Multi-task Learning for Human Affect Prediction with Auditory–Visual Synchronized Representation

Euiseok Jeong[1], Geesung Oh[1], and Sejoon Lim[2*]
[1]Graduate School of Automotive Engineering, Kookmin University
[2]Department of Automobile and IT Convergence, Kookmin University
{euiseok_jeong,gsethan17,lim}@kookmin.ac.kr

Abstract

With the development of the big data and deep learning technologies, research on predicting human affects in the wild using deep neural networks is being actively conducted. Many researchers use image and audio together to improve the affect prediction performance. However, the synchronization between image and audio data has not yet been achieved. Moreover, many different ways can be employed to annotate human affects, and the annotations in many datasets are not identical. The data cannot be utilized in supervised learning without the annotation of the task to be predicted. This study proposes a multi-task human affect prediction model with multimodal input and knowledge distillation to address the abovementioned problems. We used SoundNet, which was trained to transfer visual knowledge into auditory representations, to extract synchronized auditory–visual representations. Knowledge distillation was applied to utilize all datasets with incomplete labels. This model used image and audio data to predict the valence–arousal, expression, and action units and was validated using the Aff-Wild2 dataset. When auditory–visual synchronized representation was used, the performance improved by 11.83% and 230.16%, respectively, compared to when visual or auditory representation was used alone. When knowledge distillation was applied, the performance improved by 15.38% compared to when it was not. Consequently, the proposed model achieved a 0.95 performance for the multi-task learning task on the Aff-Wild2 test dataset. This performance is equivalent to that of the second place in the 3rd Affective Behavior Analysis in the wild Multi-task Learning Challenge.

1. Introduction

Human affective behavior research is an important aspect of the human computer interaction field and is being actively studied along with the development of the big data and deep learning technologies. However, applying the study of human affective behavioral research in the wild is difficult. To address these problems, the 3rd Affective Behavior Analysis in the wild (ABAW) competition is held in conjunction with the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2022. The 3rd ABAW includes the four following challenges: Valence–Arousal (VA) Estimation Challenge; Expression (Expr) Classification Challenge; Action Unit (AU) Detection Challenge; and Multi-task Learning (MTL) Challenge.

Audio information is closely related to human affects. It can reveal human affects in the form of words or elicit human affects in the form of music. Kuhnke et al. [21] and Deng et al. [4] used image and audio data together as input streams and showed that audio input is effective in recognizing human affects. However, no synchronization occurred between the image and audio data. Data synchronization is required to maximize the multimodal effectiveness.

For supervised learning, data cannot be utilized without the annotation of the task to be predicted. The datasets of the three Aff-Wild2 tasks have the same image and audio input type or shape, but different annotations. It is necessary to utilize as much data as possible to improve performance.

To address the abovementioned problems, we propose herein a multi-task model applied with knowledge distillation with auditory–visual synchronized representations.

We imported a pre-trained SoundNet [1] to extract natural synchronized representations from the audio data. SoundNet was trained to transfer visual knowledge into sound modality. We also applied knowledge distillation to utilize all datasets with different annotations. The teacher model was trained using only data with ground truths for the task. After the teacher model training, the teacher model output was used as a soft label to transfer dark knowledge to the student model. Even if there is no ground truth in the data, the student model can learn dark knowledge using soft labels.

We imported SoundNet and FER model [24] as a backbone networks. The auditory representation extracted from

the audio data, while the visual representation extracted from the image data using pre-trained backbone networks. The auditory and visual representations were concatenated and fed into a feature extractor. The feature extracted from the feature extractor was fed into the task network of each task to make predictions for each of the three tasks.

By training only the shallow network with frozen backbone networks, we achieved a 0.9531 performance for the ABAW MTL task test dataset, where the performance was measured as the sum of the metrics for the VA, EXPR, and AU tasks proposed by [13].

The primary contributions of this study are as follows:

- We extracted natural synchronized auditory representations from the audio data to maximize the effectiveness of the multimodal input.
- We applied knowledge distillation to train with incomplete labels.
- We achieved a good performance by training only the shallow network with frozen backbone networks.

2. Related Work

These days, research for recognizing human affects in various ways has been actively studied.

2.1. ABAW

Recently, data-based research on human effective behavior has been growing rapidly, and ABAW competition is contributing greatly to the development of human effective research [13–20, 31]. The first competition of the ABAW competition was held at the 2020 International Conference on Automatic Face and Gesture Recognition (FG), the second at the 2021 International Conference on Computer Vision (ICCV), and the third at the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). The ABAW competition consists of three challenges: 2D valence–arousal estimation (VA), 8 categorical representation classification (EXPR), and 12 facial action unit detection (AU). In the third competition, an MTL challenge for a multi-tasking model that performs three tasks at the same time has been added, making it a total of four challenges. Prior to the addition of MTL challenge, in the first and second competitions, most of the top teams proposed deep learning-based models that could perform three tasks simultaneously [3, 4, 21, 29, 30, 33, 34].

The methods of the top-ranked teams can be classified as input data as well as output. All top-ranked teams used image data from the video. Some teams used only image data [3,29,30,34], but some teams used audio data [4,21,33] along with image data to improve performance. In the 3rd ABAW MTL challenge, two of the top four teams, including our method, used the audio input data [35].

Kuhnke et al. [21] utilized ResNet [6], a representative network, for image classification to extract auditory representations. Deng et al. [4] and Zhang et al. [33] adopted a network that achieved good performance in audio-related tasks with speech activity detection [10] and audio classification [7] as backbone networks. Zhang et al. [35] imported Bert model [5] to extract word embedding features. The team mentioned above improved performance by using auditory representations with visual representations, but simply concatenated visual and auditory representations. Better performance can be expected when audio and images are synchronized naturally through the SoundNet [1].

2.2. Knowledge distillation

The knowledge distillation was proposed by Hinton et al. [8]. The output of the pre-trained teacher model is scaled by softmax function with temperature and used as a soft label to train the student model. The student model learns inter-class similarity(dark knowledge) from the soft label of the teacher model and achieves performance similar to the teacher model despite being shallower than the teacher model. Zhang et al. [32] improved performance through a self-distillation technique with a student model with the same structure as the teacher model. At the 2nd ABAW competition in 2021, several teams [4, 26, 27] applied knowledge distillation and were also listed on the leaderboard. In particular, Deng et al. [4] made it possible to train deeper dark knowledge using the knowledge distillation, the ensemble, and the generation technique in which a trained student model becomes a teacher model and trains a new student model.

3. Problem definition

$\{X,Y\}$ represents the train data. X represents the input data; and Y represents the ground truth. The model function f inputs X and outputs Y. For convenience of notation, the batch size of all tasks is assumed to be b, but can be defined by being divided by the same number of iterations, even if the number of data of the task is different. X consists of X_{img} and X_{aud}. X_{img} represents the image data. X_{aud} represents audio data. $X = \{X_{img} \in R^{b\times N_{img}\times H\times W\times C_{img}}, X_{aud} \in R^{b\times sr\cdot T_{aud}\times C_{aud}}\}$. N_{img} is the number of input images. W is the width. H is the image height. H and W are same in this dataset. C_{img} is the number of channels in the image. In audio data, sr represents the sample rate of the audio data. T_{aud} is the audio data time before the prediction time. C_{aud} is the number of audio channels.

X_{img} contains images for the past T_{img} seconds from the time of prediction. N_{fps} represents the number of images included in the 1 second video. Of the total $T_{img} \cdot N_{fps}$ images, the N_{img} image was extracted with stride S. Y comprises four types: $Y = \{Y_{va} \in$

$R^{b\times 2}, Y_{expr} \in R^{b\times 8}, Y_{au} \in .R^{b\times 12}, Y_{mtl} \in R^{b\times 22}\}$. Y_{va} represents continuous valence and arousal in the [1, 1] range. Y_{expr} is a one-hot encoded vector for the following eight emotion categories: neutral, anger, disgust, fear, happy, sad, surprised, and others. Y_{au} contains 12 face motion unit labels, namely AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU23, AU24, AU25, and AU26. In the teacher–student training algorithm, the teacher model is denoted by f^{tea}. The soft label of the teacher model for the task i is $f_i^{tea}(X)$.

4. Methodologies

This section describes the architecture, knowledge distillation, loss function, and learning algorithm of the proposed model.

4.1. Architecture

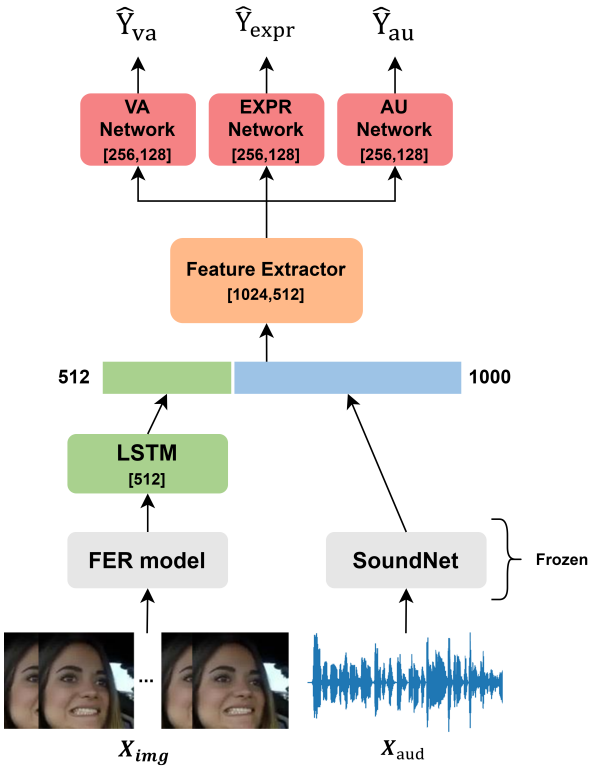

Figure 1. Model architecture. FER model and SoundNet are frozen. The numbers in the model block refer to the unit numbers of fully connected layers.

Figure 1 depicts the architecture of the proposed model. We used a CNN architecture based on the FER model of the DRER to extract the visual representation [24]. The FER model was also used in CAPNet [23] to extract the visual representation. The FER model consisted of ResNeXt [28] and SENet [9] and was pre-trained with AffectNet [22]. It is a model that predicts valence and arousal; hence, we removed the last layer from the pre-trained model and used the output of the FER model 512-dimensional vector as a visual representation. The N_{img} image was fed into the FER model and a visual representation of the shape (N_{img},512) was extracted. The extracted visual representation is fed into the LSTM layer to capture temporal features.

We adopted the pre-trained SoundNet proposed by Aytar et al. [1] to extract the auditory representation. SoundNet was trained to transfer visual knowledge into auditory modalities by leveraging a huge amount of unlabeled videos [1], making it particularly effective in cross-model approaches that use both image and audio data [11]. In the SoundNet Architecture, the raw waveform fed into the sound feature extractor and the 1000-dimensional vector was extracted. We adopted SoundNet without fine-tuning as the backbone network. Thus, the audio data for T_{aud} seconds was fed into SoundNet, and the auditory representation extracted from SoundNet was a 1000-dimensional vector.

The auditory and visual representations were concatenated and fed into a feature extractor. The concatenated representations included a natural synchronization feature between the image and audio data because auditory representations have visual knowledge. The feature extractor consisted of fully connected layers. Each task had its own network of tasks and comprised a fully connected layer. The extracted features were fed into each task network. Each task network output a prediction for each task.

4.2. Knowledge distillation

We trained a teacher model using only data with ground truths to apply knowledge distillation. After the teacher model training, the output of the trained teacher model was divided by the temperature parameter t and fed into the softmax activation layer. The softmax layer output was used as the soft label. The student model was trained using ground truths and soft labels to learn dark knowledge from the teacher model.

4.3. Loss function

The loss between the output of the model and the ground truths is computed if the batch data have ground truths for the task. This is called *supervision loss*. If we find no ground truths for the task in the batch data, we compute the loss between the student model output and the teacher model output, called soft labels. This is called *distillation loss*.

Supervision loss We used the concordance correlation coefficient (CCC) loss defined as follows for the valence–arousal estimation task:

$$CCC(y, \hat{y}) = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad (1)$$

where y denotes the ground truths and $\hat{y}$ denotes the predicted values. σ and μ are the standard deviations and the mean calculated for each batch, respectively. The supervision loss for the valence–arousal estimation task was computed through the average of the CCC values of the ground truth and the predicted values of each of the valence and arousal. The supervision loss for the valence–arousal estimation task is computed as follows:

$$\mathcal{L}_{va}^{s} = 1 - \frac{CCC(Y_v), f_v^{tea}(X) + CCC(Y_a), f_a^{tea}(X)}{2} \tag{2}$$

We used cross entropy loss for the expression classification task. Cross entropy is defined as follows:

$$CE(y, \hat{y}) = -\sum_{c=1}^{C} ylog(\hat{y}) \tag{3}$$

C is the number of classes. The supervision loss for the expression classification task is as follows:

$$\mathcal{L}_{expr}^{s} = CE(Y_{expr}, f_{expr}^{tea}(X)) \tag{4}$$

We used binary cross entropy for the AU detection task. Binary cross entropy is defined as follows:

$$BCE(y, \hat{y}) = [ylog(\hat{y}) + (1-y)log(1-\hat{y})] \tag{5}$$

The supervision loss for the action unit detection task is defined as follows:

$$\mathcal{L}_{au}^{s} = BCE(Y_{au}, f_{au}^{tea}(X)) \tag{6}$$

The supervision loss for the MTL task is the sum of each task loss:

$$\mathcal{L}_{mtl}^{s} = \mathcal{L}_{va}^{s} + \mathcal{L}_{expr}^{s} + \mathcal{L}_{au}^{s} \tag{7}$$

Distillation loss The loss between the output of the pretrained teacher model and that of the student model is computed to transfer the dark knowledge of the teacher model to the student model. The distillation loss for each task is obtained as follows.

$$\mathcal{L}_{va}^{d} = 1 - CCC(f_{va}^{stu}(X), f_{va}^{tea}(X)) \tag{8}$$

$$\mathcal{L}_{expr}^{d} = CE(f_{expr}^{stu}(X), f_{expr}^{tea}(X)) \tag{9}$$

$$\mathcal{L}_{au}^{d} = BCE(f_{au}^{stu}(X), f_{au}^{tea}(X)) \tag{10}$$

The MTL task dataset of Aff-Wild2 has annotations for every task; thus, the distillation loss of the MTL task is not computed.

Train loss When training the teacher model, the loss of task *i* is defined using only supervision loss.

$$Loss_i^{tea} = \sum_{n=1}^{b} \mathcal{L}_i^s \tag{11}$$

When training the student model, the loss of task i is defined using supervision loss and distillation loss.

$$Loss_i^{stu} = \sum_{n=1}^{b} \{\gamma_i \cdot (\alpha \cdot \mathcal{L}_i^s + \mathcal{L}_i^d) + \sum_{j \neq i} \beta \cdot \gamma_j \cdot \mathcal{L}_j^d\} \tag{12}$$

where γ is the task weight for each task proposed by Deng et al. [4]. During training, the number of epochs that do not improve the validation performance is counted for each task. The larger the count, the bigger the weight of the loss of work to boost training. The weight of that task loss is $\gamma_i = e^{0.5 n_i}$ when the counted number of i tasks is n_i. α is the hyperparameter of the weight between supervision loss and distillation loss for tasks with a ground truth label. β is a hyperparameter for the weight of distillation loss for tasks without a ground truth label.

4.4. Train procedure

The teacher model was trained to minimize the loss described by Equation (11) using only the ground truth. After the teacher model training, we trained the student model to minimize the loss described by Equation (12) using the teacher model's soft labels and ground truth. Algorithm 1 and Algorithm 2 describe the training procedures for the teacher and student models, respectively.

Algorithm 1 Train teacher model procedure

Require:
parameters: θ_t(teacher)
Epoch: N
Task: $T = [va, expr, au, mtl]$
learning rate: lr

n_{epoch} = 0
while $n_{epoch} < N$ **do**
  while not epoch end **do**
    for $i \in T$ **do**
      $loss = Loss_i^{tea}$
      $\theta_t \leftarrow \theta_t - lr \cdot \frac{\partial loss}{\partial \theta_t}$
    end for
  end while
  $n_{epoch} \leftarrow n_{epoch} + 1$
end while

5. Experiments

Table 1 lists the hyperparameters. We trained for 20 epochs and stopped training if validation performance did

Algorithm 2 Train student model procedure

Require:
parameters: θ_s(student)
Epoch: N
Task: $T = [va, expr, au, mtl]$
learning rate: lr

n_{epoch} = 0
while $n_{epoch} < N$ **do**
  while not epoch end **do**
    for $i \in T$ **do**
      if i is mtl **then**
        $loss = \mathcal{L}^s_{mtl}$
      else
        $loss = Loss^{stu}_i$
      end if
      $\theta_s \leftarrow \theta_s - lr \cdot \frac{\partial loss}{\partial \theta_s}$
    end for
  end while
  $n_{epoch} \leftarrow n_{epoch} + 1$
end while

Hyperparameter		Value
Train	Epochs	20
	Early stop	5
	Optimizer	Adam
	Learning rate	0.0001
	t	2.5
Input data	Batch size	256
	T_{aud}	10
	T_{img}	2
	S	10
	N_{fps}	30
	N_{img}	6
	sr	22050
	C_{aud}	2
	C_{img}	3
	H	112
Loss function	α	10
	β	0.9
Model	Activation function	swish
	Dropout rate	0.5

Table 1. Hyperparameters

not improve for five epochs. We used the Adam optimizer and set the learning rate to 0.0001. The temperature parameter t was set to 2.5. In the input data, the batch size was 256, and T_{aud} was set to 10. Oh et al. [23] experimentally showed that the best value for T_{img} was 2 seconds, and S was 10. Therefore, we set T_{img} to 2 and S to 10. N_{img} was set to 6. In the Aff-Wild2 dataset, N_{fps}, the video frame rate, was 30; sr was 22,050; C_{aud} was 2; C_{img} was 3; and H and W were both 112. In the loss function, α was set to 10, and β was set to 0.9. In the feature extractor and task networks, the number of units in the first layer of the feature extractor layer was set to 1024, while the second layer was set to 512. In task networks, the unit number of the first layer was set to 256, while the second layer was set to 128. A swish activation function and batch normalization were applied. A 0.5 random dropout was applied in the feature extractor.

5.1. Metrics

We used the metric proposed in [13] to evaluate the model performance. The metric for MTL is the sum of metrics for the valence–arousal, expression, and action unit tasks. The valence–arousal metric is the CCC, and the expression recognition metric is the F1 score across all eight categories (i.e., macro F1 score). The metric for action unit detection is the average F1 score over all 12 AUs (i.e., macro F1 score). The number of annotated image data for the MTL task in the Aff-Wild2 database was 172,360, which was significantly smaller than 2,816,832 for the VA task, 2,603,921 for EXPR, and 2,603,921 for AU. For a more reliable evaluation, each task was evaluated while performing the evaluation using annotations in the MTL task.

5.2. Results

An ablation study was performed on the input data to analyze the effect of using auditory representations. We then evaluated the performances of the teacher and student models. We also compared the performance of the Aff-Wild2 test dataset with those of the other teams participating in the 3rd ABAW MTL challenge.

Input data We trained three models to evaluate the model performance according to the input data (i.e., a model that uses only image input, a model that uses only audio input, and a model that uses both audio and image input). Table 2 presents the performance evaluation result of the teacher model according to the input data. The added audio input improved the model performance. The MTL task performance improved by 11.83% and 230.16%, respectively, compared to when visual or auditory representation was used alone. From the perspective of STL performance, the expression classification task affected by sound [4,12,21,25] showed the greatest performance improvement at 9.8%. However, the action units that were not directly affected by sound showed the least performance improvement.

Knowledge distillation Table 3 lists the performance evaluation result of the teacher and student models. In the

Input data	STL				MTL
	AU	EXPR	VA	Total	
I	0.50	0.51	0.61	1.62	1.86
A	0.07	0.12	0.08	0.27	0.63
I+*A*	0.51	0.56	0.64	1.71	2.08

Table 2. Ablation study results of the input data. *I* represents the image input. *A* represents the audio input. STL represents the performance evaluated using each task annotation. MTL represents the performance evaluated using the MTL task annotation.

Model	STL				MTL
	AU	EXPR	VA	Total	
Teacher	0.51	0.56	0.64	1.71	2.08
Student	0.51	0.57	0.63	1.71	2.40

Table 3. Performance evaluation result of the teacher and student models. STL represents the performance evaluated using each task annotation. MTL represents the performance evaluated using the MTL task annotation.

Model	MTL		Rank
	Validation	Test	
ours	**2.40**	**0.95**	-
NISL 2022 [2]	1.66	1.13	1
HSE-NN	-	0.81	3
N.F.V.H. [35]	1.54	0.68	4
baseline [13]	0.30	0.28	-

Table 4. Top performance for the validation and test datasets of our model and the teams that participated in the 3rd ABAW MTL challenge.

student model, the evaluated performance on a single-task data set ground truth (STL) did not significantly improve. On the contrary, that evaluated on the multitasking dataset ground truth (MTL) improved by 15.38%. The dark knowledge transferred by the teacher model helped improve the MTL performance.

Evaluation on the testset We evaluated the test dataset of the Aff-Wild2 MTL task based on previous experimental results using a student model that employed both audio and image input. Table 4 shows the performances of our model and the other models evaluated using the test dataset of the Aff-Wild2 MTL task. Our model achieved a 0.95 performance, which outperformed that of the baseline model [13]. This performance value was 0.18 lower than the first [2] and 0.14 and 0.27 higher than the third and fourth [35], respectively. The performance on the test dataset was 1.45 smaller than that on the validation dataset.

6. Conclusions

In this study, we proposed a multi-task model with auditory–visual synchronized representation and knowledge distillation. Natural synchronized representations were extracted from audio data using SoundNet, while visual representations were extracted from image data using FER model. The representations were concatenated, and predictions were simultaneously performed on the three tasks of VA, EXPR, and AU through the feature extractor and task networks. We also applied knowledge distillation to train on data with incomplete data. Only shallow feature extractors, task networks, and LSTM layers were trained. In summary, we achieved a 0.95 performance on the Aff-Wild2 MTL task test dataset. Our future work will include filling the gap between validation and test performance and fine-tuning the backbone network to improve the performance.

7. Acknowledgments

This research was supported by BK21 Program(5199990814084) through the National Research Foundation of Korea(NRF) funded by the Ministry of Education and Korea Institute of Police Technology (KIPoT) grant funded by the Korea government(KNPA) (No.092021C26S03000, Development of infrastructure information integration and management technologies for real time traffic safety facility operation).

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 1, 2, 3

[2] Didan Deng. Multiple emotion descriptors estimation at the abaw3 challenge. *arXiv preprint arXiv:2203.12845*, 2022. 6

[3] Didan Deng, Zhaokang Chen, and Bertram E Shi. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 592–599. IEEE, 2020. 2

[4] Didan Deng, Liang Wu, and Bertram E Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3557–3566, 2021. 1, 2, 4, 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal,

Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 2

[8] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[10] Fei Jia, Somshubra Majumdar, and Boris Ginsburg. Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6818–6822. IEEE, 2021. 2

[11] Yuma Kajihara, Shoya Dozono, and Nao Tokui. Imaginary soundscape: Cross-modal approach to generate pseudo sound environments. In *Proceedings of the Workshop on Machine Learning for Creativity and Design (NIPS 2017), Long Beach, CA, USA*, pages 1–3, 2017. 3

[12] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019. 5

[13] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 2, 5, 6

[14] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021. 2

[15] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 637–643. IEEE, 2020. 2

[16] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 2

[17] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multitask learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 2

[18] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 2

[19] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 2

[20] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 2

[21] Felix Kuhnke, Lars Rumberg, and Jörn Ostermann. Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE, 2020. 1, 2, 5

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 3

[23] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a past facial image sequence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3550–3556, 2021. 3, 5

[24] Geesung Oh, Junghwan Ryu, Euiseok Jeong, Ji Hyun Yang, Sungwook Hwang, Sangho Lee, and Sejoon Lim. Drer: Deep learning–based driver's real emotion recognizer. *Sensors*, 21(6):2166, 2021. 1, 3

[25] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014. 5

[26] Manh Tu Vu, Marie Beurton-Aimar, and Serge Marchand. Multitask multi-database emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3637–3644, 2021. 2

[27] Lingfeng Wang, Shisen Wang, Jin Qi, and Kenji Suzuki. A multi-task mean teacher for semi-supervised facial affective behavior analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3603–3608, 2021. 2

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[29] Sachihiro Youoku, Yuushi Toyoda, Takahisa Yamamoto, Junya Saito, Ryosuke Kawamura, Xiaoyu Mi, and Kentaro Murase. A multi-term and multi-task analyzing framework for affective analysis in-the-wild. *arXiv preprint arXiv:2009.13885*, 2020. 2

[30] Sachihiro Youoku, Takahisa Yamamoto, Junya Saito, Akiyoshi Uchida, Xiaoyu Mi, Ziqiang Shi, Liu Liu, Zhongling Liu, Osafumi Nakayama, and Kentaro Murase. Multi-modal affect analysis using standardized data within subjects in the wild. *arXiv preprint arXiv:2107.03009*, 2021. 2

[31] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 2

[32] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019. 2

[33] Su Zhang, Yi Ding, Ziquan Wei, and Cuntai Guan. Continuous emotion recognition with audio-visual leader-follower attentive fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3567–3574, 2021. 2

[34] Wei Zhang, Zunhu Guo, Keyu Chen, Lincheng Li, Zhimeng Zhang, and Yu Ding. Prior aided streaming network for multi-task affective recognitionat the 2nd abaw2 competition. *arXiv preprint arXiv:2107.03708*, 2021. 2

[35] Wei Zhang, Zhimeng Zhang, Feng Qiu, Suzhen Wang, Bowen Ma, Hao Zeng, Rudong An, and Yu Ding. Transformer-based multimodal information fusion for facial expression analysis. *arXiv preprint arXiv:2203.12367*, 2022. 2, 6