BENet: A lightweight bottom-up framework for context-aware emotion recognition

Tristan Cladière, Olivier Alata, Christophe Ducottet, Hubert Konik, Anne Claire Legrand

BENet: A lightweight bottom-up framework for context-aware emotion recognition

Tristan Cladière, Olivier Alata, Christophe Ducottet, Hubert Konik, and Anne-Claire Legrand

Laboratoire Hubert Curien, Université Jean Monnet, Saint-Etienne, France
tristan.cladiere@univ-st-etienne.fr
https://laboratoirehubertcurien.univ-st-etienne.fr/en/index.html

Abstract. Emotion recognition from images is a challenging task. The latest and most common approach to solve this problem is to fuse information from different contexts, such as person-centric features, scene features, object features, interactions features and so on. This requires specialized pre-trained models, and multiple pre-processing steps, resulting in long and complex frameworks, not always practicable in real time scenario with limited resources. Moreover, these methods do not deal with person detection, and treat each subject sequentially, which is even slower for scenes with many people. Therefore, we propose a new approach, based on a single end-to-end trainable architecture that can both detect and process all subjects simultaneously by creating emotion maps. We also introduce a new multitask training protocol which enhances the model predictions. Finally, we present a new baseline for emotion recognition on EMOTIC dataset, which considers the detection of the agents. Our code is available at https://github.com/TristanCladiere/BENet.git.

Keywords: Emotion recognition · Detection · Bottom-Up · Multitask · Deep learning

1 Introduction

Understanding emotions is a difficult yet essential task in our daily life. They can be defined as discrete categories or as coordinates in a continuous space of affect dimensions [4]. For the discrete categories, Ekman and Friesen [5] defined six basic ones: anger, disgust, fear, happiness, sadness, and surprise. Later, contempt was added to the list [12]. Concerning the continuous space, valence, arousal, and dominance form the commonly used three-dimensional frame [13].

Regarding non-verbal cues, facial expression is one of the most important signal to convey emotional states and intentions [11]. However, the context is also essential in some cases, because it can be misleading to infer emotions using only the face [1]. For images, the context can include many things, and the recent authors have built different deep learning architectures to process it. Lee *et al.* [10] proposed an attention mechanism to extract features from everything else than the face. Zhang *et al.* [20] inferred emotions with a Graph Convolutional Network, using the features generated by a Region Proposal Network

as nodes. Kosti *et al.* [9] created a two-stream Convolutional Neural Network which extracts body and scene features, and fuses them. Similarly, Bendjoudi *et al.* [2] used a two-stream network, and studied the synergy between continuous and categorical loss functions. Mittal *et al.* [14] combined agent, scene, and depth features with multiplicative fusion. Here, the agent features are computed from facial landmarks and body pose, both obtained with off-the-shelves models as pre-processing steps. Instead of depth features, Hoang *et al.* [6] designed a reasoning stream that explores relationships between the main subject and the adjacent objects in the scene, using an existent and pre-trained objects detector. Wang *et al.* [17] introduced the tubal transformer, a shared features representation space that facilitates the interactions among the face, body, and context features. Yang *et al.* [19] developed an adaptive relevance fusion module for learning the shared representations among multiple contexts, some of whom depend on external models.

Although the above approaches provide good results, they are all composed of multiple streams that use different kinds of inputs and are processed sequentially. Therefore, both training and inference become slower and more complicated, especially when memory and computational power are restricted. Moreover, none of them consider the detection task, which not only makes their method not directly usable for real world applications, but also gives emotion scores that are not representative of the whole process. Indeed, many pre-processing steps depend on the bounding boxes provided by the annotations, but in real scenario they would be obtained with a person detector, which may be inaccurate or even miss subjects. Based on these observations, we propose a totally different approach, designed to be later embedded in a robot for real-time uses, and leading to the three main contributions presented in this work. Firstly, we built a model that simultaneously assesses the emotions of all subjects present in an image. It is end-to-end trainable, relies on a single shared backbone, takes directly the raw image as input, and does not require specialized pre-trained modules. Secondly, we made the model multitask capable, which means that the same architecture can also predict the bounding boxes of the subjects by itself, estimate the emotions of a particular agent using only its person-centric features, and give all the emotions in the image using only background features. Thirdly, we share a new baseline that evaluates simultaneously the detection and the recognition parts of our model.

2 Proposed Method

In this section, the components of our multitask approach will be detailed. Each head of the architecture is dedicated to a specific task. First, the bottom-up head is introduced. It allows to estimate the emotions of multiple people simultaneously, unlike the usual methods which treat them sequentially. Next, the detection head is presented. Combined with the bottom-up one, these blocks

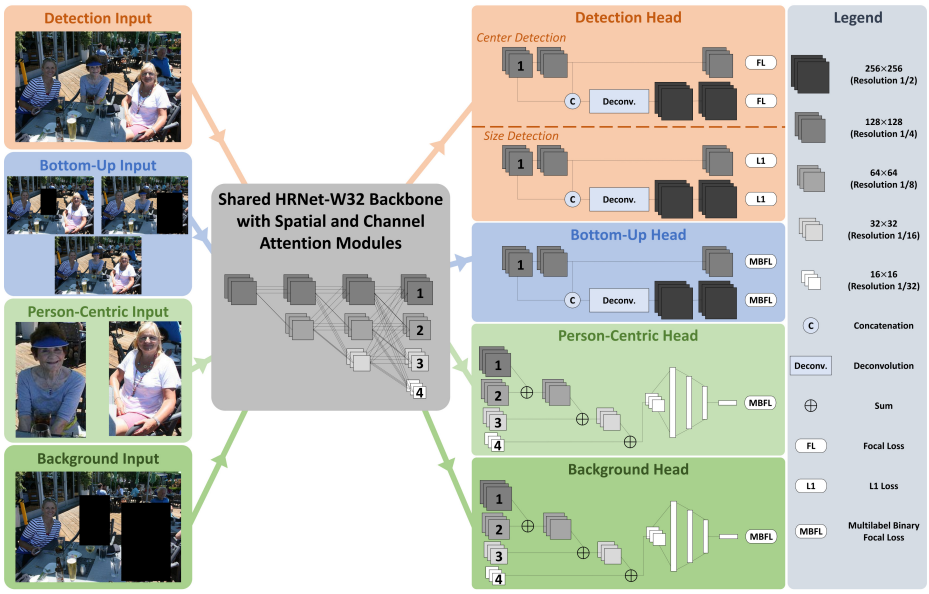

Fig. 1. Overview of our architecture

make the model fully autonomous, since it becomes independent of the annotated bounding boxes. After, the person-centric head is described. It is specifically used to predict the emotions of a single subject given as input. Here we only rely on the subject's features, without processing neither other people nor the background information. On the contrary, the background head is finally shown. It makes a global prediction using only the scene features, i.e. everything except the annotated subjects. An overview of our architecture is given in Fig. 1.

2.1 Bottom-Up approach

The authors cited in section 1 use methods considered as top-down approaches: they first have to detect the subject (or use the ground truth) before inferring his emotions. With these approaches, each subject is treated sequentially and independently, which is slow and redundant for images with multiple people. Therefore, we proposed a new way to handle this problem, that can be considered as a bottom-up solution. Inspired by [3], a bottom-up head is used to produce E discrete emotion maps directly from the raw image, as illustrated in Fig. 2. The value of E depends on the number of discrete emotion categories used in the considered database. On these maps, only the value of the pixel at the center of each subject's bounding box is imposed: 1 if the emotion is present, 0 otherwise. For all the other pixels, the model is free to output anything that helps it in making its predictions. However, at test time, the bounding boxes coordinates are necessary to extract and attribute the predictions. They can be given either by the ground truth or by a person detector.

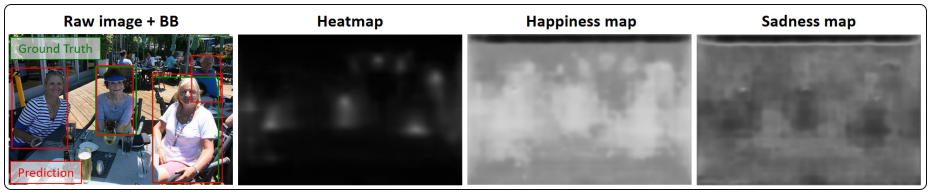

Fig. 2. Example of the heatmap produced by the detection head, and two emotion maps given by the bottom-up head. Normalisation is between 0 (black, emotion is absent) and 1 (white, emotion is present). The predicted and annotated bounding boxes are also added to the raw image.

2.2 Detection head

To become fully autonomous, our model needs to be able to automatically extract its predictions from the emotion maps. Thus, similar to [21], a bottom-up detection head is added and trained to predict the center of the bounding boxes by creating a heatmap (see Fig. 2), and to regress their dimensions. During the inference, these centers coordinates can be extracted and used to retrieve the emotions of the subjects, and also the predicted dimensions of the bounding boxes. Therefore, the model both detects all the subjects in the raw image and estimates their emotions in a single forward pass.

At this point, the whole architecture is still end-to-end trainable, and the two tasks can be jointly trained. Moreover, the framework is composed of a unique backbone, which means that the heads share common features. It benefits the bottom-up approach since the model becomes better at predicting emotions at the correct coordinates of the emotion maps.

2.3 Person-centric and background features

The bottom-up approach introduced in section 2.1 uses as input the raw image to produce its emotions maps. The architecture has a global view of the scene, that contains both person-centric and background features. Depending on the situation, one of these features may be prevailing over the other one, and the model should still be able to perform well with this single source of information. This is why two heads were added to the framework, one that will be specialised in person-centric features, the other in background ones.

To extract features from the main subject, some authors used pre-trained deep-learning architectures to detect his face, his facial landmarks, and to estimate his posture from the portion of the input image corresponding to his bounding box [10, 14, 6, 17, 19]. These outputs served as intermediate information that further helps to infer the emotions of the subject, but they are also dependent on external resources. In our case, a simpler method is used: a classification block inspired by [16] is added to the model, and is referred as the

person-centric head. The combination of our backbone and this specific head is very similar to the work of [16]. The flexibility given by such architecture seems profitable for processing in-the-wild images, including close range faces and far range silhouettes, mainly due to its multi-resolutions design.

For background information, the corresponding head has the same design as the person-centric one, but both the input and the training objective are different. Following [19], all the annotated subjects in the raw image are masked, forcing the model to rely on other sources of features (see Fig. 1). However, rather than predicting the emotions of a single person, the architecture has to estimate all the emotions present in the image, i.e. each emotion that is labelled for at least one subject. Given N people annotated for E emotions on a single image, a one-hot-encoded matrix of dimension $N \times E$ is therefore created, and the maximum along the N axis is taken, resulting in a vector of shape $1 \times E$.

Here again the whole architecture is end-to-end trainable, and all the tasks are jointly trained. In this configuration, the shared backbone learns to extract rich common features, so that each task benefits from the others.

3 Framework details

In this section, the multitask architecture is first presented, then the data used and their processing to jointly train all heads are detailed, and finally the loss functions are explained.

3.1 Network architecture

Our Bottom-up Emotions Network (BENet) uses HRNet-W32 [16] as backbone. It contains four stages with four parallel convolution streams. The resolutions are 1/4, 1/8, 1/16, and 1/32, while the widths (numbers of channels) of the convolutions are C, $2C$, $4C$, and $8C$ ($C = 32$). We also integrated spatial attention modules and channel attention modules, inspired by [18].

For the detection and the bottom-up heads, a structure very similar to [3] has been implemented. The main point is to use a deconvolution module on top of the highest resolution feature maps in HRNet, increasing the resolution from 1/4 to 1/2. Such process mainly helps to detect smaller people in the image, since bottom-up approaches must deal with subjects of very different scales. At the end, the model is trained to output its predictions at two resolutions, 1/4 and 1/2.

For the two classification heads, the design of [16] is used, but the bottleneck expansion is reduced by a factor 4. Indeed, we do not need to have too many channels, considering that there are far fewer emotions in emotion recognition than classes in image classification. It also helps to reduce the global number of parameters of the architecture, which is profitable considering that there is not a lot of available data.

3.2 Databases

EMOTIC database [8] contains 23,571 images of 34,320 annotated people in unconstrained environments. Each image has at least one subject, which is annotated with a bounding box, 26 discrete categories, and 3 continuous dimensions of emotions. The subject can be assigned multiple labels. The standard partition of the dataset is 70% for the training set, 10% for the validation one, and the remaining 20% are used for testing.

HECO database [19] regroups 9,385 images and 19,781 annotated people, with rich context information and various agent interaction behaviours. The annotations include 8 discrete categories and 3 continuous dimensions, but also the novel *Self-assurance* (Sa) and *Catharsis* (Ca) labels, which describe the degree of interaction between subjects and the degree of adaptation to the context. Unfortunately, the authors do not provide any partition of their dataset, which makes the evaluations difficult to compare. Thus, we only used HECO as extra data for training.

3.3 Data processing

The data augmentation is divided into two parts. The first part is designed to randomly apply a specific pre-transformation on each image of the training batch. Depending on the pre-transformations drawn, the images will be dispatched at the end of the backbone, and fed to the corresponding heads. Thus, each image is designed to train a specific task. These pre-transformations are named *ExtractSubject*, *MaskAllSubjects*, and *RandomMaskSubjects*. They will be briefly explained, and examples of the inputs are shown in Fig. 1.

ExtractSubject uses the ground truth to crop the image around the bounding box of a given subject. It will be used to train the model to extract person-centric features. This pre-transformation can only be drawn if the bounding box of the selected subject does not contain other people.

MaskAllSubjects uses the ground truth to mask all the annotated subjects in the image. With such images, the background head will have to extract features from everything except the people. To ensure that there is still enough information left for the model to learn useful features, this pre-transformation can only be applied on images in which the sum of the areas of the bounding boxes do not exceed 40% of the total area of the input.

RandomMaskSubjects is the transformation used to train the bottom-up head. If there are multiple annotated subjects in the image, we will randomly mask them, but always make sure to keep at least one. The idea is to augment and diversify the combinations of emotions presented to the model.

The last option is to keep the raw image. In this case, it will be used to train the detection head. Given a batch size B, each image will be pre-processed by picking one of the above pre-transformations, with probabilities of (namely) 0.25, 0.25, 0.25, and 0.25. This equiprobability is chosen as the default experiment.

The second part of the data augmentation consists of adding random gaussian noise, random blur, random colour jittering, random horizontal flip, and random perspective transformations to the pre-transformed images.

3.4 Loss functions

Since the architecture is multitask, and also multi-resolutions for the bottom-up and detection heads, a loss function must be defined for each task at each resolution.

For the detection, following [21], the focal loss is used to train the generation of heatmaps, and the L1 loss for the regression of the bounding boxes dimensions. The focal loss is defined as follows:

$$L_{center} = \frac{-1}{N} \sum_{xy} \begin{cases} \left(1 - \hat{Y}_{xy}\right)^{\alpha} \log\left(\hat{Y}_{xy}\right) & \text{if } Y_{xy} = 1 \\ \left(1 - Y_{xy}\right)^{\beta} \left(\hat{Y}_{xy}\right)^{\alpha} \log\left(1 - \hat{Y}_{xy}\right) & \text{otherwise} \end{cases} \tag{1}$$

where $\alpha = 2$ and $\beta = 4$ are hyper-parameters, N is the number of subjects, $\hat{Y}_{xy}$ and Y_{xy} are namely the prediction and the ground truth at pixel (x, y). The normalization by N is chosen in order to normalize all positive focal loss instances to 1. For the size loss, given a subject k whose bounding box coordinates are $\left(x_1^k, y_1^k, x_2^k, y_2^k\right)$, his center point lies at $p_k = \left(\frac{x_1^k + x_2^k}{2}, \frac{y_1^k + y_2^k}{2}\right)$, and his dimensions are $s_k = \left(x_2^k - x_1^k, y_2^k - y_1^k\right)$. Therefore, the size loss is defined as follows:

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{p_k} - s_k \right| \tag{2}$$

where $\hat{S} \in \mathbf{R}^{w \times h \times 2}$ are the width and height prediction maps of size $w \times h$ for a given resolution. Hence, the detection loss at this resolution is:

$$L_{det} = \lambda_{center} L_{center} + \lambda_{size} L_{size} \tag{3}$$

λ_{center} is set to 1, and λ_{size} to 0.1. Since the model gives predictions at resolutions 1/4 and 1/2, the overall detection loss is therefore:

$$L_{DET} = L_{det-1/4} + L_{det-1/2} \tag{4}$$

For the emotion recognition task, which concerns person-centric, background, and bottom-up heads, a loss similar to [2] is used. It is a multi-label and binary focal loss, which gives better results while dealing with unbalanced data. It is defined as follows:

$$L_{cat-emo} = \frac{-1}{N} \sum_{N} \sum_{i=1}^{E} Y_i \left(1 - \hat{Y}_i\right)^{\alpha} \log\left(\hat{Y}_i\right) + \left(1 - Y_i\right) \left(\hat{Y}_i\right)^{\alpha} \log\left(1 - \hat{Y}_i\right) \tag{5}$$

where N is the number of subjects in the image, E is the number of emotions, $\hat{Y}_i$ and Y_i are namely the prediction and the ground truth for the $i-th$ emotion,

and $\alpha = 2$ is a hyper-parameter. For the person-centric and background heads, we have $N = 1$ and directly a predicted array of size $1 \times E$. However, the bottom-up head outputs emotions maps of shape $E \times w \times h$, where w and h depend on the resolution considered (either 1/4 or 1/2). To extract the matrix of $N \times E$ predictions, the N bounding boxes centers given by the ground truth are used. Therefore, the global categorical emotions loss is:

$$L_{CAT} = L_{cat-emo}^{person-centric} + L_{cat-emo}^{background} + L_{cat-emo-1/4}^{bottom-up} + L_{cat-emo-1/2}^{bottom-up} \quad (6)$$

Finally, the total loss is defined as:

$$L_{TOT} = L_{DET} + L_{CAT} \quad (7)$$

4 Experiments and results

4.1 Training details

The method is built with the Pytorch toolbox [15]. The models are trained during 250 epochs, using the EMOTIC database. We kept the standard train, validation and test sets provided. When extra data are used, it means that HECO has been merged with the training set of EMOTIC. We use the Adam optimizer [7] with an initial learning rate of $1e^{-3}$. The *best model* is defined as the one with the lowest total validation loss. The *final model* is the one obtained at the end of the 250 epochs.

4.2 Evaluation metrics

Since our method includes the detection of the subjects, we propose two evaluation metrics. The first one is the standard Average Precision score (AP) for all emotion categories, that can also be averaged (mAP). The predictions are extracted using the ground truth bounding boxes. Considering that this case is independent of the model's detection head, and assuming that all other methods in the literature also use the annotations, this metric can be considered the most appropriate for comparison with the state-of-the-art.

Nevertheless, in real world applications, such annotations are not provided, so we must rely on a detector whose performance can have a significant impact on emotions scores. To know if a detection is successful or not, the commonly used method is to compute the Intersection over Union (IoU) between the detected bounding box and the ground truth, which indicates how much they are superimposed, and count as True Positive the values superior to a given threshold. The IoU values are between 0 and 1, 1 being a perfect detection. In the COCO API, the final detection score is the mean of the AP values obtained with 10 thresholds from 0.5 to 0.95. Therefore, we use this API to evaluate not only our person detector on EMOTIC, but also the performances of the whole framework during autonomous inferences. To do so, the bounding boxes successfully predicted for a given IoU threshold are used to extract the emotion

predictions from the emotions maps. Otherwise, when the detection fails for an annotated subject, his predicted emotions are treated as False Negative, i.e. a vector of zeros of shape $1 \times E$ is created. With this new evaluation protocol, the scores obtained are more representative to what could be achieve during real inferences.

4.3 Analysis of the results

To evaluate the proposed method, both *best model* and *final model* were tested. It appears that 250 epochs are enough to witness over-fitting through the validation loss. However, in some cases the *final model* still give better results. Since only the best results are reported in this paper, we specify if they come from the *final model* by underlining the value in the Tables 1, 2, 3, and 4.

In Table 1, the results for emotion recognition following different training strategies are summarized. These mAP are obtained without considering the detection part, because the annotations were used to extract the emotion predictions, instead of the integrated person detector. As we can see, when the bottom-up head (BU) is trained alone, it does not perform well. However, when the detection head (Det) is added and jointly trained, the score are improved by a good margin. There is also a little improvement by adding person-centric (PC) and background (BG) heads. The use of additional data from HECO (ED) helps to improve even more the performance of the model.

Table 1. Ablation Experiments on EMOTIC Dataset for emotion recognition. Underlined values come from the *final model* instead of the *best model*.

Heads	BU	BU+Det	BU+Det+PC	BU+Det+PC+BG	BU+Det+PC+BG+ED
mAP	23.10	27.22	27.49	27.73	**28.75**

Regarding the detection task, the best score is obtained when all the heads are trained together with extra data, as it is illustrated in Table 2. Yet, the main drawback with EMOTIC database is that it is not fully annotated. Indeed, there are many images with several people but where only a few of them are labeled. Thus, the detector tends to produce many False Positive (as illustrated in Fig. 2), that are penalized during the training and may confuse the model, and also reduce the precision during the evaluations.

Table 2. Ablation Experiments on EMOTIC Dataset for person detection, using the COCO API. Underlined values come from the *final model* instead of the *best model*.

Heads	BU+Det	BU+Det+PC	BU+Det+PC+BG	BU+Det+PC+BG+ED
AP	49.66	49.16	51.49	**51.71**

The scores presented in Table 3 correspond to the new evaluation protocol which considers the tasks of detection and emotion recognition together, introduced in section 4.2. As expected, the results are worse than those obtained with the ground truth, but surprisingly the model using all heads and giving the best results in both detection and emotion recognition is no longer the best with this new metric. This can be explained by the fact that the latter detects more subjects, even people whose emotions are particularly difficult to assess, for example those who are partially occluded or quite distant in the background. In these situations, the model is more likely to be wrong and produce more False Positives and less True Positives, which decreases its precision. However, using external data still leads to better results.

Table 3. mAP scores for emotion recognition on EMOTIC Dataset, depending on the detector predictions for thresholds from 0.50 to 0.95. (1): BU+Det ; (2): BU+Det+PC ; (3): BU+Det+PC+BG ; (4): BU+Det+PC+BG+ED. Underlined values in the "Average" column indicate that the scores in the whole row come from the *final model* instead of the *best model*.

Det. thr.	*0.50*	*0.55*	*0.60*	*0.65*	*0.70*	*0.75*	*0.80*	*0.85*	*0.90*	*0.95*	**Avg.**
(1)	25.73	25.38	25.04	24.68	24.37	24.00	23.62	22.83	21.71	19.37	<u>23.67</u>
(2)	26.19	25.85	25.51	25.32	25.13	24.81	24.20	23.24	21.91	19.37	<u>24.15</u>
(3)	26.01	25.86	25.57	25.24	24.97	24.58	23.99	23.29	22.00	19.48	<u>24.10</u>
(4)	**26.96**	**26.66**	**26.38**	**26.07**	**25.82**	**25.28**	**24.61**	**23.71**	**22.21**	**19.57**	**24.73**

Even if our framework and our objectives are quite different from the other authors, we finally compared our scores with those of the state-of-the-art in Table 4. The baseline on EMOTIC, provided by [9], is outperformed. Our model is multitask, but possible ways to fuse the predictions of the different heads have not been explored yet. Indeed, the person-centric and background heads are only used to help the model during its training, but not while inferring. Nevertheless, we still tried to average all the outputs, which requires to preprocess the raw image for the person-centric and background heads. It finally appears that the mean between the bottom-up and the person-centric outputs gives the best refined prediction, which means that 2 streams are used here. However, the most recent methods are still quite ahead, due to their rich and complex framework, and a well-made fusion.

5 Conclusion and future work

In this paper, we present a innovative method to simultaneously detect people on an image, and predict their categorical emotions. Since all subjects are treated simultaneously, our approach can be referred as a bottom-up method, and we are the first ones to explore this path. We also introduce a multitask training strategy to improve the performance of the model. Finally, we propose a new

Table 4. State-of-the-art on EMOTIC Dataset. NERR: Number of External Resources Required (off-the-shelves models).

Authors	[10]	[9]	[2]	[20]	[17]	[6]	[14]	[19]	**Ours**
nb. of streams	2	2	2	2	3	6	4	7	2
fusion module	✓	✓	✓	✓	✓	✓	✓	✓	✗
NERR	1	0	0	1	1	2	3	3	0
mAP	20.84	27.38	28.33	28.42	30.17	35.16	35.48	**37.73**	29.30

evaluation protocol that consider both detection and emotion recognition task, in order to better represent the true capabilities of the method during real life inferences. As part of future work, we would also treat continuous emotions (valence, arousal, and dominance), and explore fusion methods to combine the bottom-up predictions with the person-centric and background ones, already available in our multitask model.

Acknowledgements This work was sponsored by a public grant overseen by the French National Research Agency as part of project muDialBot (ANR-20-CE33-0008-01).

References

1. Barrett, L.F., Mesquita, B., Gendron, M.: Context in emotion perception. Current directions in psychological science **20**(5), 286–290 (2011)
2. Bendjoudi, I., Vanderhaegen, F., Hamad, D., Dornaika, F.: Multi-label, multi-task CNN approach for context-based emotion recognition. Information Fusion **76**, 422–428 (2021). https://doi.org/10.1016/j.inffus.2020.11.007
3. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020)
4. Ekman, P., Friesen, W.V.: Head and body cues in the judgement of emotion: A reformulation. Perceptual and motor skills **24**(3), 711–724 (1967)
5. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology **17**, 124–129 (1971). https://doi.org/10.1037/h0030377, place: US Publisher: American Psychological Association
6. Hoang, M.H., Kim, S.H., Yang, H.J., Lee, G.S.: Context-aware emotion recognition based on visual relationship detection. IEEE Access **9**, 90465–90474 (2021). https://doi.org/10.1109/ACCESS.2021.3091169, conference Name: IEEE Access
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: EMOTIC: Emotions in context dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 61–69 (2017)

9. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using EMOTIC dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(11), 2755–2766 (2020). https://doi.org/10.1109/TPAMI.2019.2916866, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence
10. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10143–10152 (2019)
11. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing **13**(3), 1195–1215 (2022). https://doi.org/10.1109/TAFFC.2020.2981446, conference Name: IEEE Transactions on Affective Computing
12. Matsumoto, D.: More evidence for the universality of a contempt expression. Motivation and Emotion **16**(4), 363–368 (1992). https://doi.org/10.1007/BF00992972
13. Mehrabian, A.: Framework for a comprehensive description and measurement of emotional states. Genetic, Social, and General Psychology Monographs **121**, 339–361 (1995), place: US Publisher: Heldref Publications
14. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: EmotiCon: Context-aware multimodal emotion recognition using frege's principle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14234–14243 (2020)
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019)
16. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3349–3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence
17. Wang, Z., Lao, L., Zhang, X., Li, Y., Zhang, T., Cui, Z.: Context-dependent emotion recognition. SSRN Electronic Journal (2022). https://doi.org/10.2139/ssrn.4118383
18. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, vol. 11211, pp. 3–19. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01234-2_1, series Title: Lecture Notes in Computer Science
19. Yang, D., Huang, S., Wang, S., Liu, Y., Zhai, P., Su, L., Li, M., Zhang, L.: Emotion recognition for multiple context awareness. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 144–162. Lecture Notes in Computer Science, Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-19836-6_9
20. Zhang, M., Liang, Y., Ma, H.: Context-aware affective graph reasoning for emotion recognition. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 151–156 (2019). https://doi.org/10.1109/ICME.2019.00034, ISSN: 1945-788X
21. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)