputs are transformed back to the original problem domain by using the inverse scaling.

Preprocessing for ImageNet models consisted in scaling images to an 800x600 resolution, keeping the original ratio and centering and padding with black borders where necessary, followed by normalization using default ImageNet pixel means and standard deviations. For Places365 and EmoNet models, we followed the preprocessing steps described in the respective papers. For CLIP, we use the default preprocessing chain that comes with the model, and for DINOv2 we use the same preprocessing as for the ImageNet models, but with a rescaling to 798x602.

For each task, and each model, we trained 10 models per starting learning rate lr_0 and per loss function $\mathcal{L}$. For classification, we used $\mathrm{lr}_0 \in [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$ and $\mathcal{L} \in [\mathrm{CrossEntropyLoss}, \mathrm{UnbalancedCrossEntropyLoss}]$; for regression we used $\mathrm{lr}_0 \in [10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}]$ and $\mathcal{L} \in [\mathrm{MSELoss}, \mathrm{WeightedMSELoss}]$. UnbalancedCrossEntropyLoss is a novel extension of the traditional CrossEntropyLoss, created to allow giving different weights to different misclassifications. WeightedMSELoss is a natural extension of MSELoss that takes into account class imbalance. Full technical details for both can be found in §A.11.

All experiments use the *public* dataset, Adam loss with default PyTorch parameter values, and the custom lr update rule $\mathrm{lr}_e = \mathrm{lr}_0 / \sqrt{(e//3)+1}$, with lr_e the learning rate at epoch e. By virtue of the floor division ($//$), this means we update the learning rate once every 3 epochs. The data was randomly split 80/20 train/test, making sure that each target label was also split according to this same rule.

[7] `https://gitlab.com/EAVISE/lme/emonet`

[8] More specifically the 'Pretrained heads for image classification', loaded in PyTorch using `torch.hub.load('facebookresearch/dinov2', 'dinov2_vitb14_reg_lc')`. We also experimented with the smaller `vits14` variant, obtaining results typically a few percentage points behind the `vitb14` model.

Reported metrics are: for *classification*, Average Precision (AP)—as computed using the `scikit-learn` package—and Weighted F1 (W.F1); for *regression*, Mean Average Error (MAE) computed in the original problem domain, and Spearman Rank Correlation (S.R). Training stopped when either the epoch with the best loss (or the best W.F1 score for classification) on the test set lies 6 epochs behind the current epoch, or 250 epochs were reached, with the corresponding best model put forward as the final trained model. Only results for the $(\mathrm{lr}_0, \mathcal{L})$-combination yielding the best average Weighted F1 or Mean Average Error performance over the corresponding 10 models are reported.

All our experiments were implemented in Python using PyTorch, and split over an Intel Xeon W-2145 workstation with 32GB RAM and two nVidia GeForce RTX 3060 GPUs with 12GB VRAM, and an Intel i7-12800HX laptop with 32GB RAM and an nVidia GeForce RTX 3060 Laptop GPU with 12GB VRAM. Test results are plotted in Figure 5, with the graph for train data, and tables containing the numerical results grouped in §A.12. In order to speed up training, we buffered model activations whenever possible.[9]

Apparent from these results is that these are hard problems. ImageNet-trained models slightly outperform their Places365-trained counterparts. This suggests that the natural object features extracted from the ImageNet dataset are more salient toward emotion recognition than are place-related features. In 9 out of 13 cases, our UnbalancedCrossEntropyLoss has the edge over regular CrossEntropyLoss. Predicting Arousal appears more difficult than predicting Valence, which aligns with lesser annotator agreement for Arousal than Valence, as analyzed in §A.9. As for the architectures, VGG is a clear winner, with ResNet second. Although twice as large, ResNet101 performs very similar to ResNet50. The larger depth of the DenseNet model does not translate in better performance. A breakdown of model performance per Emo8 class can be found in §A.12, showing overall best performance on "joy" and "anger". Worst performance is registered for "surprise" and "disgust" which, perhaps not surprisingly, are also the emotions for which the least annotations are available.

Interestingly, as explored in §A.14, when a model deviates from the target Emo8 annotation there is a strong tendency toward "nearby" emotions. Most often this is the adjacent leaf, with more distant leaves increasingly more unlikely. This behavior is reminiscent of the kind of disagreements we find among our human annotators (see §A.9).

4 Beyond the Baseline

To build upon the baseline established in Section 3, we built multi-stream models by applying the popular technique of late fusion [24, 25, 26, 27]. Concretely, we combine streams by concatenating their corresponding feature or output vectors, and sending the resulting vector through an extra linear layer. This section reports results for Emo8 classification; the analogous discussion for Arousal and Valence regression can be found in §A.13.

We consider the following streams for combinations: *Emo8 predictions*: for each considered architecture, we trained an Emo8 model, and took the predictions from this model as an 8-feature vector; *Baseline features*: we take the model features from the penultimate layer, vector size depends on the architecture; *EmoNet predictions*: applying the model gives us a 20-feature vector (see Section 3); *YoLo v3 trained on Open Images + Facial Emotion Recognition (OIToFER)*: we apply YoLo v3[10] [43], using LightNet [44], to each image and extract the detected "Human face" regions with probablity $p > 0.005$. We then apply the FER2013-trained ResNet18 model by X. Yuan[11] to the extracted faces, resulting in a 7-feature vector per face. We generate two 7-feature vectors from this, one containing the vector averages, the other the standard deviations, and concatenate both to obtain a final 14-feature vector; *Places365 ResNet18 predictions*: applying the ResNet18 model trained on the Places365 dataset gives us a 365-feature vector per image; *Places365 ResNet18 features*: we take the model activations from the penultimate layer, giving us a 512-feature vector.

The experimental setup is identical to Section 3, except that for time considerations, we only consider CrossEntropyLoss.[12] The test results for Emo8 classification are shown in Figure 6. Training results, as well as numerical training and test results, are included in §A.13. A first observation

[9] I.e., we precomputed the output of the frozen part of the model, and stored it on disk for easy reuse.

[10] We used the Open Images weights available from `https://pjreddie.com/darknet/yolo/`.

[11] `https://github.com/LetheSec/Fer2013-Facial-Emotion-Recognition-Pytorch`

[12] Indeed, our UnbalancedCrossEntropyLoss code is not yet optimized, and slower than CrossEntropyLoss.

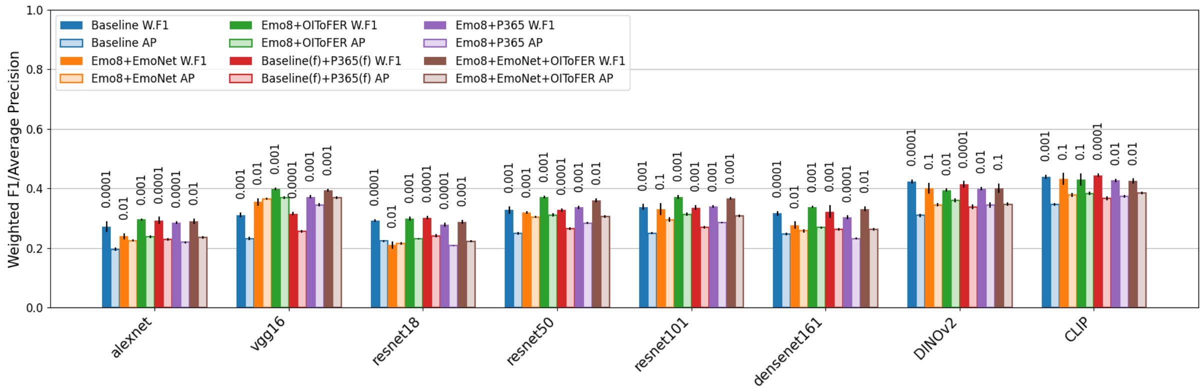

Figure 6: Test data results for extensions beyond the baseline by applying late fusion with Facial Emotion Recognition predictions (OIToFER), EmoNet predictions (EmoNet) and Places365 (P365) predictions or features. For all models, predictions on the dataset (Emo8) are concatenated and sent through a linear layer, except when '(f)' is shown, indicating model features are concatenated. The starting learning rate corresponding to each model is displayed above the training bars.

is that improving upon the baseline appears non-trivial; except for VGG16, the obtained gains are modest. Second, the highest gains clearly come from adding facial emotion features. Third, even though adding EmoNet and OIToFER features separately has a positive effect for VGG16, adding both together does not result in a compounded improvement. Fourth, the added dimensionality of concatening features instead of predictions in the case of Places365 does not result in markedly different results, in some cases even leading to worse results. Finally, not a single stream combination resulted in improved performance for CLIP and DINOv2, with the best VGG16 results nearing CLIP/DINOv2 performance.

5 Discussion

Findings The analysis of our dataset shows the annotations to conform to expectations, with Valence and Arousal values following the expected trends. Furthermore, when annotators disagree on the emotion label, they tend to choose nearby emotions in PWoE nonetheless. Our experiments show that, for the Emo8 prediction task on our dataset, modern ViT models do not seem to really outperform older CNN architectures, with VGG16 even (slightly) outperforming DINOv2 when both baselines are augmented with Facial Emotion features. For Arousal and Valence prediction however, the ViT models are clearly superior.

Limitations 1) While images in our private set have multiple annotations, we have followed the approach of Emotic [24] and gathered only a single annotation per image in our public set. This choice has allowed us to gather a larger data set, but may cause concerns about reliability. These concerns are alleviated by the clear tendency observed on the private set toward similar emotions in case of multiple labels (§A.9), combined with trained models exhibiting this same tendency to strongly favor nearby emotion leaves when deviating from the annotation (§A.14). In short: the models trained using single annotations showed similar statistics to the human multi-label annotations. 2) Concerning potential biases in the images themselves, as they were scraped from the internet the dataset inherits the same biases the internet exhibits. In particular, we have not performed any analysis concerning potential representation issues. As such, there is an unverified possibility that models trained on our dataset wrongly associate "negative" emotions more strongly with certain minority groups. 3) Since legal issues (see §A.2) prevent us from sharing the actual images, we had to resort to sharing URLs. While URLs can break, we mitigate this risk by offering multiple different URLs for the same image where possible.

Impact Statement This paper presents work whose goal is to advance the fields of Machine Learning, Psychology and Psychiatry. Our own interest lies with non-commercial applications with respect to the understanding of Emotion Recognition and Social Cognition in individuals, and how these can be affected by neurological conditions. In particular, we hope that our (future) work will be of help in assisting people with impaired Social Cognition to navigate life.

Nevertheless, the data, and possible future Machine Learning advances inspired by it, could very well lead to commercial (e.g., personalized ads tailored to one's mood) and surveillance (e.g., general crowd monitoring, detection of aggression within crowds, etc.) applications that we strongly feel warrant a public debate with regard to their desirability, and even legality.

Furthermore, the use of web-scraped images entails that not only our dataset risks inheriting biases present on the web, but that our dataset contains images of and by people (i.e., subjects and authors) that would not necessarily agree to their likeliness or work being used for the purposes described in this paper. For this reason, we offer an opt-out option to anyone who wants their likeliness or work removed from out dataset.

Conclusion We present FindingEmo, a dataset of 25k image annotations for Emotion Recognition that goes beyond the traditional focus on faces or single individuals, and is the first to target higher-order social cognition. The dataset creation process has been discussed in detail, and the annotations have been shown to align with expectations. A cross-cultural analysis of the annotations was performed, showing similar tendencies between regions and ethnicities for Valence and Emo8, with Arousal annotations somewhat less aligned. It is however important to note that the limited amount of datapoints does not allow to make strong, definitive statements. Baseline results are presented for Emotion, Arousal and Valence prediction, as well as first steps to go beyond the baseline. These results show the dataset to be complex, and the tasks hard, with even modern models like CLIP and DINOv2 struggling. This suggests that in order to solve these tasks, novel Machine Learning roads might need to be explored. Our annotation interface and code for model training are made open source.

Acknowledgments and Disclosure of Funding

This work was funded by KU Leuven grant IDN/21/010.

We are grateful to dr. Simon Vandevelde for providing us with the dataset name.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1026–1034.

[2] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty, "International evaluation of an AI system for breast cancer screening," *Nature (London)*, vol. 577, no. 7788, pp. 89–94, 2020.

[3] R. W. Picard, *Affective computing*. Cambridge, Mass: MIT Press, 1997.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[5] A. Emanuel and E. Eldar, "Emotions as computations," *Neuroscience & Biobehavioral Reviews*, vol. 144, p. 104977, 2023.

[6] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives," *Frontiers in Robotics and AI*, vol. 7, 2020.

[7] L. F. Barrett, "Discrete Emotions or Dimensions? The Role of Valence Focus and Arousal Focus," *Cognition and Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[8] L. F. Barrett, M. Gendron, and Y.-M. Huang, "Do discrete emotions exist?" *Philosophical Psychology*, vol. 22, no. 4, pp. 427–437, 2009.

[9] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell, "On the importance of both dimensional and discrete models of emotion," *Behavioral Sciences*, vol. 7, no. 4, pp. 66–, 2017.

[10] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, vol. 14, no. 4, pp. 261–292, 1996.

[12] P. Ekman, "Universal facial expressions of emotion," *California Mental Health Research Digest*, vol. 8, no. 4, pp. 151–158, 1970.

[13] R. Plutchik, "A General Psychoevolutionary Theory Of Emotion," 1980.

[14] K. R. Scherer, "What are emotions? And how can they be measured?" *SOCIAL SCIENCE INFORMATION SUR LES SCIENCES SOCIALES*, vol. 44, no. 4, pp. 695–729, 2005.

[15] Y.-I. Tian, T. Kanade, and J. Cohn, "Recognizing action units for facial expression analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[16] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[17] S. Zhang, X. Zhao, and B. Lei, "Robust facial expression recognition via compressive sensing," *Sensors (Basel, Switzerland)*, vol. 12, no. 3, pp. 3747–3761, 2012.

[18] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint Pose and Expression Modeling for Facial Expression Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 3359–3368.

[19] J. Zhu, B. Luo, S. Zhao, S. Ying, X. Zhao, and X. Zhao, "IExpressNet: Facial Expression Recognition with Incremental Classes," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2899–2908.

[20] S. Zhang, Z. Huang, D. P. Paudel, and L. Van Gool, "Facial emotion recognition with noisy multi-task annotations." IEEE, 2021.

[21] Z. Y. Huang, C. C. Chiang, J. H. Chen, Y. C. Chen, H. L. Chung, Y. P. Cai, and H. C. Hsu, "A study on computer vision for facial emotion recognition," *Scientific reports*, vol. 13, no. 1, pp. 8425–8425, 2023.

[22] H. Aviezer, Y. Trope, and A. Todorov, "Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[23] F. Kumfor, A. Ibañez, R. Hutchings, J. L. Hazelton, J. R. Hodges, and O. Piguet, "Beyond the face: how context modulates emotion processing in frontotemporal dementia subtypes," *Brain*, vol. 141, no. 4, pp. 1172–1185, 01 2018.

[24] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context Based Emotion Recognition using EMOTIC Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, arXiv:2003.13401 [cs].

[25] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 14 222–14 231.

[26] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "EmoSeC: Emotion recognition from scene context," *Neurocomputing*, vol. 492, pp. 174–187, 2022.

[27] D. Yang, S. Huang, S. Wang, Y. Liu, P. Zhai, L. Su, M. Li, and L. Zhang, "Emotion recognition for multiple context awareness," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 144–162.

[28] L. P. Mertens, "Modeling Social Cognition and its Neurologic Deficits with Artificial Neural Networks," in *Proceedings of the 25th International Conference on Multimodal Interaction*, ser. ICMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 726–730.

[29] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information Fusion*, vol. 102, p. 102019, 2024.

[30] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*. IEEE, 1998, pp. 200–205.

[31] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. IEEE, 2010, pp. 94–101.

[32] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-Aware Emotion Recognition Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10 142–10 151.

[33] I. J. Goodfellow, D. Erhan, P. Luc Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," *Neural networks*, vol. 64, pp. 59–63, 2015.

[34] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE transactions on affective computing*, vol. 10, no. 1, pp. 18–31, 2019.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017.

[36] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[39] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million Image Database for Scene Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[40] P. A. Kragel, M. C. Reddan, K. S. LaBar, and T. D. Wager, "Emotion schemas are embedded in the human visual system," *Science Advances*, vol. 5, no. 7, p. eaaw4358, 2019.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.

[42] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision Transformers Need Registers," 2023.

[43] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *ArXiv*, vol. abs/1804.02767, 2018.

[44] T. Ophoff, "Lightnet: Building Blocks to Recreate Darknet Networks in PyTorch," https://gitlab.com/EAVISE/lightnet, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.

(c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See footnote 1.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The graphs show error bars and the tables contain standard deviations.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3, penultimate paragraph.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] We referenced all relevant papers for all model architectures we used.

   (b) Did you mention the license of the assets? [No] The assets we use are commonly used, publicly released, pretrained models. As is common practice, we did not explictly mention the licenses, but do cite the corresponding papers or refer to the repository URL.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] All released new assets are included in the repository for this project, which is clearly mentioned in Section 1.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Quoting from §2.2: "Annotators were recruited through the Prolific platform, and first required to agree to an Informed Consent clause[...]"

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Quoting from §A.8: "Prolific provided us with anonymized personal data, except for 1 user. Not all datapoints are available for all users." Beyond us only having anonymized personal data, we do not share this data and only include aggregated results (e.g., distribution over age and country of origin) in our paper.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See §A.6.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] There were no participant risks identified. As mentioned in Section 1, the process to collect the annotations was approved by our insitution's Ethics Committee.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See §2.2, where we explicitly mention a £10 per participant compensation, for an expected total task duration of 1h.

A Appendix

A.1 Dataset Logo

The logo of the dataset is depicted in Figure 7.

Figure 7: Logo for the FindingEmo dataset.

A.2 Legal Compliance

Concerning the legal status of the dataset, two question arise: 1) are we allowed to share URLs to (potentially) copyrighted content, and 2) are we allowed to use (potentially) copyrighted material to train our models?

With regard to 1, we verified this with copyright experts at our institute who assured us that this is legal. With regard to 2, we point to Title II, Article 3, "Text and data mining for the purposes of scientific research", of the so-called InfoSoc Directive [13], which provides an exception to copyright obligations for (members of) research organisations. As members of KU Leuven, we fall under this law. If you are not a member of a European research or cultural heritage institution, you will need to check with your local regulation whether or not you have the right to use this material for research purposes.

We are the rightful owners of the annotations, so no potential copyright issues arise for this data. We expressly distribute the dataset under a *non-commercial* CC BY-NC-SA 4.0 license.

A.3 Additional Annotation Dimensions

These are the remaining annotation dimensions that were not mentioned in the main text for brevity.

Age group Users had to tick one or more boxes from "Children", "Youth", "Young Adults", "Adults" and "Seniors", indicating the age groups present in the image.

Deciding factor(s) for emotion Users had to tick one or more boxes from "Neutral", "Body language", "Conflict context vs. person", "Facial expression" and "Staging", indicating what prompted them to choose for a particular emotion.

Ambiguity Lastly, users could indicate by means of an integer scale $[0, 1, \ldots, 6]$ how ambiguous the emotional content exhibited by the entire photograph was, or alternatively, how much difficulty they had in annotating the picture.

A.4 Details of the Dataset Creation Process

This section describes in more detail the two phases in the dataset creation process introduced in §2.2.

[13] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0790`

A.4.1 Phase 1: Gathering Images

Phase 1 consisted in building a customized, Python-based DuckDuckGo[14] image scraper, programmed to generate random image search queries as follows. Three sets of keywords were defined: one containing a diverse set of emotions; one referring to groups of people (e.g., 'people', 'adults', 'youngsters', etc.); and one containing social settings and environments (e.g., 'birthday', 'workplace', etc.).[15] By taking all possible combinations of the elements in these sets, the system generated a multitude of queries, such as, e.g., "happy youngsters birthday". The first N results were then retrieved and filtered to exclude a number of manually blacklisted domains (e.g., stock photography providers) and by image size. Query results that passed the filtering steps were downloaded.

We started with $N = 500$ and image width 800px $< w <$ 1600px, and later extended this to $N = 1000$ and 800px $< w < 3200$. Obviously, not all downloaded images satisfied our criterion of depicting multiple people in a natural setting. Hence, as a further filtering step, one of the authors annotated 3097 images as either "keep" (useful) or "reject" (no use). These images were used in a random 80/20 split to train a CNN to perform the same task, achieving an accuracy of 77.6%. This model was used to further filter downloaded images, in particular to identify spurious images such as, e.g., drawings, images with lots of text, etc.: if the CNN labeled the downloaded image as "reject", the image was discarded. If the downloaded image was labeled as "keep", it entered the pool of images that could be selected for annotation.

In total 1,041,105 images were collected.

A.4.2 Phase 2: Gathering Annotations

The annotations were gathered using a custom web interface written in Python, HTML and JavaScript. Annotators were recruited through the Prolific platform. For this, a job would be created, which we refer to as a "run", to which users could subscribe. After doing so, they received a URL that allowed them to log on to our system and, after agreeing to an Informed Consent clause, perform the annotations. First, users were presented with detailed instructions, a copy of which are provided in §A.6, after which the data collection proper began. To be able to monitor the process closely, and to cope with hardware limitations of our server, we opted to only perform runs with a limited number of participants, most often 10 or 15. For each run, the Prolific user selection criteria were the same: fluent English speaker, (self-reported) neurotypical, and a 50/50 split male/female.

In total, annotations were collected over 51 runs. Candidates were informed of an expected task duration of 1h, including reading the instructions, and offered a £10 reward. Analysis of the durations (see §A.7) show our time estimation to be fair. We spent a total of £10k, which includes annotators whose contributions were filtered out, and most importantly, Prolific fees and taxes.

A screenshot of the interface is included in §A.5. The interface presents users with images on the left side, and dimensions to annotate on the right side. At the top left, users are presented with two buttons: one to skip an image if they so wish, and one to save the current annotation and move on to the next image.

Upon being presented an image, the first choice users needed to make was, just like the filtering CNN, whether to "keep" or "reject" the image, according to the provided instructions. Essentially, users were asked to reject images that contained no people, were watermarked, were of bad quality, etc. If users opted to "reject" an image, no further annotation was needed. This step was needed to further filter images that passed through the CNN. If the choice was "reject", no further action (besides saving) was required. Optionally, users could choose to select one of several tags indicating why they opted to reject the image from "Bad quality photo", "Copyright", "Watermark", "No interaction", "No people", "Text" and "Not applicable". Each user was asked to annotate 50 "keep" images; "rejects" did not count towards the total goal. Despite this, some users still performed full annotations on images they rejected. If users opted to "keep" the image, they were expected to annotate all other dimensions as well.

Although the frontend (i.e., user interface) remained essentially unchanged, the backend underwent some changes as annotations were collected, and some lessons were learned, which we discuss here.

[14] `https://www.duckduckgo.com`

[15] The full list of keywords is available from our code repository.

Initial iteration Initially, an image was randomly selected from the corpus, and processed by an updated "keep/reject" CNN (see §A.4.1) with an accuracy of 83.6%. If the "keep" probability p_k was < 0.75, a new random image would be selected and tested, until one was found with $p_k \geq 0.75$. If this image had already been annotated, the process would start over, until a valid image was found, which would then be shown to the annotator.

Second iteration At first, the annotating of all dimensions was not enforced; users could select the "keep" checkbox, save the annotation without annotating anything else, and move on to the next image. Most did their job diligently, but nevertheless we opted to update the interface to require all dimensions be annotated in case of a "keep", before the "Save" option became available. This frequently prompted messages from users complaining the "Save" option was not available to them. A further update explained this to users who prematurely clicked on the "Save" button.

Third iteration Over the course of the first few thousand annotations, it became clear that two emotion leaves were particularly overrepresented, namely "joy" and "anticipation", respectively accounting for 35.9% and 23.0% of all annotations by the time of Run 9. In an attempt to counter this, we came up with the following system.

Besides the "keep/reject" CNN, we trained a second CNN to predict the Emo8 label. We then first computed all "keep/reject" predictions for all images in the corpus, and followed this up by predicting Emo8 labels for all "keep"-labeled images. Upon starting the annotation server, these predictions are loaded into memory. When selecting an image to show to a user, first an emotion label is chosen, with odds inversely proportional to the number of images that were tagged (by the CNN) with a certain label. Second, out of all images tagged with this label, one that had not previously been annotated by an annotator would be chosen. The CNN used to make the predictions was retrained at several steps along the annotation gathering process. Using this system, we managed to decrease "joy" down to 28.4%, and up "sadness" from 6.3% to 10.5%.

A.5 Annotation Interface

A screenshot of the annotation interface is shown in Figure 8.

A.6 Copy of the Annotator Instructions

Welcome

It is recommended to set your browser to "full-screen" mode. Typically, this mode can be toggled by using the 'F11' key.

This interface was designed for screen resolutions with a width of 1920 pixels. In case your screen has a higher/lower resolution, the interface should automatically resize itself so as to fully fit on your screen, but this might come at the price of reduced image sharpness.

Thank you for your willingness to participate in this annotation task!

In this experiment, you will be expected to annotate 50 "good" images, i.e., annotated as "Keep", after which you will receive a URL that will direct you to the Prolific completion page for this task. Please take the time to read these annotation instructions before continuing.

Note that if for any reason you get logged out at some point, you should be able to log back in using the same URL provided to you by Prolific, and pick up right where you left.

We want to build a database of photographs with an emotional content. You will be shown randomly selected images from a large corpus, and we ask you to evaluate photographs regarding 2 consecutive issues.

First, regardless of the emotional content, all photographs should adhere to the following criteria:

- Each photograph must display a realistic situation, e.g., no drawings, no watermark, no fantasy content (i.e., digitally manipulated photos), no horror, etc.
- The formal quality of the photograph should be sufficient, i.e., no fuzzy/blurry photographs.

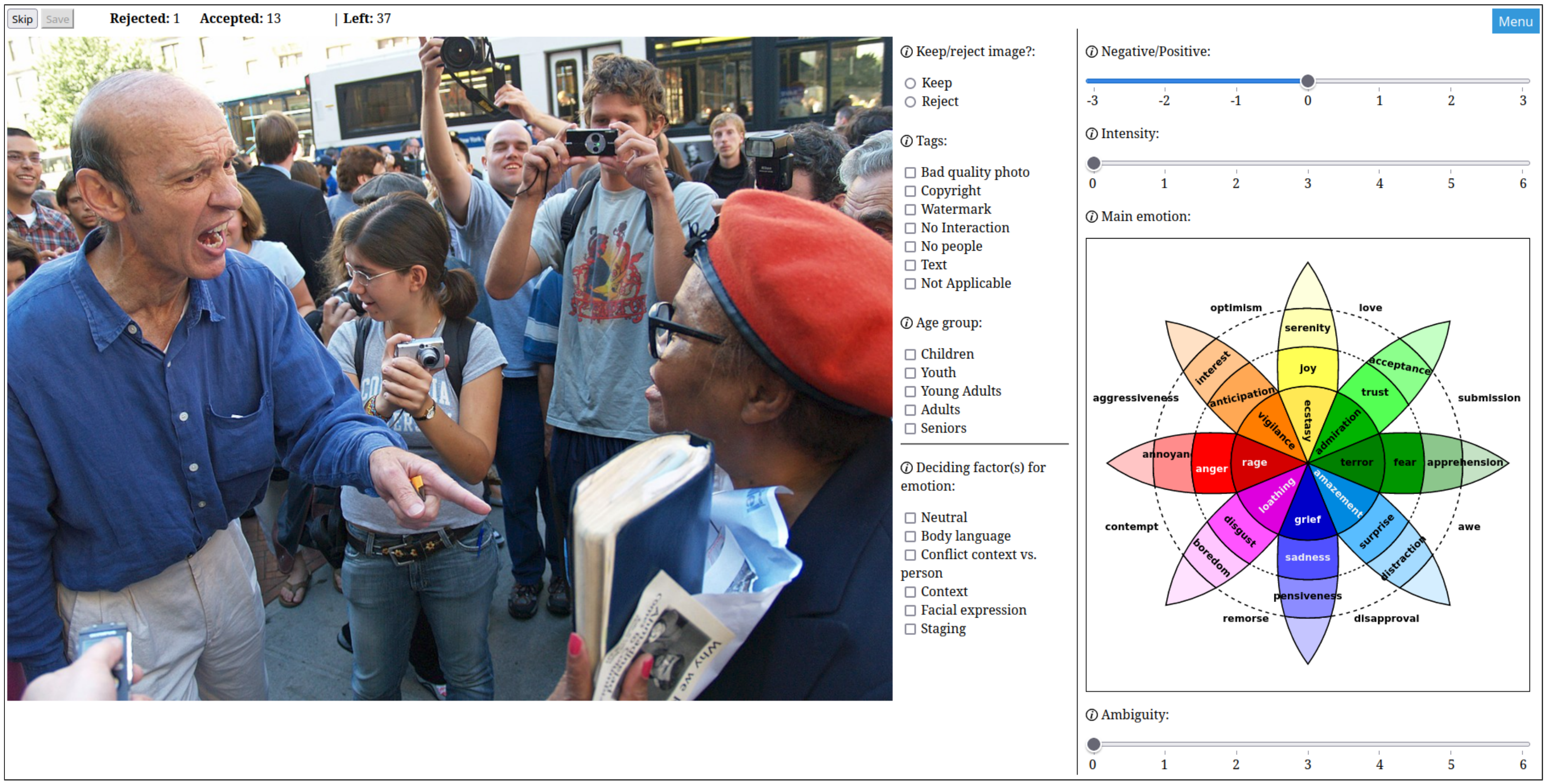

Figure 8: A screenshot of the annotation interface. Displayed photo by David Shankbone, source: WikiMedia.

- Each picture must display at least 2 people that are clearly visible. Alternatively, if only one person is shown, but this person is clearly a part of a larger context, the image can also be suitable.
- The main feature of the photograph must not consist of a textual element. For instance, if a cardboard displaying ‘stop racism’ is a central feature of the picture, the picture is not suitable.

If an image does not adhere to each of these criteria, or you are not certain, please rate it as not suitable by choosing the “Reject” option. Else, mark it as “Keep”, in which case all other dimensions, except for “tags”, need to be annotated before you can proceed! Even if you want to keep the default value of a slider, you still need to click the slider first.

Images can further be described by a number of tags:

- Bad quality photo: when a picture is too blocky/blurry.
- Copyright: a copyright, contrary to a watermark, is not repeated but appears only once. Typically, this leads to the picture being rejected, unless possibly the copyright is only small in size and could be cropped out without losing the essence of the picture.
- Watermark: a watermark is a specific pattern, typically containing the name of the copyright holder, that is repeated over an entire image.
- No interaction: the people in the picture don’t have a direct interaction.
- No people: the picture does not depict any people.
- Text: the image contains a lot of text, either typeset on top of it, or present on, e.g., banners held by subjects depicted in the picture. If the text is typeset, this is disqualifying (i.e., the picture is rejected). If the text is present in the picture itself, it is disqualifying if it is too prominent. Use your own discretion to determine what is “too prominent” and what is not. A good rule of thumb is: if your attention is immediately drawn to the textual elements when viewing the picture, then it is too prominent and the picture is disqualified.
- Not Applicable: typically used for images that are actually a collage of more than one photo, or that are rejected but don’t fit any of the other tags.

If a photograph is not rated as suitable (i.e., “Reject”), no further assessment is required; click “Save” to proceed to the next paragraph. Else, for “Keep” or “Uncertain” photos, you are also expected to annotate the age group of the main participants in the picture. These labels are of course not clear cut; feel free to use your own discretion as to which label applies best.

Second, we want you to focus on the emotional labelling of the photographs. Concretely, we ask you to annotate the image on a number of dimensions

We ask you to indicate the emotional characteristic of the ENTIRE SCENE displayed in the photograph, independent of your own political/religious/sexual orientation. So a black lives matter protest is typically negative (= the participants are not happy) independent of whether you support BLM. Specifically, we ask you to rate the valence (“Negative/Positive”) of the overall emotional gist of the photograph on a 7-point Likert scale from negative (-3) over neutral (0) to positive (+3), and also the intensity, ranging from not intense at all (0) to very intense (6) by using the appropriate sliders.

We also ask to indicate an emotional label by means of a mouse click on an emotion wheel called “Plutchik’s Wheel of Emotions”. If you can’t find the perfect emotional label then you choose the ‘next best thing’, i.e., the one that reflects it most. In case no particular emotion fits, i.e., the participants all display a neutral expression, you can opt to select no emotion, although such cases are expected to be rare. For a more detailed description of each emotion depicted in this wheel, see, e.g., `https://www.6seconds.org/2020/08/11/plutchik-wheel-emotions/`. Additional info for each emotion will be displayed when hovering over its corresponding cell.

Please also rate how straightforward the emotional content that is exhibited by the entire photograph is using the scale indicated with “Ambiguity”. For instance, if there are approximately as much emotionally positive as emotionally negative cues in the photograph, the emotional content would not be clear (6), while only positive cues or only negative cues would result in a very high clarity (0).

Finally, the options under the “Deciding factor(s) for emotion” header ask which aspects of the photo influenced you most when assessing the emotion, i.e., facial expressions, bodily expressions, the type of interaction ('Staging') among the persons (e.g., fighting, dancing, talking), type of context (e.g., wedding, funeral, protest, etc.), objects in the photograph (e.g., gun, chocolate) or a possible conflict between context and person(s) (i.e., somebody exuberantly laughing at a funeral). If none of these apply, and/or the emotion is rather neutral, the "Neutral" tag can be used, although just as for the emotion case, we expect these occasions to be rare.

If for some reason you would rather not annotate the current image being served to you, you can press the "Skip" button to be served a new picture and have the annotation interface be reset, without your current settings being saved.

If on the other hand you are happy with your current annotation, press "Save" to let it be saved and move on to the next image. If this button is greyed out, this means you have not yet annotated all necessary dimensions. Once you have reached the required number of annotations, you will automatically get to see the URL that will direct you to the Prolific completion page for this task.

At the top of this screen, you can see your annotation statistics: "Rejected/Accepted" = how many images you marked "Reject" and "Keep" respectively, and "Left" = number of "Keep" images left to annotate.

You can always check these instructions again whilst annotating by clicking the -icon next to each criterium. (Click once more to close the infobox again.)

A.7 Task Duration Analysis

A histogram of time taken per annotator to complete the task is shown in Figure 9. These are the durations as reported by Prolific. An important remark to make is that for Prolific users, the clock starts ticking once they subscribe to a job. By default, per the Prolific rules, for a job expected to take 1h users are allowed a maximum of 140 minutes to complete the job. It appears that many users subscribe to a job, and then leave their browser tab open for a while before starting the job proper. (Some never start, leading to a time-out.) Taking this into account, the shown distribution is a "pessimistic" picture, including many idled minutes. The average time taken per user, including users that were ultimately filtered out of the dataset, was 64 ± 27 minutes. With all of the above in mind, we conclude our alloted time was fair.

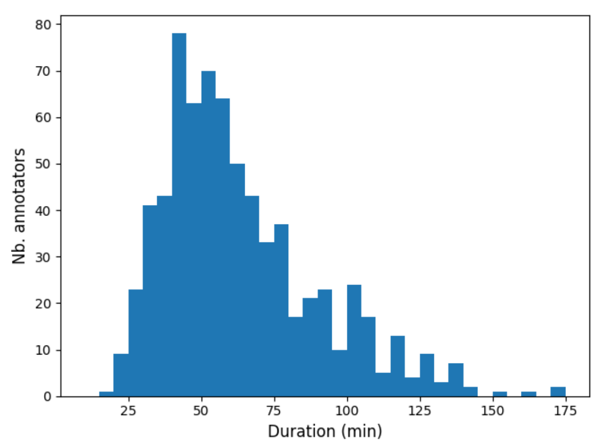

Figure 9: Distribution of minutes taken to complete the task. The plot does not include 7 outliers.

A small negative correlation manifests between the task completion time and the annotator score (SpearmanR$= -0.122$, $p = 0.002$ for s, SpearmanR$= -0.086$, $p = 0.029$ for s_{alt}).

A.8 Annotator Statistics

Annotations were collected from 655 annotators. Prolific provided us with anonymized personal data, except for 1 user. Not all datapoints are available for all users.

Of the annotators, 337 are male, 317 are female, and 1 unknown. 651 annotators were spread over 49 countries, with country for the remaining 4 unknown. Most popular were South Africa (176 annotators), Poland (127 annotators) and Portugal (104 annotators). From there, numbers drop rapidly, with follow-up Greece accounting for only 32 annotators. The full distribution of annotators per country is shown in Figure 10. The age distribution of the 653 users who shared that info is shown in Figure 11, indicating a large bias towards the early 20's. 654 annotators shared their ethnicity, consisting in 424 users identifying as White, 166 identifying as Black, 34 identifying as Mixed, 18 identifying as Other and 12 identifying as Asian.

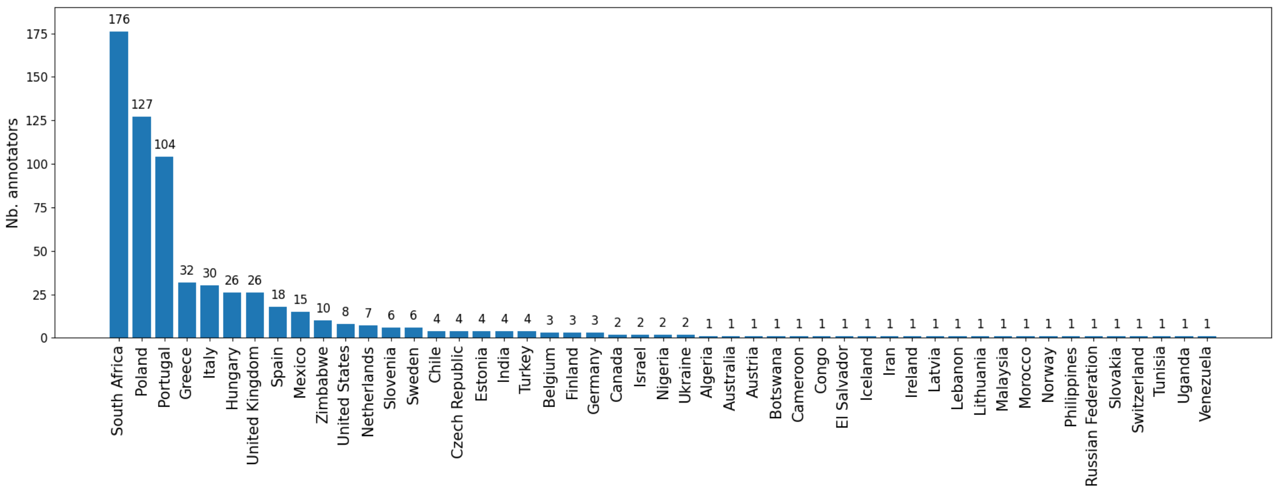

Figure 10: Distribution of country of origin of 651 annotators.

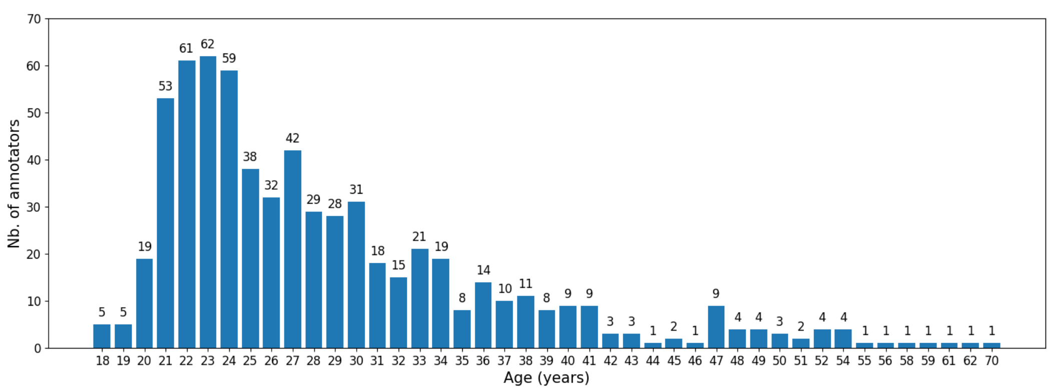

Figure 11: Distribution of age of 653 annotators.

A.9 Inter-annotator Agreement

Recall from Section 2 that we hold a private set of 1,525 images that have each been annotated by multiple users[16], amounting to a combined 6115 annotations. Table 3 shows how many images have been annotated by N different annotators. Of these, 1294 images have a majority of "Keep" annotations, 137 are mainly "Reject" and 94 are undecided.

[16]This set does not include the fixed overlap images.

We do not report the often used Cohen's Kappa and/or Krippendorf's Alpha scores, as these metrics are only meaningful when most pairs of annotators have both annotated a substantial set of shared images. In our case, however, by design, the number of images that have been annotated by any two annotators is low (1 or 2 at most, and very often zero). As such, we feel these metrics are not applicable. We explicitly opted to have a large number of annotators annotate a small number of images each, in order to have the annotations better be a reflection of "the population at large", rather than of a few annotators.

Table 3: Number of images ("# imgs.") that have been annotated by N different annotators ("# ants.").

# ants.	2	3	4	5	6	7	8
# imgs.	245	328	283	524	127	17	1

Focusing on the 1294 "Keep" images, Figure 12 shows how many images have been annotated with N different emotion labels, both for Emo8 and Emo24 labels. For 26.2% of images, all annotators chose the same emotion leaf, and 46.6% were annotated with 2 different Emo8 labels. For the finegrained Emo24 labels, 80.1% of images have been annotated with a maximum of 3 different labels.

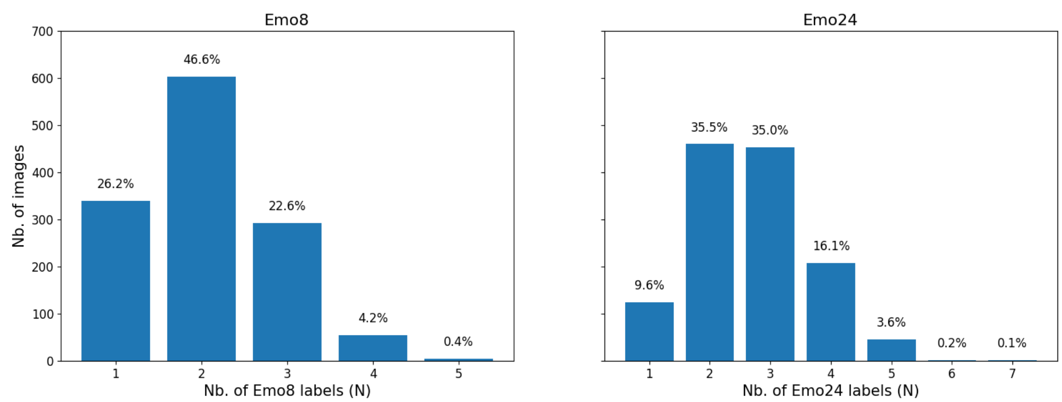

Figure 12: Number of images with N different Emo8 and Emo24 labels. The y-axis is shared between both plots.

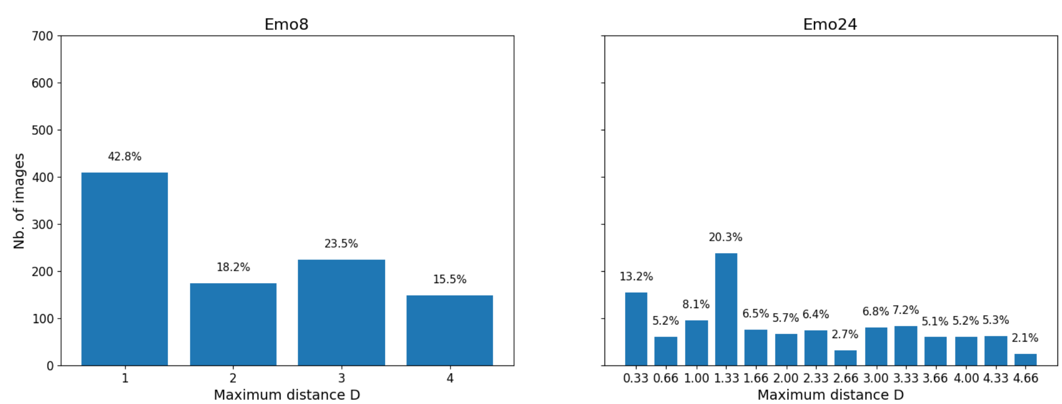

Figure 13: Number of images with a maximum distance D between their Emo8 and Emo24 labels, for images annotated with more than one label. The y-axis is shared between both plots.

Turning to the question of how different the separate emotion labels for a same image are, Figure 13 shows the distribution of maximum distance between labels, for images annotated with more than one label. The distances for 24 emotions are computed by also giving an ordinal to each emotion within a leaf, as shown in Figure 14. No less than 42.8% of the times an image has been annotated with more than one Emo8 label, those labels represent adjacent emotion leaves, while in 15.5% of the cases they represented opposite leaves, most often the pairs ("anger", "fear") and ("anticipation", "surprise").

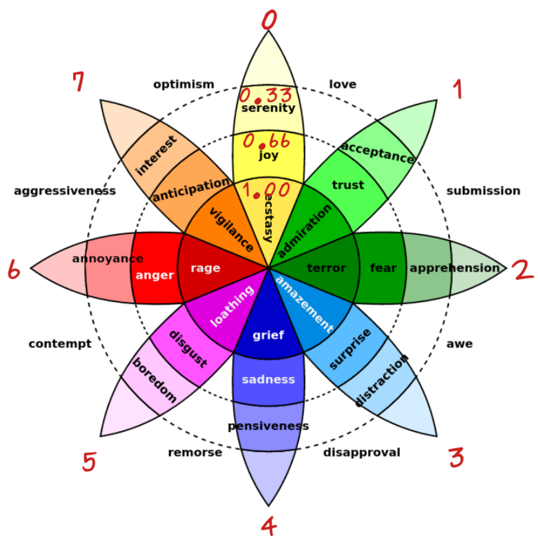

Figure 14: Plutchik's Wheel of Emotions: ordinals of emotions. The outer numbers represent the ordinal of the leaf, the numbers within the upper central leaf the ordinals of the emotions within a leaf. E.g., "joy" = 0.66 and "boredom" = 5.33. The distance between them then becomes 3.33, being the sum of the distance between the leaves (3) and the "intra-leaf" distance (0.33).

To get a better idea of what Emo8 labels often appear together, we focused on images with 2 Emo8 labels, and plotted how often each emotion pair occurs. The result is shown in Figure 15, demonstrating the pairs ("joy", "anticipation"), ("joy", "trust) and ("anticipation", "trust") make up the bulk of the pairs. As for opposite emotions, the pairs ("anticipation", "surprise") and ("anger", "fear") appear markedly more ofthen than ("joy", "sadness") and ("disgust", "trust").

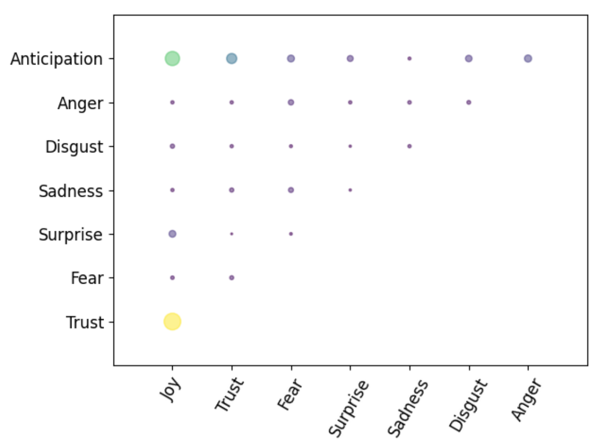

Figure 15: Prevalence of Emo8 label pairs for images annotated with 2 labels. The bigger the disc, the more often the pair appears in the dataset.

To analyze the Arousal and Valence values, we compute the maximum distance between annotated values for both dimensions over all “keep” images. For Arousal, the average maximum distance is 2.7 $\pm$ 1.4, while for Valence this is 1.8 $\pm$ 1.2. This suggests that people agree much more on the Valence dimension, than they do on the Arousal dimension. This is confirmed when we compute the average maximum distance values as a function of the number of annotations for a given image, the result of which is shown in Figure 16. For Arousal, a clear increasing maximum distance trend is visible with a stable standard deviation, going from $\pm$1.75 to more than 4. For Valence annotations on the other hand, the maximum distance appears to plateau at close to 2.

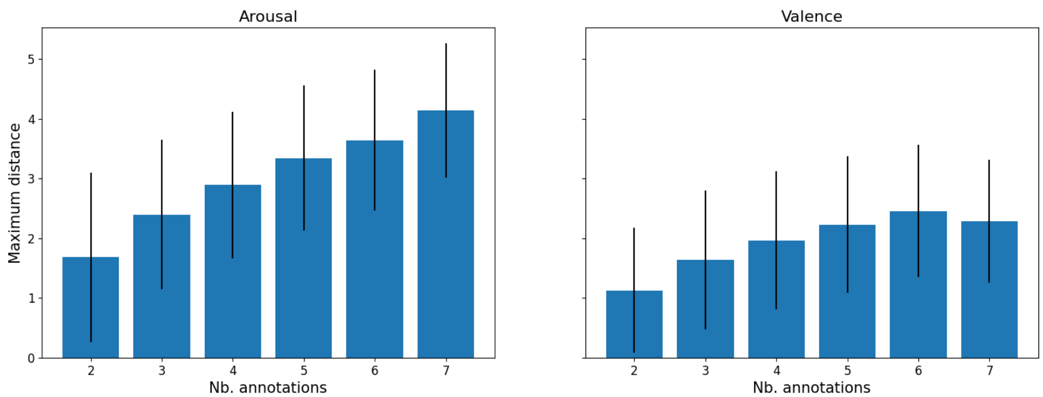

Figure 16: Distribution of maximum distance between Arousal and Valence annotations as a function of the number of annotations per image. The y-axis is shared between both plots.

Histograms comparing the distribution of Arousal, Valence and Emo8 annotations between pairs of geographic regions with at least 100 annotated images in common are included in Figure 17, with statistical data comparing both distributions per pair grouped in Table 4. Analogously for pairs of annotator ethnicities, histograms and statistics can be found in Figure 18 and Table 5 respectively.

A.10 Extra Dataset Analysis

Barplots showing the distribution of Arousal, Valence and Ambiguity annotation values for the public dataset are depicted in Figure 19. Barplots showing the distribution for the public dataset of Arousal and Valence annotations per Emo8 emotion are grouped in Figure 20, showing a clear tendency toward normal distributions. Scatterplots depicting the association between Arousal and Valence annotations per pair of opposite Emo8 emotions (e.g., “joy” and “sadness”) are collected in Figure 21 Annotation statistics per Emo24 emotion are collected in Table 6. The table is made up of three rows, each row corresponding to a ring in Plutchik’s Wheel of Emotions, from the top row corresponding to the outer (least intense) ring, to the bottom row corresponding to the inner (most intense) ring. The annotations follow this ordering, with average Arousal annotations consistently increasing from least to most intense emotion ring. Valence annotations follow suit, either increasing for positive emotions, or decreasing for negative emotions. The sole exception to this rule is center ring “Disgust” having a slightly lower average Valence rating (-1.62) than the inner ring “Loathing” (-1.57).

A.11 UnbalancedCrossEntropyLoss and WeightedMSELoss

Table 7 compares baseline results obtained using CrossEntropyLoss vs. UnbalancedCrossEntropyLoss for Emo8 classification, and MSELoss vs. WeightedMSELoss for Arousal/Valence regression. In what follows, we detail the workings of UnbalancedCrossEntropyLoss and WeightedMSELoss. We observe that UnbalancedCrossEntropyLoss presents a clear benefit over CrossEntropyLoss for the classification problem under consideration, while WeightedMSELoss typically does not manage to positively influence model performance for regression problems.

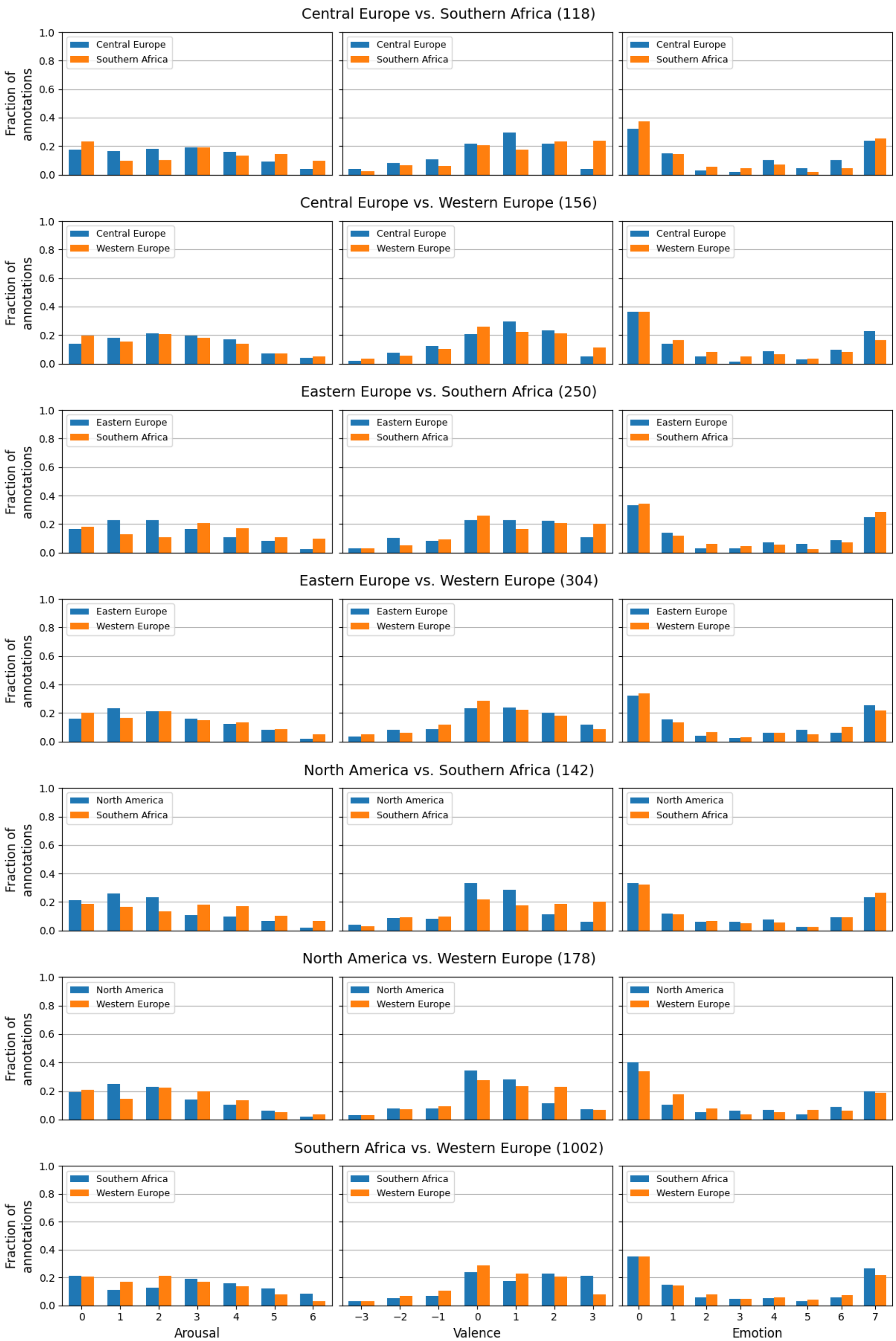

Figure 17: Distribution of annotations; a comparison between geographic regions. The y-axes are shared between all plots, the x-axes are shared between all plots within the same column. Between brackets the number of images annotated by members of both geographic regions. Emotion indices follow Figure 14.

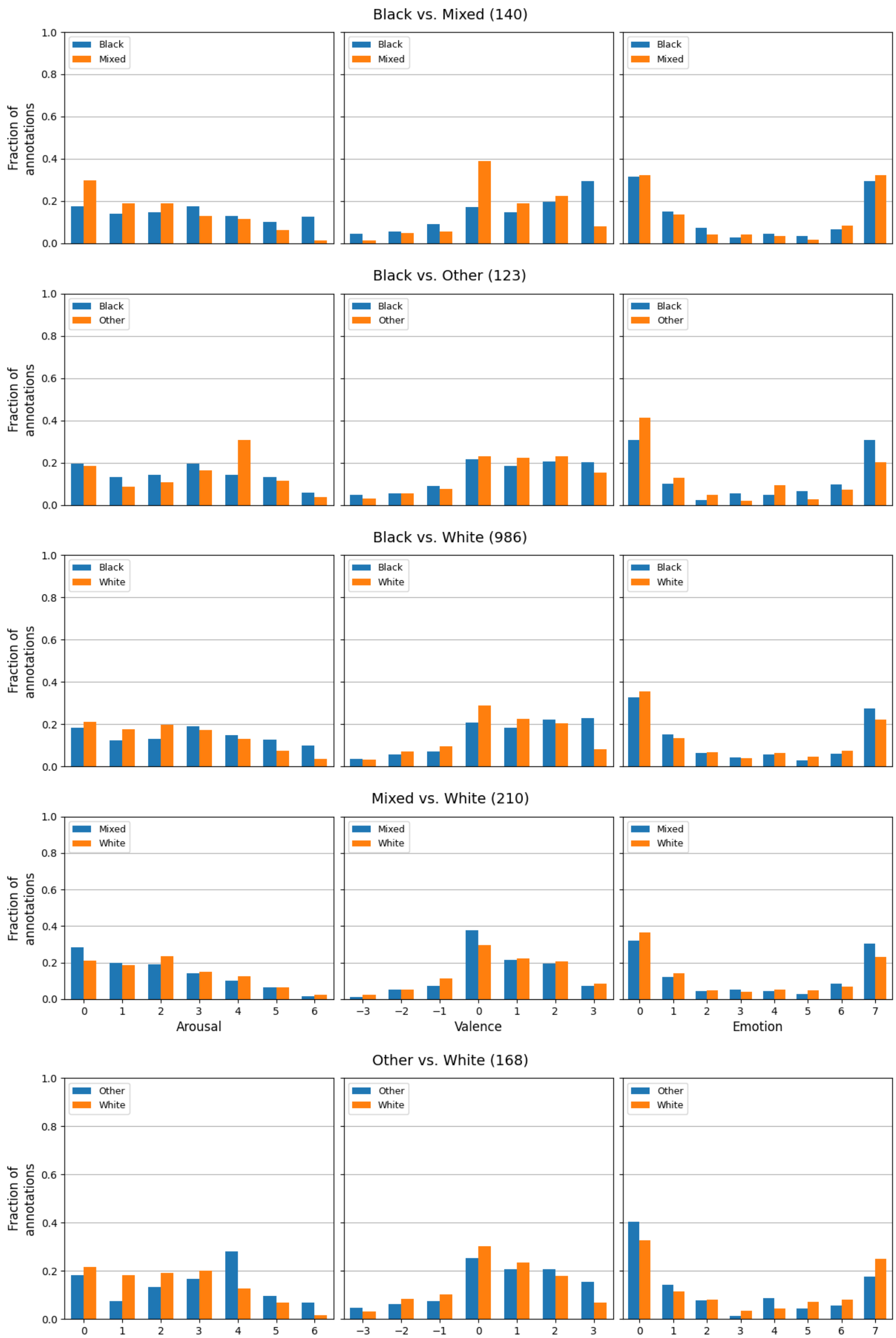

Figure 18: Distribution of annotations; a comparison between ethnicities. The y-axes are shared between all plots, the x-axes are shared between all plots within the same column. Between brackets the number of images annotated by members of both ethnicities. Emotion indices follow Figure 14.

Table 4: Comparison of annotations between geographic regions. Regions: "C/E/W.Eur." = Central/Eastern/Western Europa, "S.Afr." = Southern Africa, "N.Am." = North America. Metrics: "JS" = Jensen-Shannon distance, "KS" = two-sample Kolmogorov-Smirnov test, "Kα" = Krippendorff's Alpha, "S.R" = Spearman's R. Format: $.x^y$ with $0.x$ the metric value and y the p-value, if applicable. JS and KS computed between the histograms depicted in 17, Kα and S.R computed using vectors containing the average annotations per region, per overlapping image. Header: numbers between brackets = number of common annotated images.

	C.Eur. - S.Afr. (118)	C.Eur. - W.Eur. (156)	E.Eur. - S.Afr. (250)	E.Eur. - W.Eur. (304)	N.Am. - S.Afr. (142)	N.Am. - W.Eur. (178)	S.Afr. - W.Eur. (1002)
				Arousal			
JS	.145	.068	.188	.088	.172	.107	.132
KS	$.429^{0.575}$	$.143^{1.000}$	$.286^{0.963}$	$.143^{1.000}$	$.429^{0.575}$	$.286^{0.963}$	$.286^{0.963}$
Kα	.029	.009	$-.010$	$-.029$	.050	$-.002$	$-.012$
S.R	$.262^{0.004}$	$.102^{0.203}$	$.217^{0.001}$	$.273^{0.000}$	$.116^{0.168}$	$.170^{0.023}$	$.225^{0.000}$
				Valence			
JS	.229	.110	.121	.080	.197	.114	.150
KS	$.143^{1.000}$	$.143^{1.000}$	$.286^{0.963}$	$.143^{1.000}$	$.429^{0.575}$	$.286^{0.963}$	$.143^{1.000}$
Kα	.443	.246	.281	.234	.492	.259	.269
S.R	$.482^{0.000}$	$.584^{0.000}$	$.536^{0.000}$	$.617^{0.000}$	$.432^{0.000}$	$.575^{0.000}$	$.546^{0.000}$
				Emo8			
JS	.128	.107	.100	.083	.040	.113	.057
KS	$.250^{0.980}$	$.250^{0.980}$	$.250^{0.980}$	$.125^{1.000}$	$.250^{0.980}$	$.125^{1.000}$	$.250^{0.980}$
Kα	.264	.351	.310	.274	.323	.351	.264
S.R	$.369^{0.000}$	$.237^{0.003}$	$.283^{0.000}$	$.286^{0.000}$	$.268^{0.001}$	$.281^{0.000}$	$.340^{0.000}$

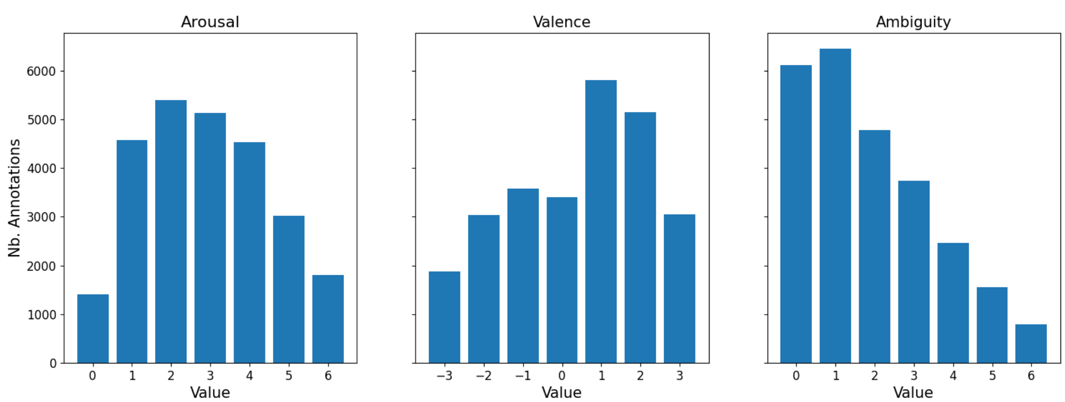

Figure 19: Distribution of Arousal, Valence and Ambiguity annotations. The y-axis is shared between all plots.

Table 5: Comparison of annotations between ethnicities. Metrics: "JS" = Jensen-Shannon distance, "KS" = two-sample Kolmogorov-Smirnov test, "Kα" = Krippendorff's Alpha, "S.R" = Spearman's R. Format: $.x^y$ with $0.x$ the metric value and y the p-value, if applicable. JS and KS computed between the histograms depicted in 18, Kα and S.R computed using vectors containing the average annotations per ethnicity, per overlapping image. Header: numbers between brackets = number of common annotated images.

	Black - Mixed (140)	Black - Other (123)	Black - White (986)	Mixed - White (210)	Other - White (168)
			Arousal		
JS	.205	.147	.133	.071	.204
KS	$.429^{0.575}$	$.429^{0.575}$	$.286^{0.963}$	$.143^{1.000}$	$.429^{0.575}$
Kα	.153	.042	$-.032$	$-.007$	$-.012$
S.R	$-.061^{0.476}$	$.112^{0.220}$	$.256^{0.000}$	$.219^{0.001}$	$.267^{0.000}$
			Valence		
JS	.249	.064	.157	.074	.116
KS	$.286^{0.963}$	$.429^{0.575}$	$.143^{1.000}$	$.286^{0.963}$	$.143^{1.000}$
Kα	.340	.414	.257	.172	.227
S.R	$.479^{0.000}$	$.433^{0.000}$	$.555^{0.000}$	$.556^{0.000}$	$.590^{0.000}$
			Emo8		
JS	.074	.157	.060	.082	.120
KS	$.250^{0.980}$	$.250^{0.980}$	$.375^{0.660}$	$.250^{0.980}$	$.125^{1.000}$
Kα	.253	.285	.242	.281	.310
S.R	$.341^{0.000}$	$.346^{0.000}$	$.337^{0.000}$	$.255^{0.000}$	$.279^{0.000}$

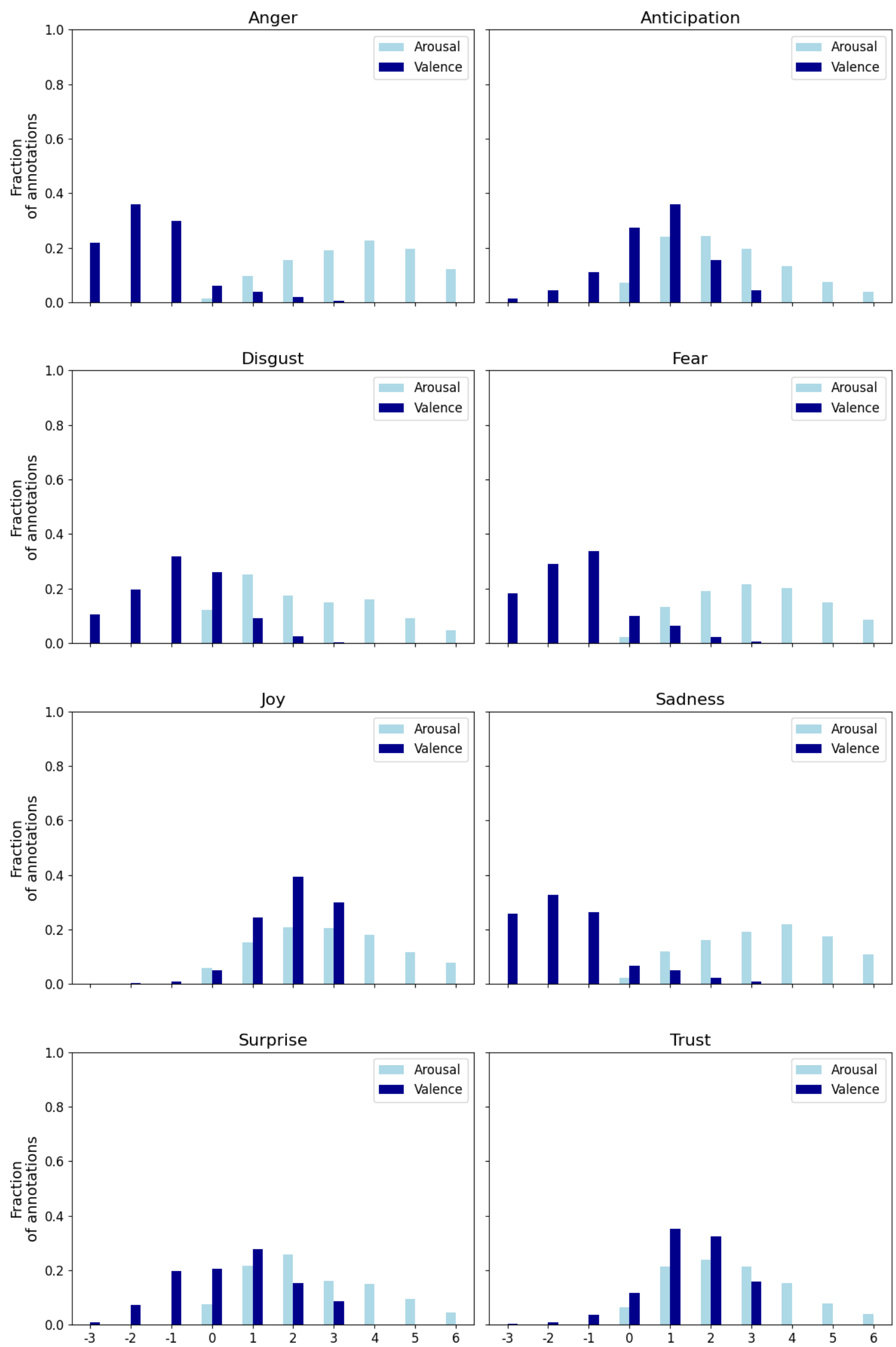

Figure 20: Distribution of Arousal and Valence annotations per Emo8 emotion. The x- and y-axes are shared between all plots.

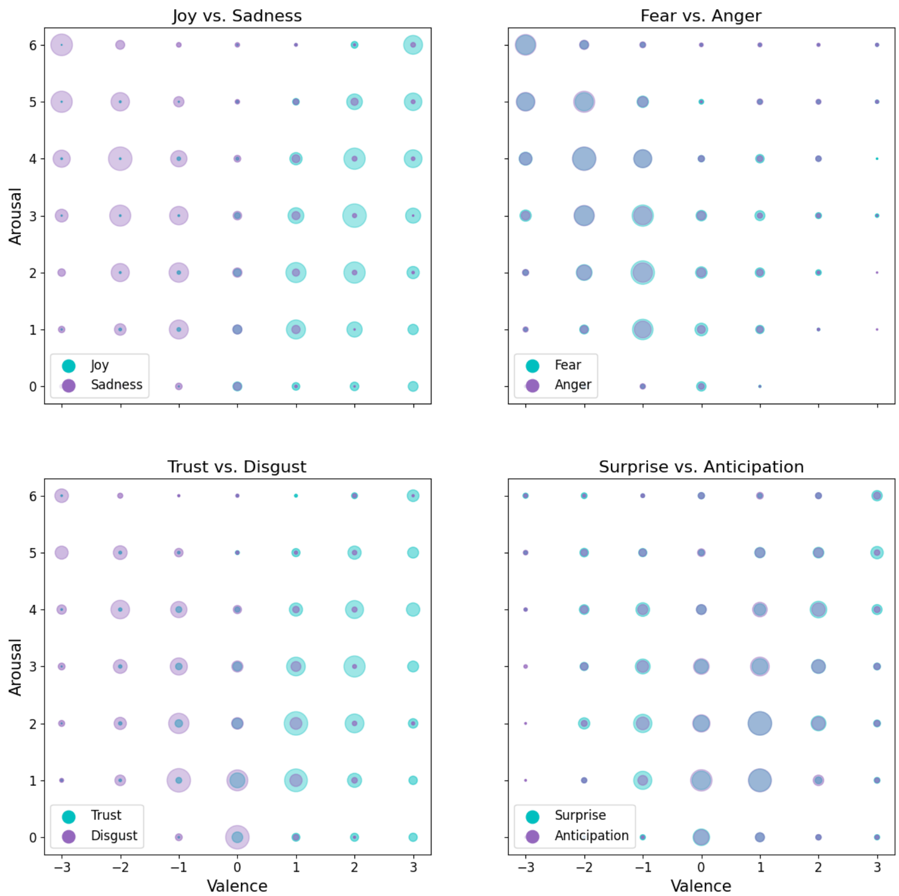

Figure 21: Association between Arousal and Valence per opposite Emo8 emotion pair. The x- and y-axes are shared between all plots. Disc sizes have been normalized between emotions.

Table 6: Average Arousal, Valence and Ambiguity annotation values for the public set, per emotion. Emotions are grouped per "ring" in Plutchik's Wheel of Emotions: from outer, least intense ring (top row) to inner, most intense ring (bottom row). The percentage of total annotations per emotion is shown in square brackets. Format: $x.xx^{y.yy}$ should be read as average = $x.xx$, standard deviation = $y.yy$.

	Serenity	Acceptance	Apprehension	Distraction	Pensiveness	Boredom	Annoyance	Interest
Nb.	1972 [7.6%]	1388 [5.4%]	1400 [5.4%]	356 [1.4%]	706 [2.7%]	587 [2.3%]	1008 [3.9%]	2688 [10.4%]
Arousal	$2.24^{1.00}$	$2.21^{1.05}$	$2.73^{1.12}$	$2.04^{1.11}$	$2.53^{1.17}$	$1.65^{0.94}$	$2.78^{1.02}$	$2.03^{0.99}$
Valence	$1.46^{1.00}$	$1.14^{1.05}$	$-0.96^{1.12}$	$-0.06^{1.11}$	$-0.90^{1.17}$	$-0.36^{0.94}$	$-1.20^{1.02}$	$0.79^{0.99}$
Ambiguity	$1.83^{1.65}$	$2.00^{1.65}$	$2.26^{1.61}$	$2.71^{1.65}$	$2.33^{1.63}$	$2.30^{1.63}$	$2.21^{1.55}$	$2.16^{1.62}$
	Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
Nb.	3971 [15.4%]	1145 [4.4%]	679 [2.6%]	311 [1.2%]	1092 [4.2%]	248 [1.0%]	985 [3.8%]	1780 [6.9%]
Arousal	$3.04^{0.86}$	$2.57^{1.05}$	$3.76^{1.13}$	$2.76^{1.41}$	$3.44^{1.18}$	$3.31^{1.08}$	$3.99^{1.11}$	$2.58^{1.26}$
Valence	$2.02^{0.86}$	$1.49^{1.05}$	$-1.79^{1.13}$	$0.35^{1.41}$	$-1.65^{1.18}$	$-1.62^{1.08}$	$-1.80^{1.11}$	$0.49^{1.26}$
Ambiguity	$1.47^{1.60}$	$1.75^{1.59}$	$1.92^{1.53}$	$2.21^{1.70}$	$1.82^{1.61}$	$2.12^{1.59}$	$1.86^{1.63}$	$2.24^{1.55}$
	Ecstasy	Admiration	Terror	Amazement	Grief	Loathing	Rage	Vigilance
Nb.	1083 [4.2%]	1016 [3.9%]	322 [1.2%]	221 [0.9%]	867 [3.4%]	165 [0.6%]	446 [1.7%]	1433 [5.5%]
Arousal	$3.97^{0.93}$	$3.06^{1.09}$	$4.34^{1.37}$	$3.14^{1.31}$	$4.12^{1.29}$	$3.93^{1.41}$	$4.54^{1.38}$	$3.13^{1.41}$
Valence	$2.27^{0.93}$	$1.70^{1.09}$	$-2.03^{1.37}$	$1.51^{1.31}$	$-2.01^{1.29}$	$-1.57^{1.41}$	$-1.92^{1.38}$	$0.22^{1.41}$
Ambiguity	$1.52^{1.83}$	$1.84^{1.66}$	$1.69^{1.65}$	$2.13^{1.61}$	$1.46^{1.64}$	$2.07^{1.79}$	$1.80^{1.72}$	$2.02^{1.66}$

A.11.1 UnbalancedCrossEntropyLoss

As stated in the main text, UnbalancedCrossEntropyLoss ($\mathcal{L}_{\text{UCE}}$) allows to give different weights to different misclassifications. E.g., it allows to penalize classifying a "joy" as a "sadness" image heavier than classifying it as "anticipation". It is defined as

$$\mathcal{L}_{\text{UCE}} = \begin{cases} w_t \log p_t & t = h \\ w_t \log p_t + w_{t,h} \log 1 - p_h & t \neq h \end{cases} \tag{1}$$

with t the target class with predicted probability p_t, h the class with the highest predicted probability p_h, w_t the weight of the target class, and $w_{t,h}$ the weight for misclassifying a sample of class t as class h. In case $t = h$, this reverts to regular CrossEntropyLoss.

To be able to use UnbalancedCrossEntropy loss, a distance needs to be defined between each pair of output classes. For the Emo8 task, we use the shortest number of leaves between two emotions. E.g., the distance between "joy" and "surprise" is 3, and the distance between "joy" and "anger" is 2.

The class weight w_i for class i was computed according to

$$w_i = \frac{N}{N_c \cdot N_i}, \tag{2}$$

with N the total number of samples, N_c the number of classes and N_i the number of samples of class i.

Finally, the weight $w_{i,j}$ for misclassifying a sample from class i as class j was computed as

$$w_{i,j} = \frac{d_{i,j}}{1 + w_j} \cdot w_i, \tag{3}$$

with w_i the weight for class i, w_j the weight for class j and $d_{i,j}$ the distance between classes i and j.

A.11.2 WeightedMSELoss

WeightedMSELoss ($\mathcal{L}_{\text{WMSE}}$) is a natural extension of the standard MSELoss to include class weights. Its mathematical formulation reads

$$\mathcal{L}_{\text{WMSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(w_i \cdot (o_i - t_i)^2 \right), \tag{4}$$

with N the number of samples, w_i the class weights as defined in Eq.2 and o_i and t_i the network output and target value for sample i respectively.

A.12 Additional Baseline Results

Baseline results for train data are depicted in Figure 22. Numerical baseline results on the train and test sets have been grouped in Tables 8 and 9 respectively. A breakdown per Emo8 for train and test sets is shown in Tables 10 and 11 respectively. For per class results, no Average Precision scores are reported, as we did not collect these.

A.13 Additional Beyond Baseline Results

The beyond baseline Emo8 classification results on train data are shown in Figure 23. Barcharts depicting the beyond baseline results for the Arousal and Valence regression tasks are grouped in Figures 24 and 25 respectively.

For the Arousal and Valence regression tasks, we dropped experiments with Places365 models in favor of precomputed Emo8 predictions using the same model architecture. E.g., for the Arousal task and AlexNet architecture, we combine the 8-feature vector obtained by applying a pretrained AlexNet Emo8 classifier with the 1-feature vector obtained by applying a pretrained AlexNet Arousal regressor. We also changed the merger network from simply concatenating the stream features to first reducing the second stream features to 1D by means of a linear layer + sigmoid, and then concatening both 1D features (from the precomputed Arousal/Valence regression + reduced second

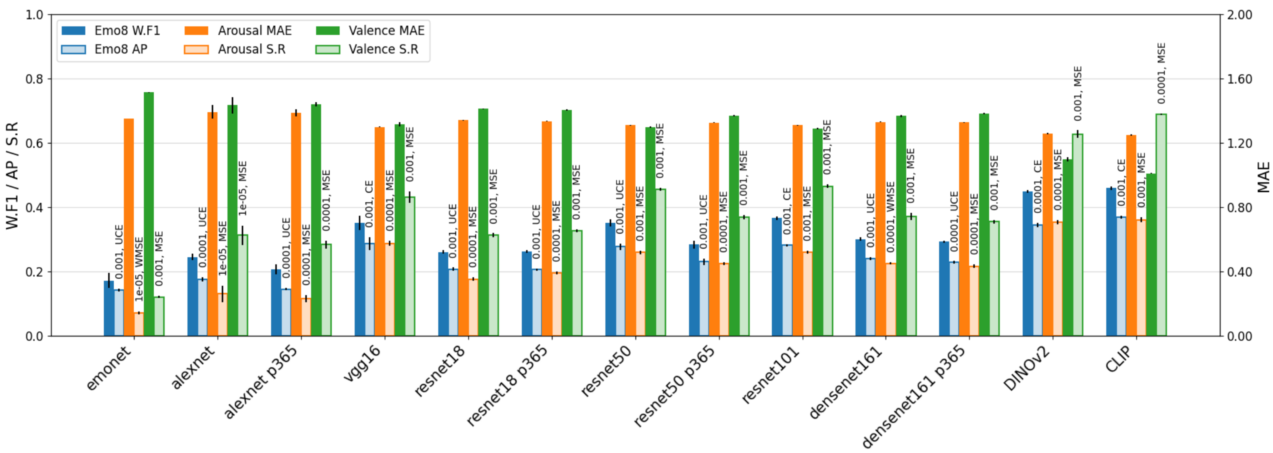

Figure 22: Train data baseline classification performance on the Emo8 classification and Arousal/Valence regression tasks. Metrics are: Weighted F1 (W.F1) and Average Precision (AP) for classification, and Mean Average Error (MAE) and Spearman R (S.R) for regression. The starting learning rate and loss corresponding to each model are displayed above the training bars. (U)CE = (Unbalanced)CrossEntropyLoss, (W)MSE = (Weighted)MeanSquaredErrorLoss, p365 = original model trained on Places365 dataset.

Table 7: Loss comparison for Emo8 classification and Arousal/Valence regression tasks, comparing test results for baseline models. Performance metrics format: $.xxx^{yy}$ should be read as average = $0.xxx$ (or $z.xxx$ if z is specified), standard deviation = $.0yy$, taken over 10 runs. Best results in bold

		emonet	alexnet	alexnet p365	vgg16	resnet18	resnet18 p365	resnet50	resnet50 p365	resnet101	densenet161	densenet161 p365	CLIP
							Emo8						
CE	Start LR	0.0001	0.0001	0.0001	0.001	0.0001	0.0001	0.001	0.0001	0.001	0.0001	0.001	0.001
	Weighted F1	$.178^{16}$	$.273^{18}$	$.237^{16}$	$\mathbf{.311^{08}}$	$.293^{04}$	$.281^{07}$	$.328^{13}$	$.289^{09}$	$\mathbf{.339^{09}}$	$.316^{08}$	$.298^{12}$	$\mathbf{.440^{07}}$
	Avg.Prec.	$\mathbf{.148^{02}}$	$\mathbf{.196^{06}}$	$\mathbf{.162^{06}}$	$\mathbf{.232^{05}}$	$\mathbf{.223^{03}}$	$\mathbf{.215^{06}}$	$.249^{05}$	$\mathbf{.228^{06}}$	$\mathbf{.250^{03}}$	$\mathbf{.247^{05}}$	$\mathbf{.231^{02}}$	$\mathbf{.346^{05}}$
UCE	Start LR	0.001	0.0001	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001	0.001
	Weighted F1	$\mathbf{.196^{18}}$	$\mathbf{.276^{18}}$	$\mathbf{.258^{14}}$	$.310^{07}$	$\mathbf{.298^{22}}$	$\mathbf{.284^{12}}$	$\mathbf{.335^{07}}$	$\mathbf{.296^{13}}$	$.335^{09}$	$\mathbf{.327^{04}}$	$\mathbf{.301^{09}}$	$.431^{08}$
	Avg.Prec.	$.146^{02}$	$.183^{10}$	$.150^{07}$	$.230^{03}$	$.223^{04}$	$.210^{03}$	$\mathbf{.251^{04}}$	$.227^{03}$	$.247^{03}$	$.244^{03}$	$.228^{06}$	$.340^{03}$
							Arousal						
MSE	Start LR	0.001	$1e-05$	0.0001	0.0001	0.0001	0.0001	0.001	0.0001	0.001	0.001	0.0001	0.001
	MAE	1.354^{01}	$\mathbf{1.333^{03}}$	$\mathbf{1.344^{03}}$	$\mathbf{1.320^{06}}$	$\mathbf{1.334^{03}}$	$\mathbf{1.333^{03}}$	$\mathbf{1.311^{06}}$	$\mathbf{1.327^{03}}$	$\mathbf{1.314^{04}}$	1.320^{03}	$\mathbf{1.326^{04}}$	$\mathbf{1.260^{06}}$
	Spearman R	$\mathbf{.094^{12}}$	$\mathbf{.203^{10}}$	$\mathbf{.173^{13}}$	$\mathbf{.238^{14}}$	$\mathbf{.214^{12}}$	$.212^{07}$	$\mathbf{.264^{10}}$	$\mathbf{.226^{05}}$	$.254^{12}$	$\mathbf{.251^{10}}$	$\mathbf{.231^{13}}$	$\mathbf{.344^{10}}$
WMSE	Start LR	$1e-05$	0.0001	0.0001	$1e-05$	0.0001	0.0001	0.0001	$1e-05$	0.0001	0.0001	0.0001	$1e-05$
	MAE	$\mathbf{1.353^{02}}$	1.350^{07}	1.361^{06}	1.324^{06}	1.338^{04}	1.335^{05}	1.318^{05}	1.332^{04}	1.317^{06}	$\mathbf{1.319^{05}}$	1.330^{04}	1.265^{07}
	Spearman R	$.072^{18}$	$.193^{10}$	$.171^{11}$	$.238^{13}$	$.208^{11}$	$\mathbf{.213^{11}}$	$.250^{09}$	$.219^{10}$	$\mathbf{.254^{11}}$	$.249^{13}$	$.227^{10}$	$.340^{09}$
							Valence						
MSE	Start LR	0.001	$1e-05$	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001
	MAE	$\mathbf{1.516^{03}}$	$\mathbf{1.363^{08}}$	$\mathbf{1.398^{09}}$	$\mathbf{1.312^{08}}$	$\mathbf{1.385^{05}}$	$\mathbf{1.382^{07}}$	$\mathbf{1.290^{07}}$	$\mathbf{1.344^{06}}$	$\mathbf{1.276^{06}}$	$\mathbf{1.336^{11}}$	$\mathbf{1.355^{06}}$	$\mathbf{1.015^{07}}$
	Spearman R	$.117^{12}$	$.383^{10}$	$.342^{10}$	$\mathbf{.434^{09}}$	$.394^{13}$	$\mathbf{.378^{11}}$	$\mathbf{.477^{07}}$	$\mathbf{.414^{09}}$	$\mathbf{.487^{08}}$	$\mathbf{.440^{14}}$	$.408^{10}$	$\mathbf{.685^{05}}$
WMSE	Start LR	0.001	$1e-05$	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001	0.001
	MAE	1.564^{02}	1.393^{09}	1.424^{10}	1.341^{11}	1.412^{09}	1.409^{10}	1.310^{12}	1.370^{09}	1.295^{08}	1.353^{07}	1.378^{08}	1.033^{09}
	Spearman R	$\mathbf{.118^{12}}$	$\mathbf{.386^{12}}$	$\mathbf{.347^{13}}$	$.432^{09}$	$\mathbf{.395^{09}}$	$.374^{09}$	$.471^{14}$	$.413^{11}$	$.483^{09}$	$.439^{07}$	$\mathbf{.410^{10}}$	$.676^{06}$

Table 8: Training results: Emo8 classification and Arousal/Valence regression performance for baseline models. Performance metrics format: $.xxx^{yy}$ should be read as average = $0.xxx$ (or $z.xxx$ if z is specified), standard deviation = $.0yy$, taken over 10 runs.

	emonet	alexnet	alexnet p365	vgg16	resnet18	resnet18 p365	resnet50	resnet50 p365	resnet101	densenet161	densenet161 p365	DINOv2	CLIP
							Emo8						
Loss	UCE	UCE	UCE	CE	UCE	UCE	UCE	UCE	CE	UCE	UCE	CE	CE
Start LR	0.001	0.0001	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001	0.0001	0.001
Accuracy	$.163^{22}$	$.231^{10}$	$.191^{15}$	$.340^{24}$	$.248^{05}$	$.247^{05}$	$.339^{11}$	$.268^{12}$	$.355^{05}$	$.287^{05}$	$.281^{04}$	$.478^{04}$	$.445^{04}$
F1	$.144^{17}$	$.211^{09}$	$.175^{14}$	$.315^{26}$	$.222^{05}$	$.223^{04}$	$.308^{13}$	$.244^{12}$	$.321^{05}$	$.258^{05}$	$.249^{04}$	$.345^{05}$	$.392^{06}$
Weighted F1	$.172^{24}$	$.245^{10}$	$.206^{15}$	$.350^{23}$	$.261^{06}$	$.263^{04}$	$.351^{12}$	$.284^{12}$	$.366^{05}$	$.301^{05}$	$.293^{04}$	$.449^{05}$	$.459^{05}$
Avg.Prec.	$.142^{05}$	$.176^{07}$	$.145^{04}$	$.287^{21}$	$.208^{05}$	$.206^{03}$	$.277^{10}$	$.230^{10}$	$.282^{03}$	$.240^{05}$	$.228^{04}$	$.344^{07}$	$.370^{05}$
							Arousal						
Loss	WMSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	WMSE	MSE	MSE	MSE
Start LR	$1e-05$	$1e-05$	0.0001	0.0001	0.0001	0.0001	0.001	0.0001	0.001	0.0001	0.0001	0.0001	0.001
MAE	1.352^{00}	1.393^{42}	1.387^{22}	1.299^{03}	1.341^{01}	1.335^{02}	1.310^{02}	1.326^{02}	1.309^{03}	1.331^{02}	1.328^{02}	1.257^{03}	1.251^{05}
Spearman R	0.071^{05}	0.130^{26}	0.115^{11}	0.288^{07}	0.177^{06}	0.196^{05}	0.259^{06}	0.224^{05}	0.260^{05}	0.225^{03}	0.216^{05}	0.354^{06}	0.361^{08}
							Valence						
Loss	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Start LR	0.001	$1e-05$	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001
MAE	1.515^{02}	1.435^{51}	1.440^{15}	1.315^{14}	1.412^{03}	1.404^{03}	1.298^{05}	1.368^{05}	1.287^{04}	1.368^{07}	1.381^{04}	1.098^{14}	1.010^{02}
Spearman R	0.121^{03}	0.313^{30}	0.284^{12}	0.431^{18}	0.313^{06}	0.326^{05}	0.455^{05}	0.369^{07}	0.465^{05}	0.371^{11}	0.355^{05}	0.627^{12}	0.689^{02}

Table 9: Test results: Emo8 classification and Arousal/Valence regression performance for baseline models. Performance metrics format: $.xxx^{yy}$ should be read as average = $0.xxx$ (or $z.xxx$ if z is specified), standard deviation = $.0yy$, taken over 10 runs.

	emonet	alexnet	alexnet p365	vgg16	resnet18	resnet18 p365	resnet50	resnet50 p365	resnet101	densenet161	densenet161 p365	DINOv2	CLIP
							Emo8						
Loss	UCE	UCE	UCE	CE	UCE	UCE	UCE	UCE	CE	UCE	UCE	CE	CE
Start LR	0.001	0.0001	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001	0.0001	0.001
Accuracy	$.192^{22}$	$.280^{21}$	$.259^{22}$	$.303^{10}$	$.301^{19}$	$.286^{16}$	$.327^{07}$	$.296^{13}$	$.329^{11}$	$.330^{13}$	$.294^{10}$	$.450^{07}$	$.427^{07}$
F1	$.152^{10}$	$.219^{09}$	$.201^{09}$	$.251^{07}$	$.240^{14}$	$.226^{07}$	$.275^{05}$	$.235^{11}$	$.277^{06}$	$.260^{05}$	$.245^{08}$	$.320^{08}$	$.365^{07}$
Weighted F1	$.196^{18}$	$.276^{18}$	$.258^{14}$	$.311^{08}$	$.298^{22}$	$.284^{12}$	$.335^{07}$	$.296^{13}$	$.339^{09}$	$.327^{04}$	$.301^{09}$	$.424^{07}$	$.440^{07}$
Avg.Prec.	$.146^{02}$	$.183^{10}$	$.150^{07}$	$.232^{05}$	$.223^{04}$	$.210^{03}$	$.251^{04}$	$.227^{03}$	$.250^{03}$	$.244^{03}$	$.228^{06}$	$.310^{05}$	$.346^{05}$
							Arousal						
Loss	WMSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	WMSE	MSE	MSE	MSE
Start LR	$1e-05$	$1e-05$	0.0001	0.0001	0.0001	0.0001	0.001	0.0001	0.001	0.0001	0.0001	0.0001	0.001
MAE	1.353^{02}	1.333^{03}	1.344^{03}	1.320^{06}	1.334^{03}	1.333^{03}	1.311^{06}	1.327^{03}	1.314^{04}	1.319^{05}	1.326^{04}	1.276^{07}	1.260^{06}
Spearman R	0.072^{18}	0.203^{10}	0.173^{13}	0.238^{14}	0.214^{12}	0.212^{07}	0.264^{10}	0.226^{05}	0.254^{12}	0.249^{13}	0.231^{13}	0.322^{12}	0.344^{10}
							Valence						
Loss	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE	MSE
Start LR	0.001	$1e-05$	0.0001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.0001
MAE	1.516^{03}	1.363^{08}	1.398^{09}	1.312^{08}	1.385^{05}	1.382^{07}	1.290^{07}	1.344^{06}	1.276^{06}	1.336^{11}	1.355^{06}	1.107^{12}	1.015^{07}
Spearman R	0.117^{12}	0.383^{10}	0.342^{10}	0.434^{09}	0.394^{13}	0.378^{11}	0.477^{07}	0.414^{09}	0.487^{08}	0.440^{14}	0.408^{10}	0.618^{09}	0.685^{05}

Table 10: Training results: Emo8 Recall, Precision and F1 metrics per emotion leaf for baseline models. Format: $0.xx^{\cdot yy}$ with $0.xx$ = average, $0.yy$ = standard deviation over 10 runs.

	Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
				alexnet				
Recall	$0.25^{\cdot 01}$	$0.20^{\cdot 01}$	$0.25^{\cdot 01}$	$0.19^{\cdot 02}$	$0.30^{\cdot 02}$	$0.20^{\cdot 01}$	$0.32^{\cdot 02}$	$0.16^{\cdot 01}$
Precision	$0.46^{\cdot 02}$	$0.21^{\cdot 01}$	$0.20^{\cdot 01}$	$0.05^{\cdot 01}$	$0.23^{\cdot 01}$	$0.07^{\cdot 01}$	$0.23^{\cdot 01}$	$0.33^{\cdot 02}$
F1	$0.33^{\cdot 02}$	$0.20^{\cdot 01}$	$0.22^{\cdot 01}$	$0.08^{\cdot 01}$	$0.26^{\cdot 02}$	$0.10^{\cdot 01}$	$0.27^{\cdot 01}$	$0.21^{\cdot 01}$
				alexnet p365				
Recall	$0.21^{\cdot 02}$	$0.17^{\cdot 01}$	$0.21^{\cdot 02}$	$0.18^{\cdot 02}$	$0.23^{\cdot 03}$	$0.19^{\cdot 03}$	$0.23^{\cdot 03}$	$0.15^{\cdot 00}$
Precision	$0.41^{\cdot 03}$	$0.18^{\cdot 01}$	$0.16^{\cdot 01}$	$0.05^{\cdot 01}$	$0.19^{\cdot 02}$	$0.06^{\cdot 01}$	$0.18^{\cdot 02}$	$0.30^{\cdot 03}$
F1	$0.27^{\cdot 02}$	$0.18^{\cdot 01}$	$0.18^{\cdot 02}$	$0.08^{\cdot 01}$	$0.21^{\cdot 03}$	$0.09^{\cdot 01}$	$0.20^{\cdot 02}$	$0.20^{\cdot 01}$
				vgg16				
Recall	$0.36^{\cdot 02}$	$0.28^{\cdot 03}$	$0.36^{\cdot 02}$	$0.36^{\cdot 07}$	$0.45^{\cdot 03}$	$0.35^{\cdot 06}$	$0.47^{\cdot 02}$	$0.23^{\cdot 02}$
Precision	$0.56^{\cdot 02}$	$0.30^{\cdot 02}$	$0.29^{\cdot 02}$	$0.12^{\cdot 02}$	$0.34^{\cdot 02}$	$0.14^{\cdot 03}$	$0.33^{\cdot 02}$	$0.47^{\cdot 03}$
F1	$0.44^{\cdot 02}$	$0.29^{\cdot 03}$	$0.32^{\cdot 02}$	$0.18^{\cdot 04}$	$0.39^{\cdot 03}$	$0.20^{\cdot 04}$	$0.39^{\cdot 02}$	$0.31^{\cdot 02}$
				resnet18				
Recall	$0.29^{\cdot 00}$	$0.20^{\cdot 01}$	$0.27^{\cdot 01}$	$0.16^{\cdot 02}$	$0.32^{\cdot 01}$	$0.21^{\cdot 02}$	$0.36^{\cdot 01}$	$0.15^{\cdot 01}$
Precision	$0.48^{\cdot 01}$	$0.22^{\cdot 01}$	$0.22^{\cdot 01}$	$0.05^{\cdot 00}$	$0.24^{\cdot 01}$	$0.07^{\cdot 01}$	$0.24^{\cdot 01}$	$0.36^{\cdot 01}$
F1	$0.36^{\cdot 01}$	$0.21^{\cdot 01}$	$0.24^{\cdot 01}$	$0.08^{\cdot 01}$	$0.27^{\cdot 01}$	$0.11^{\cdot 01}$	$0.29^{\cdot 01}$	$0.22^{\cdot 01}$
				resnet18 p365				
Recall	$0.29^{\cdot 01}$	$0.19^{\cdot 01}$	$0.25^{\cdot 01}$	$0.18^{\cdot 01}$	$0.31^{\cdot 01}$	$0.24^{\cdot 02}$	$0.35^{\cdot 01}$	$0.17^{\cdot 01}$
Precision	$0.47^{\cdot 01}$	$0.22^{\cdot 01}$	$0.21^{\cdot 01}$	$0.06^{\cdot 00}$	$0.24^{\cdot 01}$	$0.08^{\cdot 00}$	$0.24^{\cdot 01}$	$0.37^{\cdot 01}$
F1	$0.36^{\cdot 01}$	$0.21^{\cdot 01}$	$0.23^{\cdot 01}$	$0.09^{\cdot 01}$	$0.27^{\cdot 01}$	$0.12^{\cdot 01}$	$0.29^{\cdot 01}$	$0.23^{\cdot 01}$
				resnet50				
Recall	$0.38^{\cdot 01}$	$0.28^{\cdot 02}$	$0.34^{\cdot 01}$	$0.29^{\cdot 04}$	$0.46^{\cdot 01}$	$0.30^{\cdot 03}$	$0.46^{\cdot 02}$	$0.23^{\cdot 01}$
Precision	$0.57^{\cdot 01}$	$0.29^{\cdot 01}$	$0.30^{\cdot 01}$	$0.11^{\cdot 01}$	$0.34^{\cdot 02}$	$0.12^{\cdot 01}$	$0.33^{\cdot 01}$	$0.46^{\cdot 01}$
F1	$0.46^{\cdot 01}$	$0.28^{\cdot 01}$	$0.32^{\cdot 01}$	$0.16^{\cdot 02}$	$0.39^{\cdot 01}$	$0.17^{\cdot 02}$	$0.38^{\cdot 01}$	$0.31^{\cdot 01}$
				resnet50 p365				
Recall	$0.30^{\cdot 01}$	$0.20^{\cdot 01}$	$0.27^{\cdot 02}$	$0.22^{\cdot 03}$	$0.34^{\cdot 02}$	$0.26^{\cdot 03}$	$0.38^{\cdot 01}$	$0.19^{\cdot 01}$
Precision	$0.50^{\cdot 01}$	$0.23^{\cdot 01}$	$0.23^{\cdot 01}$	$0.07^{\cdot 01}$	$0.27^{\cdot 01}$	$0.09^{\cdot 01}$	$0.27^{\cdot 01}$	$0.39^{\cdot 02}$
F1	$0.38^{\cdot 01}$	$0.22^{\cdot 01}$	$0.25^{\cdot 01}$	$0.10^{\cdot 02}$	$0.30^{\cdot 01}$	$0.13^{\cdot 02}$	$0.32^{\cdot 01}$	$0.25^{\cdot 02}$
				resnet101				
Recall	$0.40^{\cdot 00}$	$0.28^{\cdot 01}$	$0.37^{\cdot 01}$	$0.28^{\cdot 02}$	$0.49^{\cdot 01}$	$0.31^{\cdot 02}$	$0.47^{\cdot 01}$	$0.25^{\cdot 01}$
Precision	$0.58^{\cdot 01}$	$0.30^{\cdot 01}$	$0.31^{\cdot 01}$	$0.11^{\cdot 01}$	$0.35^{\cdot 01}$	$0.13^{\cdot 01}$	$0.33^{\cdot 01}$	$0.47^{\cdot 01}$
F1	$0.47^{\cdot 01}$	$0.29^{\cdot 01}$	$0.33^{\cdot 01}$	$0.16^{\cdot 01}$	$0.41^{\cdot 01}$	$0.19^{\cdot 01}$	$0.39^{\cdot 01}$	$0.33^{\cdot 01}$
				densenet161				
Recall	$0.34^{\cdot 01}$	$0.23^{\cdot 01}$	$0.30^{\cdot 01}$	$0.19^{\cdot 03}$	$0.39^{\cdot 01}$	$0.25^{\cdot 02}$	$0.40^{\cdot 01}$	$0.18^{\cdot 01}$
Precision	$0.52^{\cdot 01}$	$0.24^{\cdot 01}$	$0.25^{\cdot 01}$	$0.07^{\cdot 01}$	$0.29^{\cdot 01}$	$0.08^{\cdot 01}$	$0.29^{\cdot 00}$	$0.40^{\cdot 01}$
F1	$0.41^{\cdot 01}$	$0.24^{\cdot 01}$	$0.27^{\cdot 01}$	$0.10^{\cdot 01}$	$0.33^{\cdot 01}$	$0.13^{\cdot 01}$	$0.34^{\cdot 01}$	$0.25^{\cdot 01}$
				densenet161 p365				
Recall	$0.33^{\cdot 01}$	$0.21^{\cdot 01}$	$0.28^{\cdot 01}$	$0.15^{\cdot 02}$	$0.37^{\cdot 01}$	$0.24^{\cdot 02}$	$0.42^{\cdot 01}$	$0.19^{\cdot 01}$
Precision	$0.50^{\cdot 01}$	$0.24^{\cdot 01}$	$0.23^{\cdot 01}$	$0.07^{\cdot 01}$	$0.27^{\cdot 00}$	$0.09^{\cdot 00}$	$0.26^{\cdot 01}$	$0.40^{\cdot 01}$
F1	$0.40^{\cdot 01}$	$0.22^{\cdot 01}$	$0.25^{\cdot 01}$	$0.09^{\cdot 01}$	$0.31^{\cdot 01}$	$0.13^{\cdot 01}$	$0.32^{\cdot 01}$	$0.26^{\cdot 01}$
				DINOv2				
Recall	$0.71^{\cdot 00}$	$0.22^{\cdot 01}$	$0.27^{\cdot 01}$	$0.00^{\cdot 00}$	$0.51^{\cdot 01}$	$0.04^{\cdot 01}$	$0.43^{\cdot 01}$	$0.59^{\cdot 01}$
Precision	$0.56^{\cdot 00}$	$0.41^{\cdot 01}$	$0.37^{\cdot 01}$	$0.27^{\cdot 12}$	$0.53^{\cdot 00}$	$0.32^{\cdot 04}$	$0.47^{\cdot 01}$	$0.42^{\cdot 00}$
F1	$0.62^{\cdot 00}$	$0.28^{\cdot 01}$	$0.31^{\cdot 01}$	$0.01^{\cdot 00}$	$0.52^{\cdot 01}$	$0.07^{\cdot 01}$	$0.45^{\cdot 01}$	$0.49^{\cdot 00}$
				CLIP				
Recall	$0.57^{\cdot 01}$	$0.37^{\cdot 01}$	$0.35^{\cdot 01}$	$0.25^{\cdot 03}$	$0.56^{\cdot 01}$	$0.34^{\cdot 02}$	$0.56^{\cdot 01}$	$0.33^{\cdot 01}$
Precision	$0.69^{\cdot 01}$	$0.35^{\cdot 01}$	$0.33^{\cdot 01}$	$0.11^{\cdot 01}$	$0.53^{\cdot 01}$	$0.15^{\cdot 01}$	$0.45^{\cdot 00}$	$0.52^{\cdot 01}$
F1	$0.62^{\cdot 00}$	$0.36^{\cdot 01}$	$0.34^{\cdot 01}$	$0.15^{\cdot 01}$	$0.55^{\cdot 01}$	$0.21^{\cdot 01}$	$0.50^{\cdot 01}$	$0.40^{\cdot 01}$

stream) and send these through a final linear layer + sigmoid. Finally, we only consider MSELoss, and $\text{lr}_0 \in [10^{-3}, 10^{-4}, 10^{-5}]$.

A numerical comparison of the baseline to the (overall best performing) "Baseline+OIToFER" model for all three tasks is included in Table 12. From this, it is apparent that obtainable gains are architecture-dependent, with the VGG16 and ResNet50 architectures obtaining most gains and the DenseNet161, CLIP and DINOv2 architectures barely improving, regardless of task. Obtained gains

Table 11: Test results: Emo8 Recall, Precision and F1 metrics per emotion leaf for baseline models. Format: $0.xx^{.yy}$ with $0.xx$ = average, $0.yy$ = standard deviation over 10 runs.

	Joy	Trust	Fear	Surprise	Sadness	Disgust	Anger	Anticipation
				alexnet				
Recall	$0.38^{.09}$	$0.18^{.10}$	$0.24^{.08}$	$0.05^{.05}$	$0.36^{.09}$	$0.08^{.05}$	$0.38^{.08}$	$0.23^{.08}$
Precision	$0.46^{.04}$	$0.22^{.02}$	$0.20^{.02}$	$0.04^{.01}$	$0.25^{.04}$	$0.07^{.01}$	$0.25^{.03}$	$0.35^{.02}$
F1	$0.40^{.06}$	$0.18^{.05}$	$0.21^{.03}$	$0.04^{.02}$	$0.29^{.01}$	$0.07^{.02}$	$0.29^{.02}$	$0.27^{.05}$
				alexnet p365				
Recall	$0.40^{.09}$	$0.17^{.08}$	$0.24^{.09}$	$0.10^{.09}$	$0.28^{.08}$	$0.08^{.06}$	$0.24^{.08}$	$0.20^{.09}$
Precision	$0.42^{.02}$	$0.22^{.02}$	$0.19^{.02}$	$0.05^{.01}$	$0.24^{.04}$	$0.05^{.02}$	$0.23^{.03}$	$0.33^{.03}$
F1	$0.40^{.04}$	$0.18^{.05}$	$0.20^{.03}$	$0.06^{.02}$	$0.25^{.02}$	$0.05^{.03}$	$0.22^{.04}$	$0.24^{.06}$
				vgg16				
Recall	$0.40^{.04}$	$0.20^{.03}$	$0.26^{.05}$	$0.09^{.04}$	$0.38^{.04}$	$0.14^{.05}$	$0.40^{.08}$	$0.25^{.07}$
Precision	$0.50^{.01}$	$0.24^{.02}$	$0.23^{.02}$	$0.05^{.01}$	$0.30^{.02}$	$0.07^{.02}$	$0.30^{.05}$	$0.38^{.03}$
F1	$0.44^{.02}$	$0.21^{.02}$	$0.24^{.02}$	$0.06^{.02}$	$0.33^{.02}$	$0.09^{.02}$	$0.34^{.02}$	$0.29^{.04}$
				resnet18				
Recall	$0.36^{.09}$	$0.31^{.11}$	$0.37^{.09}$	$0.04^{.04}$	$0.30^{.11}$	$0.09^{.05}$	$0.33^{.10}$	$0.27^{.07}$
Precision	$0.50^{.05}$	$0.22^{.02}$	$0.23^{.02}$	$0.04^{.03}$	$0.31^{.07}$	$0.10^{.02}$	$0.30^{.03}$	$0.36^{.04}$
F1	$0.40^{.08}$	$0.24^{.04}$	$0.28^{.02}$	$0.03^{.03}$	$0.28^{.06}$	$0.09^{.02}$	$0.31^{.04}$	$0.30^{.04}$
				resnet18 p365				
Recall	$0.36^{.08}$	$0.19^{.07}$	$0.30^{.09}$	$0.04^{.02}$	$0.30^{.10}$	$0.12^{.06}$	$0.35^{.08}$	$0.28^{.10}$
Precision	$0.47^{.03}$	$0.24^{.02}$	$0.19^{.02}$	$0.06^{.03}$	$0.27^{.04}$	$0.08^{.02}$	$0.25^{.03}$	$0.35^{.03}$
F1	$0.40^{.04}$	$0.20^{.04}$	$0.23^{.03}$	$0.04^{.02}$	$0.27^{.03}$	$0.08^{.02}$	$0.28^{.03}$	$0.30^{.07}$
				resnet50				
Recall	$0.39^{.05}$	$0.25^{.04}$	$0.32^{.06}$	$0.07^{.02}$	$0.44^{.06}$	$0.13^{.02}$	$0.44^{.04}$	$0.28^{.03}$
Precision	$0.54^{.02}$	$0.24^{.02}$	$0.27^{.03}$	$0.06^{.02}$	$0.33^{.03}$	$0.08^{.01}$	$0.31^{.02}$	$0.40^{.02}$
F1	$0.45^{.03}$	$0.24^{.02}$	$0.29^{.02}$	$0.06^{.01}$	$0.37^{.02}$	$0.10^{.01}$	$0.36^{.01}$	$0.33^{.02}$
				resnet50 p365				
Recall	$0.38^{.08}$	$0.19^{.11}$	$0.31^{.10}$	$0.06^{.06}$	$0.30^{.07}$	$0.10^{.05}$	$0.33^{.09}$	$0.30^{.11}$
Precision	$0.49^{.05}$	$0.24^{.02}$	$0.21^{.02}$	$0.04^{.02}$	$0.31^{.05}$	$0.07^{.01}$	$0.30^{.05}$	$0.35^{.03}$
F1	$0.42^{.04}$	$0.19^{.06}$	$0.24^{.03}$	$0.04^{.03}$	$0.30^{.02}$	$0.08^{.02}$	$0.30^{.03}$	$0.31^{.05}$
				resnet101				
Recall	$0.42^{.03}$	$0.23^{.04}$	$0.34^{.05}$	$0.08^{.03}$	$0.45^{.04}$	$0.15^{.05}$	$0.42^{.03}$	$0.25^{.04}$
Precision	$0.54^{.01}$	$0.25^{.01}$	$0.26^{.02}$	$0.05^{.01}$	$0.33^{.02}$	$0.08^{.01}$	$0.31^{.01}$	$0.43^{.03}$
F1	$0.47^{.02}$	$0.24^{.02}$	$0.29^{.02}$	$0.06^{.01}$	$0.38^{.01}$	$0.10^{.02}$	$0.36^{.01}$	$0.31^{.03}$
				densenet161				
Recall	$0.48^{.07}$	$0.17^{.06}$	$0.26^{.07}$	$0.06^{.05}$	$0.39^{.09}$	$0.13^{.09}$	$0.43^{.05}$	$0.27^{.05}$
Precision	$0.49^{.03}$	$0.28^{.02}$	$0.25^{.02}$	$0.04^{.02}$	$0.31^{.04}$	$0.08^{.02}$	$0.31^{.02}$	$0.38^{.01}$
F1	$0.48^{.02}$	$0.20^{.04}$	$0.25^{.04}$	$0.04^{.03}$	$0.34^{.02}$	$0.09^{.02}$	$0.36^{.01}$	$0.31^{.04}$
				densenet161 p365				
Recall	$0.38^{.03}$	$0.20^{.03}$	$0.27^{.04}$	$0.05^{.03}$	$0.38^{.02}$	$0.15^{.02}$	$0.43^{.03}$	$0.22^{.03}$
Precision	$0.49^{.02}$	$0.23^{.01}$	$0.23^{.01}$	$0.04^{.01}$	$0.28^{.01}$	$0.08^{.01}$	$0.25^{.02}$	$0.38^{.01}$
F1	$0.43^{.02}$	$0.21^{.02}$	$0.25^{.02}$	$0.04^{.02}$	$0.32^{.02}$	$0.10^{.01}$	$0.32^{.01}$	$0.28^{.03}$
				DINOv2				
Recall	$0.67^{.05}$	$0.23^{.05}$	$0.25^{.07}$	$0.00^{.00}$	$0.47^{.02}$	$0.02^{.02}$	$0.42^{.05}$	$0.55^{.04}$
Precision	$0.56^{.03}$	$0.36^{.03}$	$0.32^{.01}$	$0.00^{.00}$	$0.49^{.04}$	$0.17^{.08}$	$0.42^{.03}$	$0.40^{.01}$
F1	$0.61^{.01}$	$0.27^{.04}$	$0.28^{.04}$	$0.00^{.00}$	$0.48^{.02}$	$0.04^{.03}$	$0.42^{.02}$	$0.46^{.01}$
				CLIP				
Recall	$0.57^{.03}$	$0.33^{.03}$	$0.35^{.05}$	$0.17^{.02}$	$0.54^{.02}$	$0.21^{.03}$	$0.52^{.03}$	$0.34^{.04}$
Precision	$0.67^{.01}$	$0.33^{.01}$	$0.32^{.02}$	$0.08^{.01}$	$0.52^{.03}$	$0.11^{.01}$	$0.43^{.02}$	$0.48^{.01}$
F1	$0.61^{.02}$	$0.33^{.02}$	$0.33^{.03}$	$0.11^{.01}$	$0.53^{.02}$	$0.14^{.01}$	$0.47^{.01}$	$0.40^{.02}$

are highest for the Emo8 task (absolute 13.7% AP gain, or relative 59%, for VGG16). Gains for the Arousal and Valence tasks are somewhat less straightforward to compare, as changes in MAE and Spearman R do not always agree. Compare, e.g., for ResNet50, for the Valence task an absolute 0.121 MAE and 0.120 Spearman R improvement, or relative 9.0% and 22.5% respectively, to an absolute 0.039 MAE and 0.087 Spearman R improvement, or relative 2.9% and 38% respectively for the Arousal task.

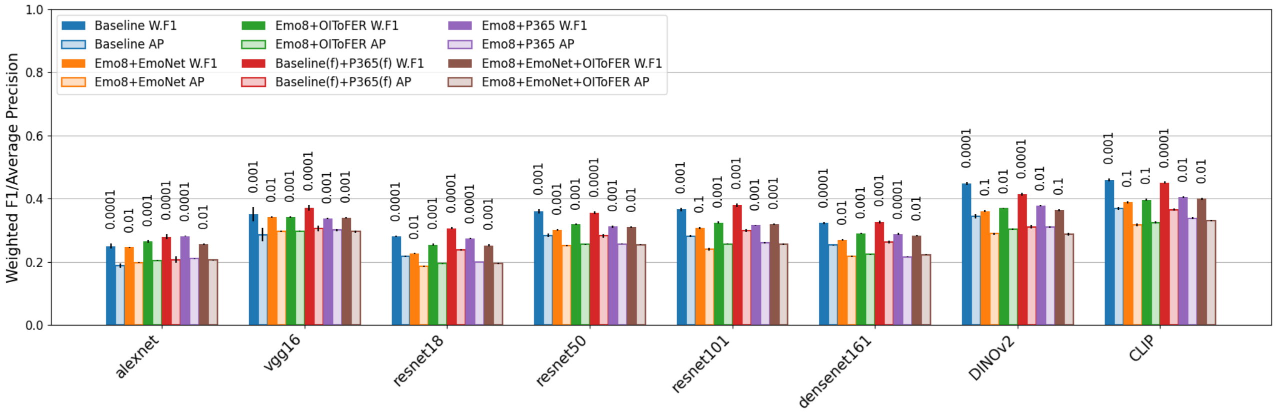

Figure 23: Training data results for extensions beyond the ImageNet baseline by applying late fusion with EmoNet predictions (EmoNet), Facial Emotion Recognition predictions (OIToFER) and Places365 (P365) predictions or features. For all models, predictions on the dataset are concatenated and sent through a linear layer, except when '(f)' is shown, indicating model features are concatenated. The starting learning rate corresponding to each model is displayed above the training bars.

A.14 A Note on the Fuzziness of Emotion Recognition

As shown in §A.9, if there is disagreement between annotators concerning the emotion depicted in an image, then it is typically among similar emotions. So, although annotators often disagree, the different labels provided for a same image are far from random, instead showing clear tendencies toward a specific region of the emotion spectrum.

This fuzzyiness in assigned labels is a feature of human psychology. Emotion recognition is hard, nuanced, and multidimensional. With more raters, one would obtain a distribution of responses, but still no perfect agreement. To compound this issue, the estimate of the distribution per image would be poor unless one has many raters per image (tens of raters for tens of thousands of images!). This is, for many reasons not least of which financially, highly impractical.

Having one rater per image gives uncertainty if one is interested in a single image, but the average performance across many images is still meaningful. To demonstrate this, we perform the following experiment: for several architectures, we train an Emo8 prediction model on our dataset using the approach described in Section 3, and once trained, we let the model make predictions for each image in the full public dataset (i.e., train + test splits). We train 5 models per $\mathrm{lr}_0 \in [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}]$ using CrossEntropyLoss, and keep the one with the highest Weighted F1 score as the winner. For these models, we list in Table 13 how often the annotated emotion was ranked N (out of 8), and in Table 14 we show the distance between the top predicted and annotated emotions. Except for AlexNet, all other models show a nice downward sloping behavior as either the rank (Table 13) or distance (Table 14) increases. In other words, the "mistakes" made by these models are clearly not random, but show behavior that is similar to those observed in the human annotators. This confirms that even with a single annotation per image, already valuable results and insights can be obtained.

A.15 Author Responsibility Statement

We, the authors, confirm that we bear all responsibility in case of any violation of rights during the collection of the data or other work, and that we will take appropriate action if and when needed,

Table 12: Baseline vs. +OIToFER: A comparison of Emo8 classification and Arousal/Valence regression performance on the test data.

		alexnet	vgg16	resnet18	resnet50	resnet101	densenet161	DINOv2_B	CLIP
					Emo8				
Baseline	Start LR	0.0001	0.001	0.0001	0.001	0.001	0.0001	0.0001	0.001
	Accuracy	$.271^{25}$	$.303^{10}$	$.288^{04}$	$.317^{17}$	$.329^{11}$	$.308^{09}$	$.450^{07}$	$.427^{07}$
	F1	$.221^{10}$	$.251^{07}$	$.242^{04}$	$.271^{10}$	$.277^{06}$	$.261^{08}$	$.320^{08}$	$.365^{07}$
	Weighted F1	$.273^{18}$	$.311^{08}$	$.293^{04}$	$.328^{13}$	$.339^{09}$	$.316^{08}$	$.424^{07}$	$.440^{07}$
	Avg.Prec.	$.196^{06}$	$.232^{05}$	$.223^{03}$	$.249^{05}$	$.250^{03}$	$.247^{05}$	$.310^{05}$	$.346^{05}$
+OIToFER	Start LR	0.001	0.001	0.001	0.001	0.001	0.001	0.01	0.1
	Accuracy	$.299^{04}$	$.390^{05}$	$.315^{06}$	$.366^{05}$	$.367^{07}$	$.343^{04}$	$.451^{03}$	$.475^{10}$
	F1	$.243^{03}$	$.363^{05}$	$.239^{06}$	$.321^{04}$	$.321^{07}$	$.278^{05}$	$.288^{05}$	$.322^{16}$
	Weighted F1	$.296^{04}$	$.399^{05}$	$.299^{07}$	$.372^{04}$	$.372^{07}$	$.339^{04}$	$.395^{05}$	$.430^{20}$
	Avg.Prec.	$.237^{05}$	$.369^{05}$	$.232^{02}$	$.311^{05}$	$.313^{06}$	$.269^{03}$	$.360^{07}$	$.384^{06}$
					Arousal				
Baseline	Start LR	$1e-05$	0.0001	0.0001	0.001	0.001	0.001	0.0001	0.001
	MAE	1.333^{03}	1.320^{06}	1.334^{03}	1.311^{06}	1.314^{04}	1.320^{03}	1.276^{07}	1.260^{06}
	Spearman R	$.203^{10}$	$.238^{14}$	$.214^{12}$	$.264^{10}$	$.254^{12}$	$.251^{10}$	$.322^{12}$	$.344^{10}$
+OIToFER	Start LR	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	MAE	1.330^{03}	1.290^{05}	1.329^{04}	1.288^{05}	1.279^{06}	1.308^{04}	1.251^{06}	1.245^{07}
	Spearman R	$.220^{11}$	$.305^{12}$	$.226^{10}$	$.313^{09}$	$.326^{11}$	$.272^{09}$	$.369^{11}$	$.374^{14}$
					Valence				
Baseline	Start LR	$1e-05$	0.001	0.001	0.001	0.001	0.001	0.001	0.0001
	MAE	1.363^{08}	1.312^{08}	1.385^{05}	1.290^{07}	1.276^{06}	1.336^{11}	1.107^{12}	1.015^{07}
	Spearman R	$.383^{10}$	$.434^{09}$	$.394^{13}$	$.477^{07}$	$.487^{08}$	$.440^{14}$	$.618^{09}$	$.685^{05}$
+OIToFER	Start LR	0.01	0.001	0.01	0.01	0.01	0.01	0.001	0.01
	MAE	1.324^{06}	1.237^{06}	1.324^{07}	1.223^{06}	1.213^{06}	1.274^{06}	1.077^{08}	1.005^{07}
	Spearman R	$.438^{07}$	$.522^{06}$	$.444^{07}$	$.534^{06}$	$.541^{07}$	$.483^{09}$	$.649^{07}$	$.693^{05}$

Table 13: Percentage of times, with respect to the full public dataset, the annotated emotion was ranked N in the model predictions.

	0	1	2	3	4	5	6	7
alexnet	22.5	15.7	13.5	11.8	11.0	10.1	8.9	6.5
vgg16	41.8	20.3	13.2	9.0	6.0	4.7	3.1	2.0
resnet18	31.3	19.2	13.3	10.6	8.4	6.9	5.8	4.5
resnet50	38.1	19.4	13.5	9.5	7.0	5.5	4.0	2.9
resnet101	38.4	20.2	13.2	9.4	6.8	5.1	4.0	3.0
densenet161	33.3	18.7	13.4	10.4	8.4	6.5	5.2	4.0

Table 14: Percentage of times, with respect to the full public dataset, the distance between the annotated and predicted emotions was equal to N in the model predictions.

	0	1	2	3	4
alexnet	22.5	19.7	19.1	23.7	15.0
vgg16	41.8	21.4	15.2	14.0	7.7
resnet18	31.3	26.1	17.0	17.0	8.6
resnet50	38.1	21.7	16.2	15.7	8.3
resnet101	38.4	23.1	16.9	14.1	7.5
densenet161	33.3	22.8	18.4	16.8	8.7

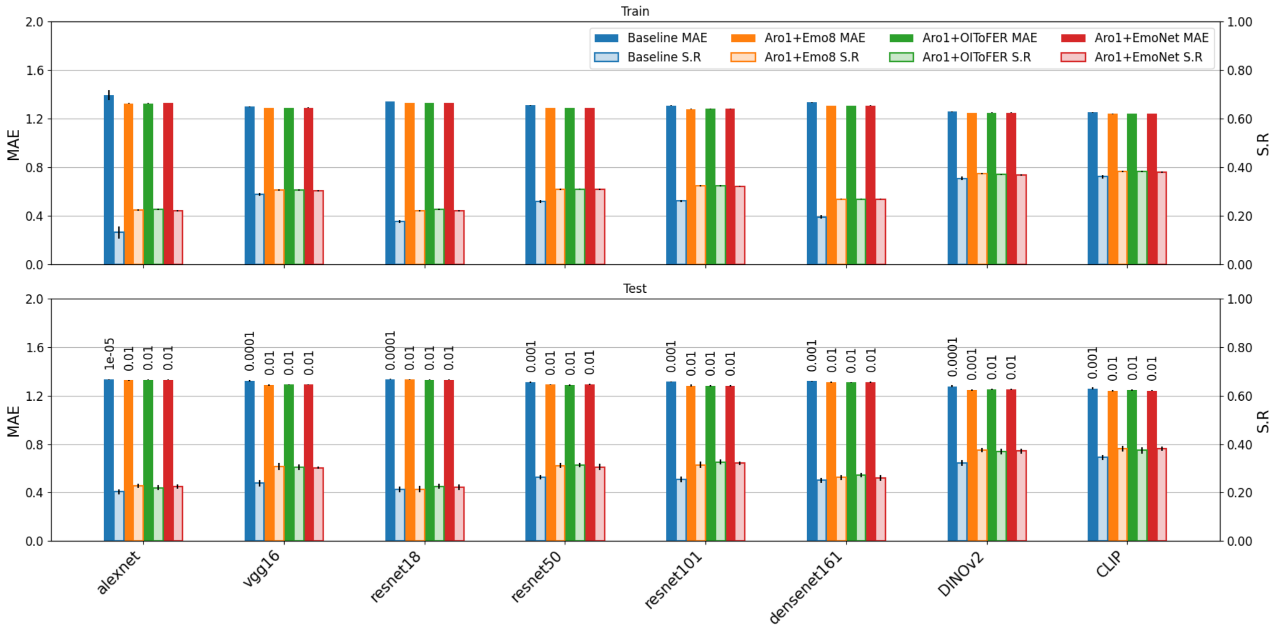

Figure 24: Arousal regression results for extensions beyond the baseline by applying late fusion with precomputed Emo8 predictions of the same architecture (Emo8), Facial Emotion Recognition predictions (OIToFER) and EmoNet predictions (EmoNet). For all models, precomputed predictions on the dataset (Aro1) are concatenated and sent through a linear layer. Metrics are: Mean Average Error (MAE) and Spearman R (S.R). The starting learning rate corresponding to each model is displayed above the training bars.

e.g., to remove data with such issues. We also confirm the licenses provided with the data and code associated with this work: an MIT license for all code; a CC BY-NC-SA 4.0 license for the dataset (concretely, the list of URLs and the annotations).

In particular, and as clearly and explicitly stated on our repository (under "Legal Compliance and Privacy"), we invite any rightful copyright holders or persons depicted in any of the images that do not want their work/likeness to be used within the context of this dataset to contact us, so that we can remove that specific material from the dataset.

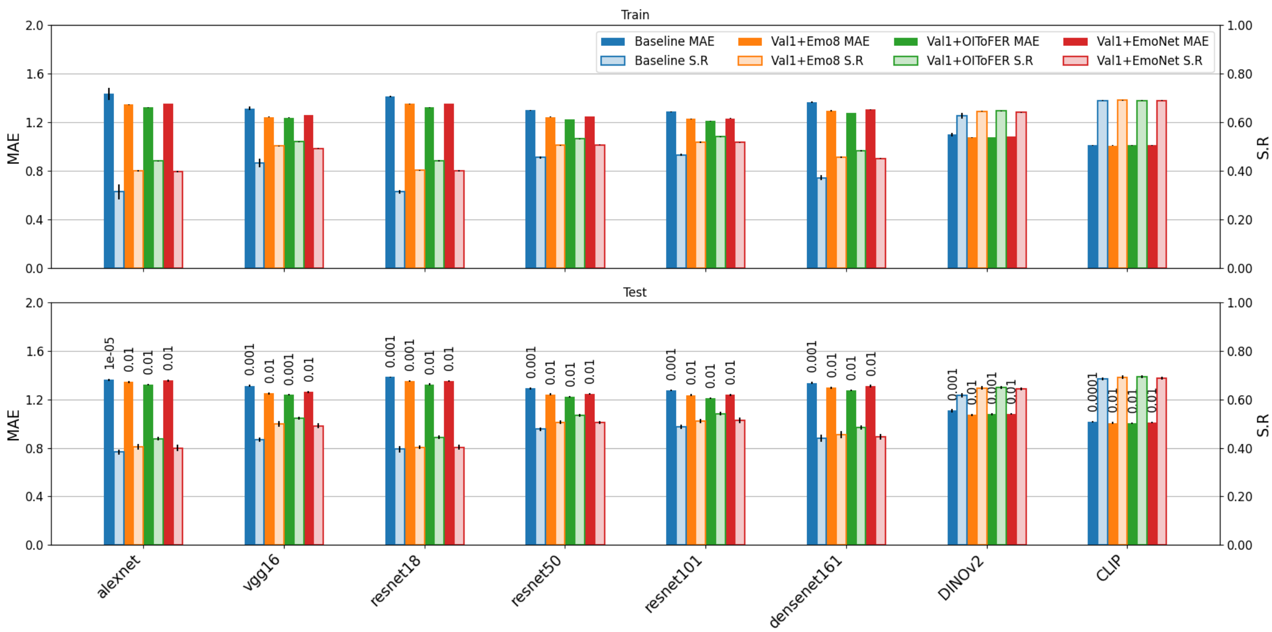

Figure 25: Valence regression results for extensions beyond the baseline by applying late fusion with precomputed Emo8 predictions of the same architecture (Emo8), Facial Emotion Recognition predictions (OIToFER) and EmoNet predictions (EmoNet). For all models, precomputed predictions on the dataset (Val1) are concatenated and sent through a linear layer. Metrics are: Mean Average Error (MAE) and Spearman R (S.R). The starting learning rate corresponding to each model is displayed above the training bars.