Multimodal Group Emotion Recognition In-the-wild Using Privacy-Compliant Features

Anderson Augusma, Dominique Vaufreydaz, Frédérique Letué

MULTIMODAL GROUP EMOTION RECOGNITION IN-THE-WILD USING PRIVACY-COMPLIANT FEATURES

AUTHOR VERSION

Anderson Augusma [1,2], **Dominique Vaufreydaz** [1], **Frédérique Letué** [2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

[1] Univ. Grenoble Alpes, CNRS, LJK, 38000 Grenoble, France

Figure 1: *Examples from the VGAF dataset. From left to right, examples of Positive, Neutral, and Negative classes.*

ABSTRACT

This paper explores privacy-compliant group-level emotion recognition "*in-the-wild*" within the EmotiW Challenge 2023. Group-level emotion recognition can be useful in many fields including social robotics, conversational agents, e-coaching and learning analytics. This research imposes itself using only global features avoiding individual ones, i.e. all features that can be used to identify or track people in videos (facial landmarks, body poses, audio diarization, etc.). The proposed multimodal model is composed of a video and an audio branches with a cross-attention between modalities. The video branch is based on a fine-tuned ViT architecture. The audio branch extracts Mel-spectrograms and feed them through CNN blocks into a transformer encoder. Our training paradigm includes a generated synthetic dataset to increase the sensitivity of our model on facial expression within the image in a data-driven way. The extensive experiments show the significance of our methodology. Our privacy-compliant proposal performs fairly on the EmotiW challenge, with 79.24% and 75.13% of accuracy respectively on validation and test set for the best models. Noticeably, our findings highlight that it is possible to reach this accuracy level with privacy-compliant features using only 5 frames uniformly distributed on the video.

***K*eywords**: Multimodal, Privacy safe, Transformer networks, Group emotion recognition in-the-wild.

1 Introduction

Emotion recognition research is of interest in multimodal interaction for numerous applications like social robotics, conversational agents, e-coaching, or learning analytics. Among others, one challenge is the automatic recognition of group-level emotions in ecological or "*in-the-wild*" scenarios, which involves considering group dynamics, individual behavior, emotional expressions, postures, environmental elements, and different kinds of activities [18]. This task is further complicated by cultural and ethnic factors and some technical issues like the quality of recording point of view. Adding concerns about ethics and privacy limits some choices in the data that can be employed as input and outputs of the machine learning algorithms.

This article presents a privacy-compliant architecture to classify audio-visual data labeled with group-level emotions: the VGAF dataset [18]. This machine learning model materializes our participation in the EmotiW 2023 [2] challenge. As our former research, Petrova et al. [15], we impose ourselves some privacy rules on the input of our machine learning model. Unlike other approaches [12, 17, 19, 20], our privacy-compliant model must not use any information that can be helpful to identify or track any person on the videos. It avoids individual data like postures or facial landmarks, for instance, focusing only on global features computed on images and sound.

The paper is structured as follows. After the presentation of related work (section 2) and the VGAF dataset (section 3), section 4 introduces our methodology. Section 5 details our extensive experiments while section 6 discusses the results in regard to the state-of-the-art.

2 Related Work

Progresses in machine learning in the last decade permit to address emotion recognition in real-world or "*in the wild*" scenarios, and more recently group level emotion prediction in videos. While some research remains monomodal [14, 15, 17], most of them use multimodal inputs [1, 6, 16]. Surrace et al. [21] used a method that examines both general and specific information. They focused on the overall scene in a top-down strategy. They also extracted personal information in a bottom-up strategy, isolating faces from wider images. Other studies [1, 8–12, 17, 19, 23, 25] also used individuals' information for emotion recognition. For example, Liu et al. [12] used a network for recognizing facial emotions and body poses, while Fisher et al [25] developed a method to detect facial expressions. This research goes in a different direction by focusing on global features to avoid harming as far as possible privacy.

The use of synthetic data for model training is a well-established practice in machine learning research. This data-driven method is often employed by researchers to enhance the generalizability or, a contrario, the specificity of their models. For instance, Dwibedi et al. [4] harnessed real videos to construct synthetic repetition videos for training their "Counting Out Time" model. Marín-Jiménez et al. [13] altered the relative head positions of individuals in images to comply with a real scenario. Their aim was to improve the generalizability of their Look At Each Other (LAEO) model. In research specifically related to the EmotiW challenge, Petrova et al. [15] employed monomodal synthetic static images to enable their model to concentrate on individuals' faces while disregarding background. Their method creates static images to train their feature extraction. This research extends this approach to generate coherent video sequences coupled with class-related sound.

3 VGAF Dataset

The VGAF dataset [18] is a collection of web videos that include a wide variety of genders, ethnicities, event types, numbers of people, and poses. These videos are grouped into three categories of group level emotions: Positive, Neutral, and Negative (as shown in Figure 1). The classification of each video was determined by a voting process involving several annotators, with a majority vote deciding the final classification. The dataset is divided into a Training set, a Validation set, and a Test set, with 2661, 766, and 756 videos in each set respectively. The Validation set is unevenly distributed, with more videos in the Positive and Neutral categories compared to the Negative category.

The videos include groups of at least two people and feature a range of events like interviews, festivals, parties, and protests, among others. They vary in resolution, and are split up into 5-second labeled clips. The frame rate varies from 13 to 30 frames per second. The camera focuses on different positions in each video, resulting in frames that highlight different group organisations. Thus, the technical and content varieties of the videos, the association of images and sounds to identify emotions, make the classification complexity of the VGAF dataset. One of the dataset's complexities is that some videos feature children playing and hitting tables or chairs, generating sounds that may be confused with protestation noises. However, these two events are different because playing with children creates a positive atmosphere, while protestations generate a negative atmosphere.

4 Methodology

4.1 Model architecture

The proposed neural architecture is illustrated in Figure 2. This model encompasses two monomodal branches producing embedding for video and audio frames. The video branch exploits a pre-trained, then fine tuned, vision transformer (ViT-Large, 14 patches) [3] ending by a custom linear layer with an output dimension of 1024, i.e. the embedding dimension of the model (d_{model}). The audio branch consumes a sequence of Mel-Spectrograms as input through four specialized CNN blocks. As depicted, these blocks are composed of 2D convolutions stack with Relu and a final max pooling. The final section of the audio branch integrates a Transformer Encoder layer, as outlined in [22] with fixed positional-encoding and four self-attention heads. The *feed-forward* hidden-size is set at twice the value of d_{model}, while output dimension remains d_{model}. To integrate interaction between audio and video modalities, a late fusion mechanism concatenates embeddings from the monomodal branches jointly with a Cross Attention one. In cross-attention, the audio modality serves as query (Q), while the video modality serves as key (K) and value (V). The resulting attention weights are then multiplied with the video embeddings to obtain the final transformed representation. An average operation is then applied to the concatenated sequence averaging all frame embeddings to one of size $3 * d_{model}$. The classification task is carried out using a fully-connected layer followed by a Softmax activation, which yields to the final classification result in 3 classes.

4.2 Privacy compliant features

As stated, one self-imposed constraint is to remain as privacy compliant as possible. To do so, our approach carefully avoids all individual features, i.e. features that could identify a person directly or within a group, focusing on global features.

For the video branch, our system does not employ body pose, shape, height, nor body language features like gesture or agitation. Computed facial information like FACS, emotion or person id are excluded, like counting and tracking individuals within the scene. The video input is a set of n still video frames uniformly taken over the video and resized to 224×224. n is chosen among two extrema: 1 frame per second (*fps*) and 15 frames per second on the 5-second videos of VGAF. From our experience, 1 *fps* is a lower band to get a non-random classification. As 15 *fps* is the lowest *fps* on the training videos, we set it as max *fps* to train on actual frames without repeating some of them. Thus, n is taken in $\{5,75\}$ for each video in this research.

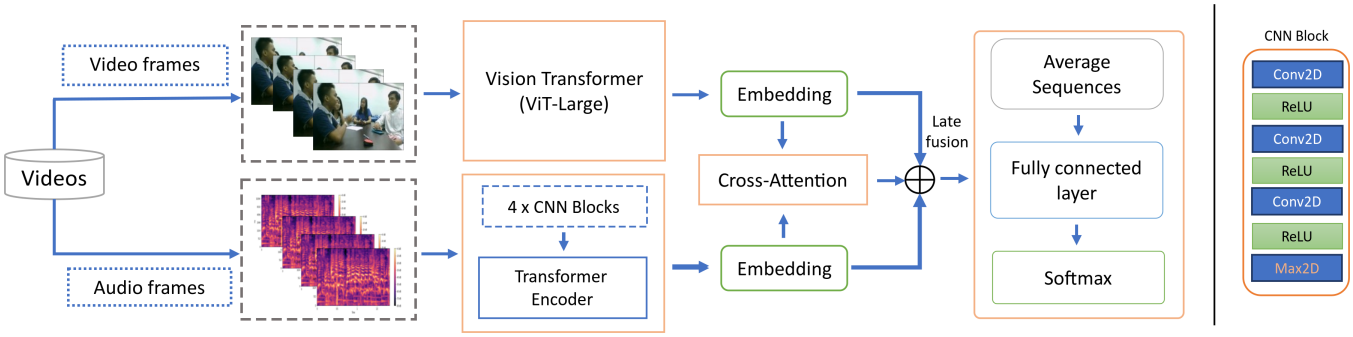

Figure 2: *At left, the proposed model is a combination of two monomodal branches, a Cross-Attention and a late fusion paradigm. The video branch uses a pre-trained vision transformer (ViT) model [3]. The audio branch encompasses 4 CNN blocks followed by a transformer encoder. At right, description of one CNN block used in the audio branch.*

For the audio processing, we avoid speaker identification, speaker diarization or any speech-to-text processing. The audio of all videos is standardized by resampling it at 16 Khz and converting it to mono channel. By compliance with video branch, we extract 5 or 75 audio frames per video. With 5 frames, the audio frame corresponds to 1 second with no overlap. In the case of 75 frames, a sliding window is set to 67 milliseconds to get the right number of frames. Last, each audio frame is converted into Mel-Spectrograms using 128 Mel filters, to produce an input image of 128×251 for the CNN blocks.

4.3 Synthetic video dataset

Inspired by former research [15], we augmented our training set with synthetic data. The purpose is to guide the neural network to concentrate on faces while searching for positive, neutral or negative contexts, ignoring the background. The original generation process involves placing real faces expressing basic emotions [5] onto random backgrounds [26] to create still images conveying emotions. To create synthetic videos, we take from 3 to 9 faces expressing the right emotions, for instance positive emotions to generate positive video, and randomly move them on a fixed background. To mitigate the impact of occlusions, we introduce masks in the generation process, which allows a maximum of ten percent of occlusion on faces. As generating audio is a very complex task in our context, we associated randomly an audio with the same class from the VGAF training set to the generated video sequence. The synthetic video dataset has the same size and class balance as the VGAF training dataset: 802 (30,16%), 923 (34,71%), 934 (35,13%) of positive, neutral and negative videos respectively.

5 Experiments

5.1 Optimal synthetic video ratio

We conducted experiments to measure the impact of synthetic data ratio on performance using 5 frames per video. The synthetic ratio was gradually increased from 10% to 50% with a 10% step of the total training data. The accuracies on the validation set of these variants can be found in Table 1. The best accuracy (75.07%) is reached using 30% of synthetic data in the training. Less or more data leads to lower performance. Following this result, in all subsequent experiments comprising standard and synthetic data, the synthetic ratio is set to 30%.

Table 1: *Evaluation of synthetic video ratio on the validation accuracy.*

Synthetic ratio	Synthetic videos	Total videos	Acc.
10%	297	2958	74.80%
20%	666	3327	74.93%
30%	**1140**	**3801**	**75.07%**
40%	1773	4434	74.41%
50%	2661	5322	74.28%

5.2 Multimodal ablation Study

To evaluate relevance of each of our training inputs, (VGAF video and audio, synthetic data), experiments of all their combinations with 5 and 75 frames were conducted. Models were trained using cross-entropy loss and the SGD optimizer with a fixed learning rate of 10^{-5}. We used the VGAF (audio and/or video) and the synthetic datasets in monomodal and multimodal variants architecture. The ViT-Large model consists of 24 pre-trained transformer encoder blocks on 14 patches. Its weights are frozen for 10 epochs, and then released to be fine-tuned on our data. Cross attention is present only when both video and audio branches of the network are active. Results on the VGAF validation set are summarized in Table 2.

Training on audio data only leads to average performance up to 56.4% with 5 frames per video. Using 75 audio frames, the performance decreases. This may be explained by the averaging effect of the sliding window applied in this case. Using only synthetic videos performs better with arround 60% of accuracy, showing interest of this approach. The best mono input performance is 74.15% using 75 video frames. As in many computer vision tasks, pre-trained vision transformers prove to be efficient to capture relevant features on the VGAF

dataset. Combining audio and videos inputs produces various performance in term of validation accuracy. Joining synthetic video and audio does not improve much the performance of individual inputs. The best multimodal performance (78.07%) is obtained by combining all inputs using 5 frames. Our best accuracy concerns a monomodal combination of VGAF and synthetic video dataset during training phase. This model outperforms all other models with a validation accuracy of 79.29%.

Five monomodal and multimodal versions of our model participated in the EmotiW challenge 2023. Performances on the test set are reported in Table 3. As anticipated, audio and synthetic data models (v1 and v2) remain around 55% of test accuracy. Unlike validation performance, video plus synthetic model (v3) is less accurate than the 2 multimodal ones (v4 and v5). Noticeably, both multimodal systems, using 5 (v4) or 75 frames (v5) per video, have the same performance even if their predictions differ: their prediction agreement, i.e. when the same class is predicted for the same video, is 88%. As expected, prediction agreements of v5 versus v1 and v3 are respectively 53% and 90%, highlighting that, for our architecture, audio provides less information than video.

Table 2: *Ablation study investigating the impact of different inputs on validation accuracy using 5 and 75 frames per video. The ablation considers all combinations of (VGAF) video, (VGAF) audio, and generated synthetic videos.*

	Input Data	5 Frames	75 Frames
	audio	56.40%	54.96%
Video	synt_video	57.44%	60.05%
	video	73.62%	74.15%
	video + synt_video	75.98%	**79.24**%
Audio + video	synt_video + audio	62.14%	58.75%
	video + audio	76.11%	77.42%
	video + audio + synt_video	**78.07**%	78.72%

6 Discussion

Excluding systems from the EmotiW 2023 [2] challenge as results were not available at the writing time of this article, one can compare the 5 versions of our model with the state-of-the-art. Accuracies on validation and test sets are reported in Table 4. Using only audio, as we limited input to Mel-spectrograms, our v1 model do not reach score of Ottl et al. [14] proposal, due to their more complex usage of OpenSMILE features [7] coupled with Deep Spectrum analysis. On video only system, the v3 model have the best performance even if, as said, its test accuracy is lower than its validation counterpart. On multimodal systems, v4 and v5 remain under the proposal of Liu et al. [12]. Nevertheless, our approach exposes fair performance while using only global thus privacy-compliant features.

Table 3: *Test Set Accuracy of 5 versions of our model.*

Vers.	Input data	Nb Frames	Acc.
v1	audio	5	55.29%
v2	synt_video	75	54.23%
v3	video + synt_video	75	74.73%
v4	video + synt_video + audio	5	75.13%
v5	video + synt_video + audio	75	75.13%

Table 4: *Comparison with SOTA systems on the VGAF dataset. Columns detail usage of individual features, accuracy on official validation set and, when reported, on the test set. The table is ordered by input modalities and then by validation accuracy.*

		Ind. feat.	Val. Acc.	Test Acc.
Audio	*Ours v1*		*56.40%*	*55.29%*
	Ottl et al. [14]		59.40%	62.30%
Video	Petrova et al. [15]		52.36%	59.13%
	Ours v2		*60,05%*	*54.23%*
	Savchneko et al. [17]	✓	70.23%	-
	Ours v3		***79.24%***	*74.73%*
Audio + Video	Evtodienko et al [6]		60.37%	-
	Sharma et al [19]	✓	61.61%	66.00%
	Pinto et al. [16]		65.74%	-
	Wang et al. [24]		66.19%	66.40%
	Sun et al. [20]	✓	71.93%	-
	Belova et al. [1]		71.95%	-
	Liu et al. [12]	✓	74.28%	**76.85%**
	Ours v4		*78.07%*	*75.13%*
	Ours v5		*78.72%*	*75.13%*

Looking at intrinsic analysis of our models, they benefit from the pre-trained ViT network fine-tuned on the VGAF data and from our synthetic video approach but lack in the audio branch. Increasing the number of parameters like number of attention heads wages unconditionally to overfitting, showing that such architectures require a lot of training data. Gathering more group emotion data, (transfer) learning on close or generated data can be part of the solution. Our synthetic generation process is straightforward and can be enhanced in several ways. Generative Adversarial Networks can be trained to create new data. But all these approaches are limited by the complexity and diversity of group emotion videos in terms of audio and video content.

As anticipated, avoiding individual features leads to set a thread-off between performance and privacy. In this research, our aim was to investigate the maximum performance that could be obtained without contravening this rule. Depending on the target application context, some additional information could be added to improve performance.

7 Conclusion

This research introduces a privacy-compliant yet efficient method for recognizing group emotions in uncontrolled environments, proposing to eliminate the need for individual feature extraction. The privacy compliance is achieved through

a selection of non-individual characteristics in signal inputs, providing only global information on the scene. The proposed model is composed of a video and an audio branches. The video branch is based on a ViT architecture fine tuned to compute relevant embedding on image sequences. The audio branch extracts Mel-spectrograms and feed them through CNN blocks into a transformer encoder. A cross-attention mechanism is added before the late multimodal fusion and the final classification. Our training paradigm includes a generated synthetic dataset to increase the sensitivity of our model on facial expression within the image in a data-driven way.

The extensive experiments show the relevance of our methodology. Our proposal performs fairly on the EmotiW challenge, with 79.24% and 75.13% of accuracy respectively on validation and test set for the best models. Contributions of each modality differ, the audio branch having rooms for larger improvement. Noticeably, our findings highlight that it is possible to reach 75% of accuracy with privacy-compliant features using only 5 frames uniformly distributed per video.

To conclude, the targeted privacy compliance of this research comes at a cost in performance. The accuracy remains slightly lower than other methods employing individual features. The next research challenge is thus to validate whatever it is possible, with individual features, to minimize the performance gap of privacy-compliant models, allowing to use them in more application contexts.

ACKNOWLEDGMENTS

This work was supported by the PERSYVAL Labex (ANR-11-LABX-0025). This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD010614233 made by GENCI. We express our gratitude to Garance Dupont-Ciabrini for her contribution to the first step of the synthetic video preparation.

References

[1] Natalya S Belova. Group-level affect recognition in video using deviation of frame features. In *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, volume 13217, page 199. Springer Nature, 2022.

[2] Abhinav Dhall, Monisha Singh, Roland Goecke, Tom Gedeon, Donghuo Zeng, Yanan Wang, and Kazushi Ikeda. Emotiw 2023: Emotion recognition in the wild challenge. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI 2023)*, 2023.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

[4] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10387–10396, 2020.

[5] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. Faces-a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42:351–362, 2 2010. ISSN 1554351X. doi:10.3758/BRM.42.1.351.

[6] Lev Evtodienko. Multimodal end-to-end group emotion recognition using cross-modal attention. *CoRR*, abs/2111.05890, 2021. URL `https://arxiv.org/abs/2111.05890`.

[7] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[8] Xin Guo, Bin Zhu, Luisa F Polanía, Charles Boncelet, and Kenneth E Barner. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. *Proceedings of the International Conference on Multimodal Interaction (ICMI 2018)*, 2018. doi:10.1145/3242969. URL `https://doi.org/10.1145/3242969.3264990`.

[9] Xin Guo, Luisa F. Polanía, Bin Zhu, Charles Boncelet, and Kenneth E. Barner. Graph neural networks for image understanding based on multiple cues: Group emotion recognition and event recognition as use cases. *CoRR*, abs/1909.12911, 2019. URL `http://arxiv.org/abs/1909.12911`.

[10] Aarush Gupta, Dakshit Agrawal, Hardik Chauhan, Jose Dolz, and Marco Pedersoli. An attention model for group-level emotion recognition. *Proceedings of the International Conference on Multimodal Interaction (ICMI 2018)*, pages 611–615, 10 2018. doi:10.1145/3242969.3264985.

[11] M. Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 9 2019. ISSN 15662535. doi:10.1016/j.inffus.2018.09.008.

[12] Chuanhe Liu, Wenqiang Jiang, Minghao Wang, and Tianhao Tang. Group level audio-video emotion recognition using hybrid networks. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI 2020)*, pages 807–812. Association for Computing Machinery, Inc, 10 2020. ISBN 9781450375818. doi:10.1145/3382507.3417968.

[13] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos, 2019.

[14] Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller. Group-level speech emotion recognition utilising deep spectrum features. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 821–826, 2020.

[15] Anastasia Petrova, Dominique Vaufreydaz, and Philippe Dessus. Group-Level Emotion Recognition Using a Unimodal Privacy-Safe Non-Individual Approach. In *EmotiW2020 Challenge at the 22nd ACM International Conference on Multimodal Interaction (ICMI2020)*, Utrecht, Netherlands, October 2020. URL `https://inria.hal.science/hal-02937871`.

[16] João Ribeiro Pinto, Tiago Gonçalves, Carolina Pinto, Luís Sanhudo, Joaquim Fonseca, Filipe Gonçalves, Pedro Carvalho, and Jaime S Cardoso. Audiovisual classification of group emotion valence using activity recognition networks. In *IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS 2020)*, pages 114–119. IEEE, 2020.

[17] Andrey V Savchenko and IA Makarov. Neural network model for video-based analysis of student's emotions in e-learning. *Optical Memory and Neural Networks*, 31(3): 237–244, 2022.

[18] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. Automatic group level affect and cohesion prediction in videos. *8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, pages 161–167, 9 2019. doi:10.1109/ACIIW.2019.8925231.

[19] Garima Sharma, Abhinav Dhall, and Jianfei Cai. Audiovisual automatic group affect analysis. *IEEE Transactions on Affective Computing*, 2021.

[20] Mo Sun, Jian Li, Hui Feng, Wei Gou, Haifeng Shen, Jian Tang, Yi Yang, and Jieping Ye. Multi-modal fusion using spatio-temporal and static features for group emotion recognition. *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI 2020)*, pages 835–840, 10 2020. doi:10.1145/3382507.3417971.

[21] Luca Surace, Massimiliano Patacchiola, Elena Battini Sönmez, William Spataro, and Angelo Cangelosi. Emotion recognition in the wild using deep neural networks and bayesian classifiers. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017)*, pages 593–597, 2017.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] Kai Wang, Debin Meng, Xiaoxing Zeng, Kaipeng Zhang, Yu Qiao, Jianfei Yang, and Xiaojiang Peng. Cascade attention networks for group emotion recognition with face, body and image cues. *Proceedings of the 2018 International Conference on Multimodal Interaction (ICMI 2018)*, pages 640–645, 10 2018. doi:10.1145/3242969.3264991.

[24] Yanan Wang, Jianming Wu, Panikos Heracleous, Shinya Wada, Rui Kimura, and Satoshi Kurihara. Implicit knowledge injectable cross attention audiovisual model for group emotion recognition. In *Proceedings of the International Conference on Multimodal Interaction (ICMI 2020)*, pages 827–834. Association for Computing Machinery, Inc, 10 2020. ISBN 9781450375818. doi:10.1145/3382507.3417960.

[25] Dai Yu, Liu Xingyu, Dong Shuzhan, and Yang Lei. Group emotion recognition based on global and local features. *IEEE Access*, 7:111617–111624, 2019. ISSN 21693536. doi:10.1109/ACCESS.2019.2932797.

[26] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. URL `http://arxiv.org/abs/1506.03365`.