Knowledge Amalgamation for Single-Shot Context-Aware Emotion Recognition

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

Université Jean Monnet Saint-Etienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

{*tristan.cladiere, olivier.alata*}*@univ-st-etienne.fr*

Keywords: Emotion Recognition, People Detection, Context, Amalgamation, Multiple Teachers.

Abstract: Fine-grained emotion recognition using the whole context inside images is a challenging task. Usually, the approaches to solve this problem analyze the scene from different aspects, for example people, place, object or interactions, and make a final prediction that takes all this information into account. Despite giving promising results, this requires specialized pre-trained models, and multiple pre-processing steps, which inevitably results in long and complex frameworks. To obtain a more practicable solution that would work in real time scenario with limited resources, we propose a method inspired by the amalgamation process to incorporate specialized knowledge from multiple teachers inside a student composed of a single architecture. Moreover, the student is not only capable of treating all subjects simultaneously by creating emotion maps, but also to detect the subjects in a bottom-up manner. We also compare our approach with the traditional method of fine-tuning pre-trained models, and show its superiority on two databases used in the context-aware emotion recognition field.

1 INTRODUCTION

Even as human beings, it is not always trivial to assess someone's emotions. In real world situations, useful visual cues for inferring emotions not only include facial expressions (Li and Deng, 2020), but also diverse information relative to the context, such as human appearance and pose, objects interacting with the subject and more generally the global context of the scene (Barrett et al., 2011). This problem has been recently addressed as Context-Aware Emotion Recognition (CAER).

Recent architectures designed for CAER address the challenge posed by the diverse and distinct nature of contextual elements. These architectures typically adopt a multiple-stream network, made of various encoding modules to extract specific features from the input image, as illustrated by Figure 1. The subsequent fusion module and classification head are then employed to predict the emotion of the main subject. For instance, in (Lee et al., 2019), the face-centric stream is supplied with a cropped image of the face, while the global context stream is fed with the entire image, excluding the face which has been intentionally concealed. Similarly, (Kosti et al., 2019) and (Bendjoudi et al., 2021) both proposed a two-stream architecture to extract person-related and scene-related features. In (Zhang et al., 2019), one of the streams is made of a graph convolutional network that uses the features generated by a region proposal network as nodes. This multi-stream methodology often relies on off-the-shelf modules for pre-processing, as illustrated by (Mittal et al., 2020), who proposed a three-stream architecture. In their approach, the person-centric stream is made of two sub-streams utilizing OpenPose and OpenFace models, while the inter-agent stream incorporates the Megadepth model to extract a depth map. The authors in (Wang et al., 2022) introduced the tubal transformer, a shared features representation space that facilitates the interactions among the face, body, and context features, but this requires to use Retinaface first. The reasoning stream in (Hoang et al., 2021) that explores relationships between the main subject and the adjacent objects in the scene relies on FasterRCNN to generate their inputs. This is also the case for one of the seven streams in (Yang et al., 2022) that are merged using an adaptive relevance fusion module. A brief summary of these approaches is provided in Table 1, along with an overview of their performance on EMOTIC (Kosti et al., 2017; Kosti et al., 2019), a widely used dataset in CAER.

Cladière, T., Alata, O., Ducottet, C., Konik, H. and Legrand, A.-C.
Knowledge Amalgamation for Single-Shot Context-Aware Emotion Recognition.
DOI: 10.5220/0013169800003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 410-419
ISBN: 978-989-758-728-3; ISSN: 2184-4321

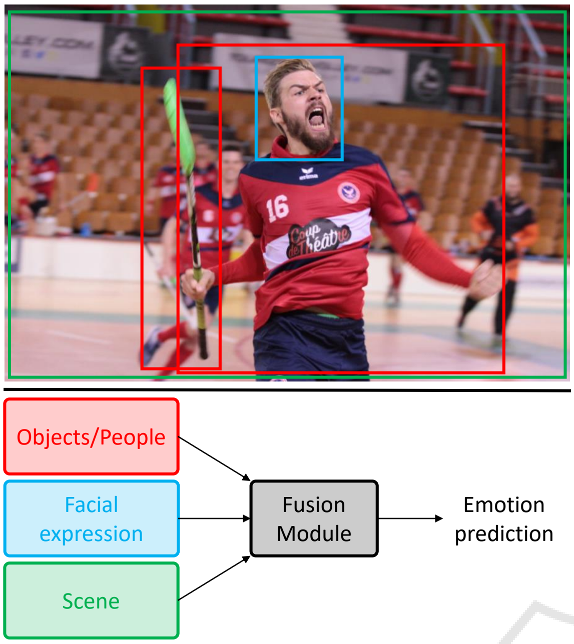

Figure 1: Illustration of the current trend followed by CAER approaches, where multiple specialized streams are used to extract complementary cues, finally combined through fusion modules to obtain a refined prediction.

Online training and implementation of such models present complexity primarily due to the need for seamless integration of all external modules into the overall architecture. An other important point is that these models inherently follow a top-down approach, where each individual in the scene must be detected and processed separately as the primary actor, while other elements within the scene are treated as contextual background. As a result, processing a single image necessitates multiple inferences, equivalent to the number of subjects present in the scene. It is worth noting that the task of individually detecting each actor is not included within the architecture itself but is instead provided by the ground truth of the dataset, following standard single task evaluation benchmarks in CAER.

The development of online approaches suitable for real-time applications with limited resources poses various challenges. Firstly, it needs to address the emotion recognition task in a bottom-up manner, i.e. capable of simultaneously estimating the emotion of all subjects within an image. Secondly, it must seamlessly integrate a single-shot person detection module to process all the actors in the scene simultaneously. Thirdly, it must depend on a constrained number of feature extraction backbones to enable end-to-end training and minimize the required resources for implementation. A promising architecture of this nature was introduced by (Cladière et al., 2023); however, the training procedure remains intricate and suffers from the limited amount of available data in CAER to efficiently train both the person detection and emotion recognition tasks.

Table 1: Mean Average Precision (mAP) obtained on EMOTIC dataset (Kosti et al., 2017; Kosti et al., 2019) by state-of-the-art methods. NERR: Number of External Resources Required (off-the-shelf models).

Authors	**Streams**	**NERR**	**mAP**
Us	1	0	27.10
(Lee et al., 2019)	2	1	20.84
(Kosti et al., 2019)	2	0	27.38
(Bendjoudi et al., 2021)	2	0	28.33
(Zhang et al., 2019)	2	1	28.42
(Wang et al., 2022)	3	1	30.17
(Hoang et al., 2021)	6	2	35.16
(Mittal et al., 2020)	4	3	35.48
(Yang et al., 2022)	7	3	37.73

In this paper, we propose to introduce the technique of knowledge amalgamation for training a single-shot CAER architecture, employing multiple pre-trained teachers specialized in various sub-tasks related to scene context analysis. This not only streamlines the training process through the use of existing off-the-shelf pre-trained architectures but also provides a simple yet effective single stream approach to learn the diverse visual features needed for a comprehensive understanding of the scene, leveraging task-oriented external datasets. To the best of our knowledge, this is the first attempt to use knowledge amalgamation in the CAER context, since it has only been tested on image classification in the literature. Indeed, (Luo et al., 2019) worked on the amalgamation of heterogeneous-architecture teacher models by learning a common feature space, wherein the student model imitates the projected features of the teachers. A comparable approach is developed by (de Carvalho et al., 2022) to limit the catastrophic forgetting problem in a class-incremental learning framework. (Ye et al., 2019) proposed a branching out method to only amalgamate the filtered knowledge from a pool of teachers to the student model. The approach proposed by (Shen et al., 2019a) is to concatenate and then compress features from multiple teachers, thus creating examples that the student model has to reproduce. Finally, (Shen et al., 2019b) introduced a selective learning scheme to select the best teachers among many, and a transfer bridge to align the features of the student and those of the teachers.

Our contributions encompass three key aspects. Firstly, we proposed a bottom-up model capable of simultaneously estimating the emotion of all subjects within an image. Secondly, we introduced a novel

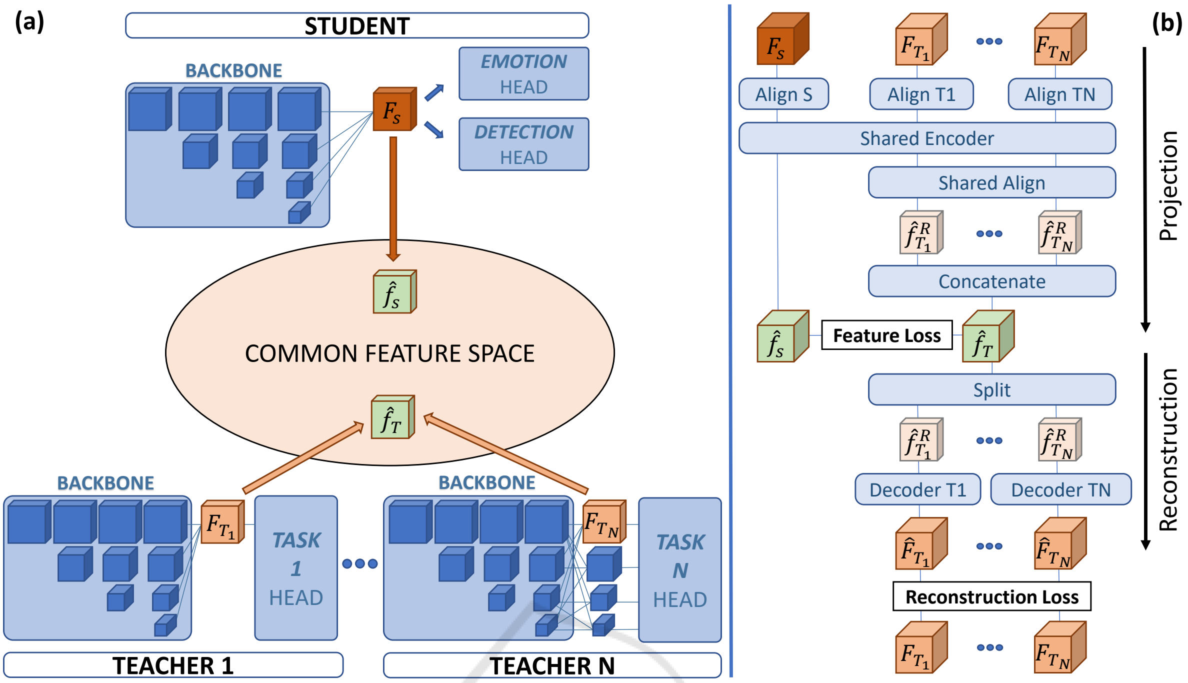

Figure 2: **(a)** Overview of the proposed approach. The student model is composed of a single backbone shared by two bottom-up heads, one for emotion recognition, the other for person detection. Given several teachers that may have different heads (depending on their specialization) but the same backbone, we want the student to reproduce their features. Therefore, both teachers and student features are projected into a common space, where the student learns to reproduce a merged version of the teachers' features. **(b)** Summary of the operations used to project the features into the common feature space and to reconstruct them.

knowledge amalgamation process, leveraging multiple pre-trained teachers, each specialized in specific sub-tasks. Thirdly, we integrated a person detection module to process all actors in a scene in a single shot. We then conducted a performance comparison between our amalgamation approach and a comparable architecture trained with fine-tuning on two CAER databases, evaluating both single-task and multi-task configurations.

2 METHOD

In order to reduce the complexity of our CAER framework, and to explore the capacities of a lightweight solution, we decided to follow the work of (Cladière et al., 2023). Indeed, instead of using multiple specialized streams, we prefer to have a condensed, single-stream architecture that is able to extract rich features, carrying both subject-related and context-related information. To ensure the "quality" of these features, the student model learns to reproduce the examples provided by one or multiple teachers, through the proposed amalgamation process. Then, our bottom-up approach allows to handle all the subjects in an image simultaneously with the generation of emotion maps, where the activated areas indicate the presence of a given emotion. Although poorly explored in CAER research, such bottom-up method offers faster treatment compared to the widely used top-down approaches that require treating each person sequentially. Finally, it is also possible to integrate the person detection task to extract the individual emotion predictions from emotion maps by simply connecting a specialized head to the backbone of the architecture, thus offering a ready-to-use solution without significantly increasing the model size. The Figure 2 summarizes our method.

2.1 Emotion Maps

To generate emotion maps (see Figure 3), we use a bottom-up head inspired by bottom-up architectures used for body pose estimation, such as HigherHRNet (Cheng et al., 2020). Ours is composed of 4 residual blocks, followed by one $2-$dimensional convolution with a kernel of size 3×3 and a padding of size 1, a batch normalization, a ReLu activation function, and a final $2-$dimensional convolution with a kernel of size 1×1 to obtain the desired number E of maps.

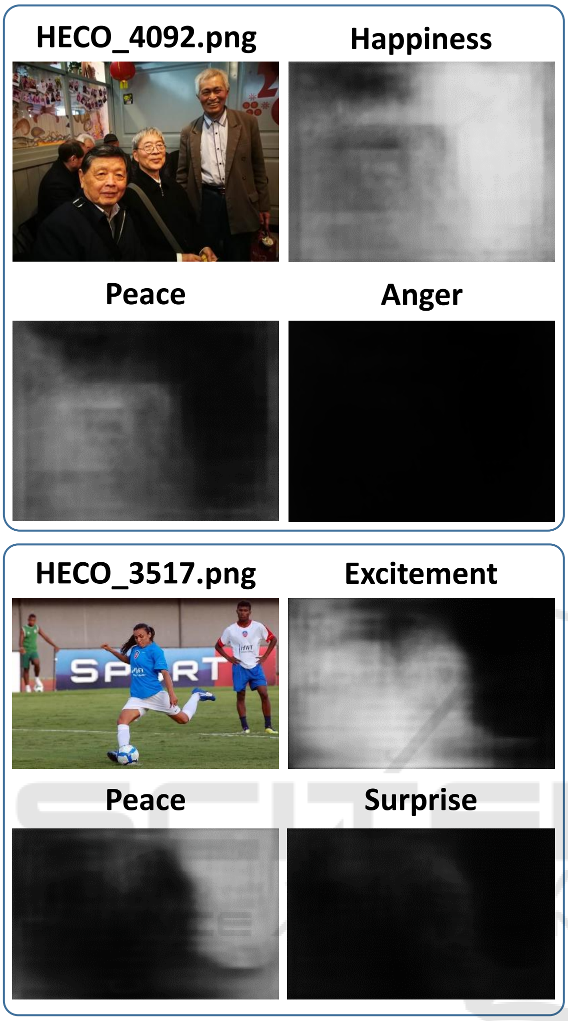

Figure 3: Examples of emotion maps. A softmax function has been applied across the maps for better visualisation.

The value of E depends on the number of discrete emotion categories used in the considered database. On these maps, we want the mean value of the pixels contained inside each subject's bounding box to be equal to 1 if the emotion is present, 0 otherwise. In other words, if a subject is labeled with a certain emotion, the pixels at its position in the corresponding emotion map must be globally activated. However, the value of the other pixels is not imposed, leaving the model free to activate them or not if this can help it make correct predictions. We have found empirically that this produces better results than forcing the background pixels to be 0. Finally, for training and testing our approach, we therefore need to use the annotated bounding boxes to extract the predictions from the emotion maps.

2.2 Detection Head

A limitation of using emotion maps is that we need to rely on the detection of the subjects to extract our predictions. For real inferences on the field, where there is no annotations, we thus need to use a person detector. We made the choice to directly incorporate it inside our architecture by adding another head.

Similar to (Zhou et al., 2019), this bottom-up detection head is trained to predict the center of the bounding boxes by creating a heatmap, and to regress their dimensions by outputting two other maps, one for their height, the other for their width. It is composed of 4 residual blocks, followed by two modules, both made of a $2-$dimensional convolution with a kernel of size 3×3 and a padding of size 1, a batch normalisation, a ReLu activation function, and a final $2-$dimensional convolution with a kernel of size 1×1 to obtain either 1 heatmap or 2 regression maps. Actually, this head is very similar to the one we use for creating emotion maps.

It is during a post-processing step that the local maxima of the heatmaps are determined, thus giving the coordinates of the center of the detected bounding boxes. These are then used to extract predictions on the dimension maps, which finally allows to reconstruct the bounding boxes. Then, the estimated bounding boxes can be used to extract emotion predictions by taking the average value of pixels in the corresponding area of the emotion maps. Therefore, using maps for these two tasks allows them to be processed in parallel, in addition to processing all subjects in a single forward pass, simply by feeding the architecture with the raw image.

2.3 Knowledge Amalgamation

Since the model is fed with the raw image, it has access to different types of information, that can be subject-related, such has the facial expressions, but also context-related, for example the place or the objects. To force the backbone to extract rich features covering these aspects, we decided to use teacher networks to generate examples of what these features should look like.

Given N teacher models, each of which denoted by T_i, we obtain N feature maps F_{T_i}. We want the student network to produce feature maps F_S that imitate those of the teachers. To do so, we transform all these features to a common space, similarly to (Luo et al., 2019) and (de Carvalho et al., 2022). Thus, we first ensure that they all are of equal dimensions by using a $2-$dimensional convolution with a kernel of size 1×1 per network, giving the aligned features noted f_{T_i} and f_S. This aligned features have a fixed number of channels C, where C is a multiple of N. Then, we use a shared extractor composed of 3 residual blocks to project f_{T_i} and f_S into $\hat{f}_{T_i}$ and $\hat{f}_S$, where the number of channels is still C.

However, when $N > 1$, we also add a last $2-$dimensional convolution with a kernel of size 1×1 shared among the teachers, that transforms the $\hat{f}_{T_i}$ to $\hat{f}^R_{T_i}$ with a reduced number of channels equal to $\frac{C}{N}$. This actually allows us to concatenate these $\hat{f}^R_{T_i}$ features to create $\hat{f}_T$, that has the same number of channels than $\hat{f}_S$ but still contains the knowledge of all the teachers. Similar process has also been used in (Shen et al., 2019a). On the other hand, with $N = 1$, we directly have $\hat{f}_T = \hat{f}_{T_i}$.

To further make the learning of the common feature space more robust, we add a learnable decoding module that must reconstructs F_{T_i}, $i = 1, ..., N$, from $\hat{f}_T$. It is to ensure that the projected features can be "mapped back". We denote the reconstructed features $\hat{F}_{T_i}$, $i = 1, ..., N$. Figure 2 (b) summarize the whole amalgamation process.

3 FRAMEWORK DETAILS

3.1 Teacher Models

Our architecture uses HRNet-W32 (Wang et al., 2020) as backbone. It contains four stages with four parallel convolution streams. The resolutions are 1/4, 1⁄8, 1⁄16, and 1⁄32 of the input image, while the widths (numbers of channels) of the convolutions are C, $2C$, $4C$, and $8C$ ($C = 32$). Such backbone has been used for many tasks, including image classification, semantic segmentation, human pose estimation, object detection and facial landmarks detection[1]. This illustrates that it is capable of extracting various kind of features, and its design also allows to keep quite high resolution information, which is convenient for our bottom-up approach. Moreover, many of these trained versions are available online, and can be directly used as teachers for our knowledge amalgamation process. Yet, we also would like to use teachers specialized in emotion recognition task, and as far as we know, HRNet-W32 has not been used in this field. This is why we decided to train two more teachers.

The first teacher has been trained for Facial Emotion Recognition (FER), because facial expressions are well known to be one of the most important non verbal cues to infer people's emotions. We used AffectNet (Mollahosseini et al., 2017) as database, which regroups more than 1,000,000 facial images from the Internet. Actually, we trained our teacher with a subset of 291,651 pictures of aligned faces (224 × 224), annotated for 8 emotions (Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise). Nevertheless, in context aware emotion recognition images, it is not always possible (or reliable) to only use the facial expression of a person to assess its emotion. We then have to focus on other clues, such as the places, the objects or the activity to extrapolate the emotion. In Visual Emotion Analysis (VEA), we aim to recognize the basic emotions expressed by images of all kind. Thus, this field is quite correlated with CAER, and this is why we trained the second teacher on this task. We selected the EmoSet database (Yang et al., 2023), which contains 3.3 million images in total. The set of 118,102 images we used has been labeled by human annotators for 8 emotions (Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sadness).

Table 2: Performance of our two teachers compared with the baselines on AffectNet and EmoSet databases.

Database	Baseline Acc. / F1-score	Teacher Acc. / F1-score
AffectNet	0.64 / 0.55	0.61 / 0.61
EmoSet	0.74 / -	0.71 / 0.72

For both teachers, the classification head we used is similar to the one proposed in (Wang et al., 2020), except that we reduce the number of channels of the second to last layer from 2048 to 1024, because of the small amount of classes we are dealing with. We use random perspective transformations and random horizontal flip as data augmentations. We train the models using the Cross Entropy loss and the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-3}, during 150 epochs.

Then, for both AffectNet and EmoSet, the selected model is the one giving the best average F1-Score on the validation set. We preferred average F1-score over accuracy because of its ability to better represent the classifier's performance when dealing with imbalanced classes, which is often the case with emotions. We also compared the performance of our teachers with the baselines of both datasets (Table 2). The accuracy of our AffectNet teacher is slightly lower than the baseline (Mollahosseini et al., 2017), but the F1-score is better, which is not so surprising since we selected our best model using this latter metric. For EmoSet, we are also lower than the baseline (Yang et al., 2023) regarding the accuracy, and the F1-Score is not provided by the authors. Yet, our teachers are simple classifiers without any attempt to be particu-

[1] https://github.com/HRNet

larly well adapted for their task, but we assume that their performances are sufficient for our experiments. Actually, using even better models could be a future work.

We finally selected a third teacher from the pool of available pre-trained models online, which has been used for Object Detection (OD) on COCO dataset (Lin et al., 2014).

3.2 Losses

For the training of the emotion recognition task in the case of multi-labels annotations, we use the following loss:

$$L_{emo} = \frac{-1}{N} \sum_{n=1}^{N} \sum_{i=1}^{E} \begin{cases} \log\left(\sigma(\hat{Y}_{n,i})\right) & \text{if } Y_{n,i} = 1 \\ \log\left(1 - \sigma(\hat{Y}_{n,i})\right) & \text{otherwise,} \end{cases} \quad (1)$$

where N is the number of people in the image, E is the number of emotions, $\hat{Y}_{n,i}$ and $Y_{n,i}$ are namely the logit and the ground truth for the $i-th$ emotion of the $n-th$ person, and $\sigma()$ the sigmoid function. When the annotation is a single class of emotion, the loss becomes:

$$L_{emo} = \frac{-1}{N} \sum_{n=1}^{N} \log\left(\frac{\exp(\hat{Y}_n)}{\sum_{i=1}^{E} \exp(\hat{Y}_{n,i})}\right), \quad (2)$$

where $\hat{Y}_n$ is the logit of the class that should be activated for the $n-th$ person (regarding the annotation), and $\hat{Y}_{n,i}$ the logit for the $i-th$ class of emotion of the $n-th$ person.

Since the predictions for a given person are extracted from the emotion maps by taking the average value $\bar{P}$ of the pixels inside its bounding box, we also add the following constraint:

$$L_{var} = \frac{1}{N_P} \sum_{i=1}^{N_P} \left(P_i - \bar{P}\right)^2, \quad (3)$$

where N_P is the number of pixels inside the bounding box, and P_i is the value of the $i-th$ pixel. Basically, this constraint amounts to imposing zero variance within the pixels of the bounding box.

For the detection task, following (Zhou et al., 2019), 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:

$$FL = \frac{-1}{N} \sum_{xy} \begin{cases} \left(1 - \hat{Y}_{xy}\right)^{\alpha} \log\left(\hat{Y}_{xy}\right) & \text{if } Y_{xy} = 1 \\ \left(1 - Y_{xy}\right)^{\beta} \left(\hat{Y}_{xy}\right)^{\alpha} \\ \log\left(1 - \hat{Y}_{xy}\right) & \text{otherwise} \end{cases} \quad (4)$$

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:

$$L_{size} = \frac{1}{N} \sum_{k=1}^{N} \left|\hat{S}(p_k) - s_k\right|, \quad (5)$$

where $\hat{S}$ are the width and height prediction maps. Hence, the total detection loss is:

$$L_{det} = FL + \gamma \times L_{size} \quad (6)$$

where $\gamma = 0.1$ following (Zhou et al., 2019). Concerning the knowledge amalgamation process, we define our feature loss as following:

$$L_{feat} = \frac{mask(\hat{f}_T - \hat{f}_S)^2}{N_{mask}}, \quad (7)$$

where $mask(A - B)$ is used to apply the difference only between the strictly positive pixels of A and the corresponding ones in B, and N_{mask} is the number of pixels in the mask. Since we have $\hat{f}_T \geq 0$ and $\hat{f}_S \geq 0$ because of the ReLu layers, it is a trick to focus on the pixels carrying the information instead of encouraging the network to output zero values. The reconstruction loss also uses the $mask(A - B)$ function on the N teacher's features F_{T_i}, $i = 1, \dots, N$ and the reconstructed ones $\hat{F}_{T_i}$, $i = 1, \dots, N$:

$$L_{rec} = \sum_{i=1}^{N} \frac{mask(F_{Ti} - \hat{F}_{Ti})^2}{N_{mask}}, \quad (8)$$

Finally, the total loss is defined as:

$$L_{TOT} = L_{det} + L_{emo} + L_{var} + L_{feat} + L_{rec} \quad (9)$$

3.3 Training Procedure

The method is built with the Pytorch toolbox (Paszke et al., 2019). The students trained with the knowledge amalgamation process are first initialized with the pre-trained weights obtained from ImageNet (Deng et al., 2009). Then, every models are trained on the target dataset during 150 epochs with a batch size equal to 20. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-3}. For all the experiments, the input images are resized to 512 × 512, and we apply random perspective transformation and random horizontal flip as data augmentations. Finally, the best model is defined as the one with the lowest total validation loss.

4 EXPERIMENTS AND RESULTS

4.1 Databases

The EMOTIC database (Kosti et al., 2017; Kosti et al., 2019) is probably the most used dataset in CAER. It is composed of 23,571 images, totaling 34,320 annotated people in unconstrained environments. Each picture contains at least one subject, which is annotated with a bounding box, and emotionally labeled using 26 discrete categories in a multi-label manner, as well as 3 continuous dimensions (Valence, Arousal, and Dominance). Using such a large number of emotion classes coupled with multi-label annotations inevitably results in quite unbalanced data, making the dataset particularly challenging. Finally, the provided partition of the dataset is 70% for the training set, 10% for the validation one, and the remaining 20% are used for testing.

The HECO database (Yang et al., 2022) is on contrary not very used yet, since it is more recent. It regroups 9,385 images and 19,781 annotated people, with rich context information and various agent interaction behaviours. Here, only 8 discrete categories are used (Anger, Disgust, Excitement, Fear, Happiness, Peace, Sadness, and Surprise), but the 3 continuous dimensions are similar to the ones used in EMOTIC. The people are also annotated with 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. Since the authors do not provide any partition of their dataset, we created our own sets by first sorting the images into two groups: a first one that contains the images where only one subject has been annotated, and a second one regrouping the remaining images. Then, for each group, all the images were randomly split into training, validation and test sets (70%, 10%, and 20%). Finally, the corresponding sets from the two groups are merged to form the final training, validation, and test partitions. With this procedure, we ensure that our bottom-up approach is trained and tested with various images representing both single and multiple subjects.

4.2 Metrics

To evaluate the model for the CAER task, we calculate the Average Precision score (AP) for all the emotion categories, and average it (mAP). The predictions are extracted from the emotion maps using the annotated bounding boxes. For the person detection task, we used the COCO API to obtain a mAP score computed over 10 Intersection over Union (IoU) thresholds. This metric was initially proposed to evaluate object detection on the COCO dataset (Lin et al., 2014), and here we are in a special case where we have only one object class (person) to detect. We finally use the new metric proposed by (Cladière et al., 2023), where the prediction of a subject's emotion is automatically counted as a false negative if the model was not previously able to correctly estimate its bounding box. This metric is actually more representative of the model's true capability during ”on-the-field” inferences, where the emotion recognition task can only be performed if a subject is first clearly detected in the image.

Table 3: mAP scores obtained on HECO and EMOTIC datasets for emotion recognition task.

HECO

Teacher	F-T	KA	Baseline
FER	26.73	**27.94**	26.62
VEA	27.14	**27.26**	
OD	26.27	**27.30**	

EMOTIC

Teacher	F-T	KA	Baseline
FER	26.15	**26.61**	25.09
VEA	25.21	**26.74**	
OD	24.86	**27.10**	

4.3 Results in Single-Task Configuration

In order to validate the concept of knowledge amalgamation (KA), we first evaluated the performance of a student model guided by a single teacher on the HECO and EMOTIC databases. These results are then compared to those obtained when the teacher models are fine-tuned (F-T) on these same databases. We define the baseline as the score obtained with a similar model initialized with the weights inherited from ImageNet then fine-tuned on the target dataset. The results are given in Table 3. For the two emotion-related teachers (FER and VEA), the fine-tuning on HECO and EMOTIC gives better results than the baseline, which is not the case for the OD teacher. It confirms that using pre-trained models from correlated tasks before fine-tuning on the considered database is beneficial. However, with our KA method, whatever the teacher model used, the student network will outperform it on both HECO and EMOTIC, which highlights the superiority of this approach. This may be explained by the fact that the model will quickly over-fit in the case of fine-tuning, probably forgetting the pre-learned task, whereas with KA the student learns jointly to reproduce useful features and exploit them for the new task.

4.4 Results in Multi-Tasks Configuration

Similar experiments have been conducted with the detection head added to the student model. In this configuration, it is evaluated regarding its performances in Emotion Recognition (ER), Person Detection (PD), and ”On-the-field” Emotion Recognition (OER), where the emotion recognition score depends on the model’s person detection capabilities (Cladière et al., 2023). The results on HECO are presented in Table 4, and those on EMOTIC in Table 5. The values in bold correspond to the best score between F-T, KA and the baseline. The underlined values are the best score between F-T and KA when none of them are beating the baseline.

On HECO, KA is always better than F-T and the baseline, except for PD task using the VEA teacher where KA is better than F-T but still lower than the baseline. On EMOTIC, the results are more mixed. For the FER teacher, KA always outperforms F-T and the baseline, but for the VEA teacher, while beating the F-T method, KA does not surpass the baseline in ER and PD tasks. Concerning the OD teacher, F-T leads to the best scores for all the tasks.

It is interesting to note that it is always using KA with the FER teacher that the best PD scores are obtained. This could be explained by the fact that this teacher was only trained with facial images, thus giving feature maps activated at the level of the subjects’ heads. Thus, the student model only has to learn to make the connection between the head already highlighted in the features and the rest of the body to predict a bounding box. This is also beneficial for many close-range images inside EMOTIC and HECO where only faces are visible, which could disrupt a detection model pre-trained with more far-range examples.

Table 4: mAP scores obtained on HECO dataset for ER task, PD task, and OER task after multi-tasks training.

HECO ER

Teacher	**F-T**	**KA**	**Baseline**
FER	25.86	**26.75**	26.42
VEA	26.58	**26.82**	
OD	27.66	**27.67**	

HECO PD

Teacher	**F-T**	**KA**	**Baseline**
FER	33.59	**35.75**	35.11
VEA	33.84	34.08	
OD	32.71	**35.26**	

HECO OER

Teacher	**F-T**	**KA**	**Baseline**
FER	19.72	**20.36**	20.10
VEA	19.42	**20.25**	
OD	20.71	**21.07**	

Table 5: mAP scores obtained on EMOTIC dataset for ER task, PD task, and OER task after multi-tasks training.

EMOTIC ER

Teacher	**F-T**	**KA**	**Baseline**
FER	21.87	**22.61**	22.33
VEA	22.00	22.33	
OD	**22.74**	22.55	

EMOTIC PD

Teacher	**F-T**	**KA**	**Baseline**
FER	49.02	**53.74**	52.12
VEA	50.43	51.24	
OD	**52.39**	50.37	

EMOTIC OER

Teacher	**F-T**	**KA**	**Baseline**
FER	20.37	**20.93**	20.68
VEA	20.13	**20.83**	
OD	**20.96**	20.46	

4.5 Results with Combinations of Teachers

Since the amalgamation process allows to distill the knowledge of multiple teachers inside a single student, we tested our approach with different combinations of teachers to train a student for CAER task. As we can see in Table 6, all the combinations lead to higher scores than the baseline on both datasets. However, FER+VEA did not perform better than FER and VEA alone, although one would have thought that the two teachers complement each other and benefit the CAER. Actually, this could be explained by the fact that the AffectNet and EmoSet databases were not annotated by the same people, which risks leading to variations in the ground truths, but also with a different set of emotions. As a result, the two teachers could provide contradictory features instead of complementary ones, thus disrupting the amalgamation. On the other hand, the FER+OD and VEA+OD combinations give even better results on HECO than using the teachers independently, which demonstrate the potential of the amalgamation process to merge different yet complementary knowledge.

On EMOTIC, it is finally OD alone which gives the best score. Furthermore, the fact that combining all the teachers does not lead to the best score, while improving the FER+VEA combination on both datasets, illustrates that there is room for improve-

Table 6: mAP scores obtained on HECO and EMOTIC datasets for emotion recognition task when using up to 3 teachers during the knowledge amalgamation process.

HECO

Teacher(s)	KA	Baseline
FER	27.94	26.62
VEA	27.26	
OD	27.30	
FER + VEA	26.78	
FER + OD	**28.53**	
VEA + OD	28.09	
FER + VEA + OD	27.24	

EMOTIC

Teacher(s)	KA	Baseline
FER	26.61	25.09
VEA	26.74	
OD	**27.10**	
FER + VEA	26.36	
FER + OD	26.54	
VEA + OD	26.82	
FER + VEA + OD	26.68	

ment of the method, especially in the design of the common feature space. Indeed, by studying in detail the distribution of the features from different teachers in this common space, we could enhance their fusion by preventing non-constructive cases, such has contradictory, redundant or unbalanced teachers.

5 CONCLUSION AND FUTURE WORKS

Unlike the other authors from the CAER field, we have opted in this paper for a single stream model, which does not require the use of off-the-shelf modules, or any pre-processing steps. Moreover, the solution we presented is a bottom-up approach instead of a top-down one, which allows to simultaneously estimate the emotion of all subjects within an image. Thus, our framework is very condensed and straightforward, and this is why its deployment on the field seems more feasible than typical solutions from the literature. We also introduced a knowledge amalgamation protocol to distill multiple specialized teachers into a single student network. When relying on a single teacher for guiding the student model, we systematically obtained better results than directly fine-tuning the teacher. On HECO, we further increased the score by combining two teachers, which is unfortunately not the case on EMOTIC. This illustrates the potential of our approach, even if more work can be done to design a more robust and consistent fusion between the features from many teachers. Finally, we integrated a detection head to our model, which therefore become autonomous since the predictions from the emotion maps can be extracted without using the annotations. Even in this multitask configuration, our knowledge amalgamation method is still almost always better than simply fine-tuning the teachers, for the three tasks we evaluated.

As part of future works, we would use more efficient teachers in their respective domain, which do not necessarily share the same architecture as the student. Indeed, we think that using better features as example would benefit the amalgamation process even more, and that features coming from different architectures could be more varied and complementary than those obtained with similar teachers. We would also make further investigations on the distribution of the features in the common feature space, to ensure that they are well aligned even if they come from different tasks. By having a more in-depth comprehension of this space, we assume that a refined version of our first attempt of knowledge amalgamation with several teachers could become more robust and reach even better scores.

ACKNOWLEDGMENTS

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

Barrett, L. F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. *Current directions in psychological science*, 20(5):286–290.

Bendjoudi, I., Vanderhaegen, F., Hamad, D., and Dornaika, F. (2021). Multi-label, multi-task cnn approach for context-based emotion recognition. *Information Fusion*, 76:422–428.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., and Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395.

Cladière, T., Alata, O., Ducottet, C., Konik, H., and Legrand, A.-C. (2023). Benet: A lightweight bottom-up framework for context-aware emotion recognition. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 100–111. Springer.

de Carvalho, M., Pratama, M., Zhang, J., and Sun, Y. (2022). Class-incremental learning via knowledge amalgamation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 36–50. Springer.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Hoang, M.-H., Kim, S.-H., Yang, H.-J., and Lee, G.-S. (2021). Context-aware emotion recognition based on visual relationship detection. *IEEE Access*, 9:90465–90474.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017). Emotic: Emotions in context dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 61–69.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2019). Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766.

Lee, J., Kim, S., Kim, S., Park, J., and Sohn, K. (2019). Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152.

Li, S. and Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Luo, S., Wang, X., Fang, G., Hu, Y., Tao, D., and Song, M. (2019). Knowledge amalgamation from heterogeneous networks by common feature learning. *arXiv preprint arXiv:1906.10546*.

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shen, C., Wang, X., Song, J., Sun, L., and Song, M. (2019a). Amalgamating knowledge towards comprehensive classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3068–3075.

Shen, C., Xue, M., Wang, X., Song, J., Sun, L., and Song, M. (2019b). Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3504–3513.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Wang, Z., Lao, L., Zhang, X., Li, Y., Zhang, T., and Cui, Z. (2022). Context-dependent emotion recognition. *Journal of Visual Communication and Image Representation*, 89:103679.

Yang, D., Huang, S., Wang, S., Liu, Y., Zhai, P., Su, L., Li, M., and Zhang, L. (2022). Emotion recognition for multiple context awareness. In *European Conference on Computer Vision*, pages 144–162. Springer.

Yang, J., Huang, Q., Ding, T., Lischinski, D., Cohen-Or, D., and Huang, H. (2023). Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394.

Ye, J., Wang, X., Ji, Y., Ou, K., and Song, M. (2019). Amalgamating filtered knowledge: Learning task-customized student from multi-task teachers. *arXiv preprint arXiv:1905.11569*.

Zhang, M., Liang, Y., and Ma, H. (2019). Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE.

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.