Ensemble of Multi-task Learning Networks for Facial Expression Recognition In-the-Wild with Learning from Synthetic Data

Jae-Yeop Jeong[1], Yeong-Gi Hong[1], Sumin Hong[2], JiYeon Oh[2], Yuchul Jung[3], Sang-Ho Kim[3], and Jin-Woo Jeong[1(✉)]

[1] Department of Data Science, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
{jaey.jeong,yghong,jinw.jeong}@seoultech.ac.kr
[2] Department of Industrial Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
{17101992,dhwldus0906}@seoultech.ac.kr
[3] Kumoh National Institute of Technology, Gumi 39177, Republic of Korea
{jyc,kimsh}@kumoh.ac.kr

Abstract. Facial expression recognition in-the-wild is essential for various interactive computing applications. Especially, "Learning from Synthetic Data" is an important topic in the facial expression recognition task. In this paper, we propose a multi-task learning-based facial expression recognition approach where emotion and appearance perspectives of facial images are jointly learned. We also present our experimental results on validation and test set of the LSD challenge introduced in the 4th affective behavior analysis in-the-wild competition. Our method achieved the mean F1 score of 71.82 on the validation and 35.87 on the test set, ranking third place on the final leaderboard.

Keywords: Facial expression recognition · Leaning from synthetic data · Multi-task learning · Ensemble approach

1 Introduction

In the affective computing domain, understanding and prediction of natural human behaviors, such as gaze, speech, and facial expression are essential for more efficient and accurate affective behavior analysis. Advances in this technology will facilitate the development of practical real-world applications in the field of human-computer interaction, social robots, and a medical treatment [32]. Among various modalities to investigate the human being's affective states, facial expression has been considered one of the most promising and practical channels.

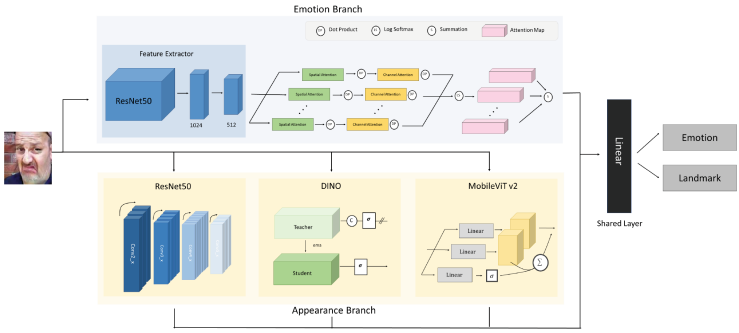
J.-Y. Jeong, Y.-G. Hong, S. Hong and J. Oh—Contributed equally to this work.

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13806, pp. 60–75, 2023.
https://doi.org/10.1007/978-3-031-25075-0_5

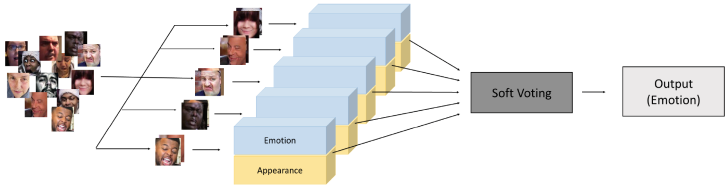
To achieve a more robust and accurate affective behavior analysis, a number of studies have been proposed in recent years [5,18,24,25,29,42] in facial expression recognition (FER). However, there are still many rooms to improve the robustness and performance of facial expression recognition techniques. One of the most challenging research areas is facial expression recognition in-the-wild. Generally, it is well known that a number of well-aligned high-resolution face images are necessary to have high performance in facial expression recognition. Compared to face images gathered in a controlled setting, however, in-the-wild face images have much more diversity in terms of visual appearance, such as head pose, illumination, and noise, etc. Moreover, gathering large-scale in-the-wild facial images is much more time-consuming. Therefore, facial expression recognition in-the-wild is much more challenging but should be addressed thoroughly for realizing practical FER-based applications. Meanwhile, face image generation/synthesis for facial expression recognition tasks has been steadily getting much attention as one of the promising techniques to address this problem, because it can generate unlimited photo-realistic facial images with various expressions and conditions [1,13,49]. By learning from synthetic data, the problem of collecting large-scale facial images in-the-wild would be mitigated, thereby accelerating the development of real-world applications.

The 4th competition on Affective Behavior Analysis in-the-wild (ABAW), held in conjunction with the European Conference on Computer Vision (ECCV) 2022 [20], is a continuation of the 3rd Workshop and Competition on Affective Behavior Analysis in-the-wild in CVPR 2022 [21]. The ABAW competition contributes to the deployment of in-the-wild affective behavior analysis systems that are robust to video recording conditions, diversity of contexts and timing of display, regardless of human age, gender, ethnicity, and status. The 4th ABAW competition is based on the Aff-Wild2 database [27], which is an extension of the Aff-wild database [26,48] and consists of the following tracks: 1) Multi-Task-Learning (MTL) and 2) Learning from Synthetic Data (LSD).

In this paper, we describe our method for the LSD challenge and present our results on the validation and test set. For the LSD challenge, some frames from the Aff-Wild2 database [27] were selected by the competition organizers and then used to generate artificial face images with various facial expressions [22,23,28]. In total, the synthetic image set consists of approximately 300K images and their corresponding annotations for 6 basic facial expressions (anger, disgust, fear, happiness, sadness, surprise), which will be used in model training/methodology development. In this LSD challenge, participating teams were allowed to use only the provided synthetic facial images when developing their methodology, while any kind of pre-trained model could be used unless it was not been trained on the Aff-Wild2 database. For validation, a set of original facial images of the subjects who also appeared in the training set was provided. For evaluation, the original facial images of the subjects in the Aff-Wild2 database test set, who did not appear in the given training set, are used. For the LSD challenge, the mean F1 score across all 6 categories was used as a metric.



(a) Overall network architecture



(b) Ensemble architecture

Fig. 1. Overview of the proposed architecture

2 Method

To achieve high performance on the task of facial expression recognition, extraction of the robust feature from facial images is essential. To this end, we propose a multi-task learning approach, jointly optimizing different multiple learning objectives. Figure 1a depicts an overview of the proposed architecture used in our study. As shown in Fig. 1a, the framework was designed to solve both facial expression recognition task (i.e., Emotion branch) and face landmark detection task (i.e., Appearance branch) using the provided synthetic training data. Figure 2 shows examples of the provided training images for the LSD track of the 4th ABAW competition.

Based on this network architecture, we adopted an ensemble approach when computing the final predictions to achieve a more generalized performance for unseen data. Figure 1b describes our ensemble approach. Each model with a different configuration is trained with sub-sampled data sets (i.e., 20% out of the entire set) and produces its own output. Finally, we aggregate the probabilities of each model through soft voting for the final prediction. More details on the CNN models used in each branch can be found in Sect. 2.2.

2.1 Data Pre-processing

In this section, we describe the data pre-processing steps used in our study. The diversity of training data largely affects the performance of a classification model

Fig. 2. Synthetic training data

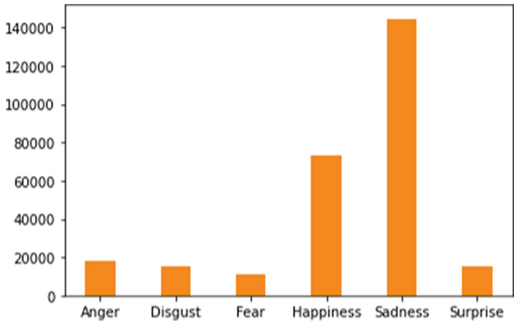


Fig. 3. Training data statistics

Fig. 4. Example of face landmark annotations

in terms of robustness and generalization. However, as shown in Fig. 3, the provided training data set has a problem of imbalance data distribution. For example, the number of images in the "Sadness" class is about 13x bigger than that of the "Fear" class (i.e., 144,631 vs 10,923). Furthermore, subject-wise data distribution in each emotional class is not balanced as well. Accordingly, we adopted two strategies for model training to overcome the aforementioned limitations: 1) a multi-task learning approach to utilize more diverse image/feature representations and 2) data augmentation techniques to compensate for the visual diversity of the original training images.

Database. In order to train our multi-task learning framework, each training image is given 1) facial expression annotation and 2) face landmark annotation. First, the facial expression label consists of 6 basic emotional categories (i.e.,

anger, disgust, fear, happiness, sadness, surprise), which were offered by the 4th ABAW competition organizers. Second, the landmark annotations were obtained through the DECA framework, which is a state-of-the-art on the task of 3D shape reconstruction from in-the-wild facial images [11]. The DECA framework was originally designed to tackle the problem of wrinkle modeling by 3D head reconstruction with detailed facial geometry from a single input image. Before the model reconstructs 3D head model, it predicts 2D face landmark data in order to learn eye closure landmarks, and we can use it as our appearance annotations. Figure 4 shows our example of landmark annotation which is composed of 68 coordinate points with (x, y) of a face.

Data Augmentation. It is well known that data augmentation techniques have a huge influence on the performance of deep learning models [43]. There exist various data augmentation methods based on affine transformation, which have already shown their effectiveness to help deep learning models achieve more robust performance, but a number of advanced data augmentation studies have still been reported [9,47,50]. Among them, we apply mix-up [50] method and one of the mix-up variants called mix-augment [39] method as our main augmentation strategy since they have already been validated in several computer vision studies [3,44]. Mix-up [50] is a data augmentation method generating new samples by mixing two raw original images. The result of Mix-up augmentation is a virtual image $\widetilde{x}$ in which the two original images are interpolated. The image is then given a new label, $\widetilde{y}$, represented in the form of label smoothing. An example of Mix-up augmentation on facial images with "sad" and "fear" emotional classes is presented in Fig. 5. Then, formulation of Mix-up augmentation can be represented as:

$$\widetilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\widetilde{y} = \lambda y_i + (1 - \lambda) y_j$$

where x_i and $x_j \in X$ are two random sampled images in a batch, y_i and $y_j \in Y$ are their labels, and $\lambda \sim B(\alpha, \alpha) \in [0, 1]$ is a parameter used for reconciling images i and $j \in N$ (batch size) [50]. The loss function for Mix-up augmentation is the categorical cross entropy(CCE) of two labels ($\bar{y}$ and $\widetilde{y}$), defined as [50] :

$$\mathcal{L}_{CCE}^{v} = \mathop{\mathbb{E}}_{\bar{y},\widetilde{y}}[-\widetilde{y} \cdot \log \bar{y}]$$

where $\bar{y}$ represents the predicted output for $\widetilde{x}$.

Facial expression recognition in-the-wild has unique characteristics compared to the existing FER problem targeting static, well-aligned facial images, because in-the-wild facial images have various angles, head poses, and illumination, and so on. In such an environment, mix-up interpolation can hinder the training of deep learning models [39]. To tackle this issue, one of the variants of Mix-up, called Mix-augment [39], was proposed. The Mix-augment method is a Mix-up-based data augmentation technique specially designed for facial expression recognition in-the-wild. The difference between Mix-up and Mix-augment is that

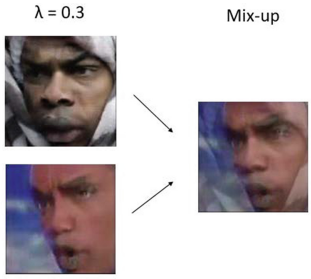


Fig. 5. Mix-up example using synthetic facial images

Mix-augment computes the cross-entropy loss for two original sampled images (x_i,x_j) in addition to the original mix-up loss. The Mix-augment loss is therefore defined as:

$$\mathcal{L}_{Ma} = \mathcal{L}_{CCE}^{v} + \mathcal{L}_{CCE}^{r_i} + \mathcal{L}_{CCE}^{r_j}$$

where $\mathcal{L}_{CCE}^{v}$ is the Mix-up loss for v which is a virtual sample generated with x_i and x_j, and $\mathcal{L}_{CCE}^{r_{i,j}}$ is the CCE loss for i/j-th real images [39].

2.2 Model Architecture

As depicted in Fig. 1, our architecture is composed of two branches: 1) emotion and 2) appearance. In each branch, we utilize a pre-trained backbone for extraction of more robust and generalized features. Finally, we employ a series of shared fully connected layers right after two branches to exploit all knowledge extracted from different kinds of learning tasks. In addition, we employ an ensemble approach to our framework for more robustness.

Emotion Branch. As depicted in the emotion branch of Fig. 1a, we adopt a deep learning-based facial expression recognition approach called "DAN" [46] which is a state-of-the-art method on the AffectNet database [37]. The DAN architecture consists of two components: Feature Clustering Networks (FCN) and attention parts.

First, a series of facial images are fed to the FCN module for feature extraction. During the feature extraction process, a new loss function called affinity loss supervises to maximize the inter-class margin and minimize the intra-class margin [46]. In simple terms, the features belonging to the same class are refined to become closer to each other, otherwise, far away from each other. The affinity loss is presented as:

$$\mathcal{L}_{af} = \frac{\sum_{i=1}^{M} \left\| x_i^{'} - c_{y_i} \right\|_2^2}{\sigma_c^2}$$

where $x_i \in X$ in input feature is i-th input vector and $y_i \in Y$ in target space is the target, $c \in R^{m \times d}$ is a class center, m is the dimension of Y, d is the dimension of class centers, and σ_c is the standard deviation between class centers [46].

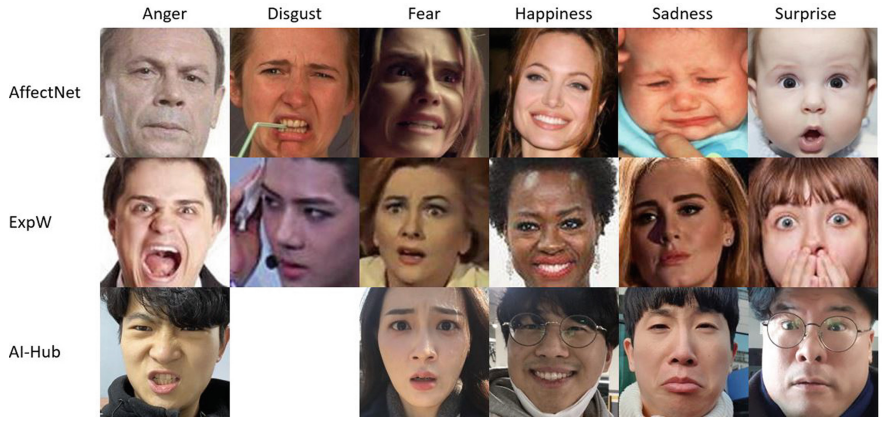


Fig. 6. Example images of each data set used for pre-training

The attention part consists of Multi-head cross Attention Network (MAN) and Attention Fusion Network (AFN). The MAN module consists of a combination of parallel cross-head attention units which include spatial and channel attention units. The AFN module merges the attention maps from the MAN module in an orchestrated fashion without overlapping, using the partition loss which is defined as:

$$\mathcal{L}_{pt} = \frac{1}{NC} \sum_{i=1}^{N} \sum_{j=1}^{C} \log(1 + \frac{k}{\sigma_{ij}^2})$$

k is the number of cross-attention heads, C is the channel size of the previous attention maps, N is the number of samples, and σ_{ij}^2 is the variance of the j-the channel on i-th feature sample [46].

The DAN architecture used in our emotion branch was pre-trained on the following data sets; AffectNet [37], Expw [51], and AIHUB data set [2]. The example images included in the pre-training data sets are shown in Fig. 6. For better performance, we replaced the original feature extractor of the DAN architecture (i.e., ResNet18 [16] pre-trained on MS-Celeb-1M [15]) with ResNet50 [16] pre-trained on VGGFace2 [7], as presented in [19]. Furthermore, to prevent overfitting and obtain a high generalization performance, we applied various data augmentation techniques in the emotion learning branch, such as Horizontal flip, RandomErasing, RandomAffine, Mix-Up, and Mix-Augment [39].

Appearance Branch. The goal of the appearance branch is to extract the robust feature in terms of visual appearance of facial images by solving a landmark detection task. For this, we employed various pre-trained models in the appearance branch and analyzed overall performance according to the network configuration. We selected several popular models pre-trained on various large-scale data sets through 1) supervised learning and 2) self-supervised learning (SSL) mechanisms. First, we employed a task-specific backbone related to face

information, hence, the ResNet50 [16] model pre-trained on VGGFace2 [7] data set which is designed to support a face recognition task. Second, we considered a backbone model pre-trained on the general purpose large-scale image data set (i.e., ImageNet [30]), which has been leveraged in a variety of tasks (e.g., object detection [6,17] and segmentation [12,40]). For this option, we used a lightweight vision transformer architecture called MobileViTv2 [35]. MobileVITv2 is a hybrid network that combines the advantages of CNN and vision transformers, which has similar or fewer parameters than existing lightweight CNN architectures, but it is still more effective in terms of classification accuracy. Next, in the case of the SSL-based model, we selected ResNet50 pre-trained on VGGFace [38] data set, which was learned through one of the latest SSL frameworks called DINO [8]. In summary, our candidate backbones for the appearance branch are classified into: a) ResNet50 [16] pre-trained on VGGFace2 [7], b) DINO ResNet50 [8] pre-trained on VGGFace [38], and c) MobileVITv2 [35] pre-trained on ImageNet [30]. As shown in the appearance branch of Fig. 1, we train only a single backbone for learning appearance features, rather than utilizing multiple backbones together. Similar to the emotion branch, we also applied a data augmentation strategy in the appearance branch. Due to the characteristics of landmark data, however, we applied Colorjitter only to prevent unnecessary spatial transformations.

In summary, our multi-task learning model is configured with the emotion branch based on DAN and the appearance branch with one of the following backbones: a) VGGFace2 ResNet50, or b) DINO ResNet50, or c) MobileViTv2.

Multi-task Learning. After feature extraction is done from each branch, all the feature vectors are incorporated into the shared fully connected layers to share the knowledge from both branches. Finally, our multi-task learning framework produces two outputs for FER and landmark detection tasks, y_{expr} and y_{land} respectively. The loss functions for y_{expr} and y_{land} are defined as follows. First, the loss for the emotion branch $\mathcal{L}_{Em}$ is basically based on DAN (i.e., affinity loss and partition loss) along with Mix-augment or Mix-up loss, which is defined as:

$$\mathcal{L}_{Em} = \begin{cases} \mathcal{L}_{Ma} + \mathcal{L}_{af} + \mathcal{L}_{pt}, & \text{if mix-aug used} \\ \mathcal{L}^{v}_{CCE} + \mathcal{L}_{af} + \mathcal{L}_{pt}, & \text{if mix-up used} \end{cases}$$

Second, the loss for the appearance branch $\mathcal{L}_{Ap}$ is the Mean square error (MSE) to compare the difference between the ground truth landmark points and predicted landmark points. Finally, the total loss for our multi-task learning architecture is defined as the sum of the appearance loss ($\mathcal{L}_{Ap}$) and the emotion loss ($\mathcal{L}_{Em}$):

$$\mathcal{L}_{Total} = \mathcal{L}_{Em} + \mathcal{L}_{Ap}$$

During the inference, only the output from the emotion branch is used for classification of facial expression for the given validation/test sample.

Table 1. Hyperparameter setting

Hyper-parameter	Value
Batch Size	128
Optimizer	ADAM
Learning Rate	$1e^{-4}$
Learning Rate Scheduler	Exponential Decay
Epochs	12
Optimizer Weight Decay	$1e^{-4}$
Number of Cross Attention Head	4
α of Mix-up and Mix-aug	0.1

Ensemble Approach. In general, an ensemble approach shows better performance than an individual model in terms of accuracy, robustness, and generalization by aggregating multiple models' outputs [10]. The ensemble approach we used is a bagging approach [4] in which each (weak) classifier is trained with sub-sampled training data. As can be seen from Fig. 1b, the weak classifiers are trained with a subset of training data so that they can learn particular representations by observing a different part of training data. In other words, each classifier has a different decision-making capability in terms of classification. By aggregating the final probability from each weak classifier, the final result can be more accurate, robust and generalizable. In this work, we used a soft voting approach (i.e. probability-based voting) to combine the predictions from each model.

3 Experiments and Results

In this section, we describe the results of our experiments. All experiments were performed on a high-end server equipped with six NVIDIA RTX 3090 GPUs, 128Gb RAM, and an Intel i9-10940X CPU. We conducted model training, validation, and evaluation with Pytorch framework.

3.1 Training Setup

Our model was trained for 12 epochs with batch size of 128 and we used ADAM optimizer with a learning rate of 0.0001. We adopted the Exponential decay learning rate scheduler to prevent overfitting. Also, the number of cross-head attention heads for the DAN model in the emotion branch was set to 4. We set the hyper-parameter α of mix-up and mix-aug methods to 0.1. For more details of our training setup, refer to Table 1.

Table 2. F1 scores of individual models on the validation set

Method	Emotion	Appearance	Basic Aug	F1(%)	Mean F1(%)
Mix-aug	DAN(ResNet50)	VGGFace2(ResNet50)	O	67.30	**68.71**
			X	**68.66**	
		DINO(ResNet50)	O	**69.58**	
			X	69.30	
Mix-up		VGGFace2(ResNet50)	O	**69.51**	68.46
			X	67.28	
		DINO(ResNet50)	O	**70.57**	
			X	66.46	

3.2 Performance Evaluation

To validate the best model configuration for the LSD task, we trained our multi-task learning model with different settings and evaluated their performances on the validation and test set. In this section, we describe the performance of our individual models first and then explain the results of our ensemble approach.

Evaluation on Validation Set. We explain how well each individual model trained with the entire training data set works on the official validation set first. As mentioned above, we applied a set of basic augmentations (i.e., Colorjitter, Horizontal flip, and RandomErasing) along with Mix-augment or Mix-up augmentation methods in the emotion branch. Table 2 reveals the performance difference between the model configurations according to the use of data augmentation methods (models with MobileVitV2-based appearance branch was not tested due to some technical issues). In the case of the methods with Mix-augment, there was no consistent pattern found in the performance change according to the use of basic augmentation methods. Specifically, the use of basic augmentation skills was not useful for the performance of DAN with VGGFace2(ResNet50) model. Rather, without basic augmentation methods, its performance increased from 67.30 to 68.66 in terms of F1 score. On the contrary, DAN with DINO(ResNet50) with basic augmentations showed a slightly higher performance than the one without basic augmentations. However, it should be noted that both models yielded robust performance in terms of F1-score (i.e., 69.58 and 69.30). On the other hand, as shown in the second row of Table 2, the performance of the models with Mix-up augmentation was significantly improved by the use of basic data augmentation methods. Specifically, DAN with VGGFace2(ResNet50) achieved a performance improvement of 2.23%p (from 67.28 to 69.51) while DAN with DINO(ResNet50) has a performance gain of 4.11%p (from 66.46 to 70.57). In summary, it was found that the mean F1-score of the models with Mix-augment method (68.71) is slightly higher (0.3%p) than that of the models with Mix-up method (68.46). The author of Mix-augment [39] argued that Mix-augment can provide more robust performance than Mix-

Table 3. F1 scores of weak models on the validation set

Method	Emotion	Appearance	F1(%)	Mean F1(%)
Mix-aug	DAN(ResNet50)	VGGFace2(ResNet50)	**70.08**	**69.65**
		DINO(ResNet50)	69.19	
		MobileViTv2	69.66	
Mix-up		VGGFace2(ResNet50)	68.72	69.20
		DINO(ResNet50)	68.42	
		MobileViTv2	**70.46**	

up in the task of facial expression recognition in the wild. The findings from our study seem to be consistent with the argument from [39], even though the best F1-score was reported from the model with Mix-up augmentation. We expect this difference came from the difference in the model architecture used, target data set, and hyper-parameter settings.

Second, we discuss the performance of our multi-task learning framework trained with sub-sampled training data. Hereafter, we denote these models as "weak" models while the models trained with the entire training data set as "strong" ones. Table 3 summarizes the performance of the weak models with different configurations. All the weak models used basic augmentations by default. Here, "F1" denotes the average F1 score of five weak models trained with a particular network configuration. The mean F1 score in Table 3 indicates the average scores of all the weak models trained with Mix-aug or Mix-up method. Compared with the performance of individual strong models (i.e., trained with a whole data set), the weak models demonstrated highly competitive performances even though they are trained with sub-sampled data only. Specifically, the mean F1 score of weak models with Mix-aug and Mix-up is 69.65% and 69.20%, respectively, higher than that from the strong models. In the case of Mix-aug-based configuration, DAN with VGGFace2(ResNet50) scored the best with an F1 score of 70.08 which also outperforms all the strong models trained with the Mix-aug method. On the contrary, DAN with MobileViTv2 scored the best with an F1 score of 70.46, in the case of Mix-up-based configuration. To sum up, the best-resulting model configuration among weak models is DAN with MobileViTv2 trained with Mix-up method.

Finally, from these results, we designed five ensemble configurations for the final performance evaluation as follows:

- The first configuration ("MU MobileViTv2 Weak") is an ensemble of weak models of DAN with MobileViTv2 using Mix-up augmentation which has shown the best performance among the weak models.
- The second configuration ("MU Strong & MU Weak") denotes an ensemble of two best-performing strong models and all weak models trained with Mix-up method. We assumed that strong models and weak models can compensate for each other.

Table 4. F1 scores of Ensemble approach on validation and test set ("MU" denotes "Mix-up", "MA" denotes "Mix-aug", "Ensemble" denotes the result from the aggregation of models we trained)

Ensemble Method	Val. F1(%)	Test. F1(%)
MU MobileViTv2 Weak	70.39	33.74
MU Strong & MU Weak	69.91	34.96
MA Strong & MA Best Weak	**71.82**	**35.87**
MU & MA Best Weak	71.33	34.93
Best Strong & MA VGG Weak & MU MobileViTv2 Weak	71.65	35.84
Baseline	50.0	30.0

- The third configuration ("MA Strong & MA Best Weak") is an ensemble of two best-performing strong models and best performing weak models trained with Mix-aug method.
- The fourth configuration ("MU & MA Best Weak") is an ensemble of best-scoring weak models trained with Mix-up and Mix-aug methods.
- The last configuration ("Best Strong & MA VGG Weak & MU MobileViTv2 Weak") consists of best-performing strong models and weak models of both VGGFace2(ResNet50) with Mix-aug and MobileViTv2 with Mix-up.

Table 4 shows the configuration and the performance of our final ensemble models on the validation and test set. Through the experiment, we found the following interesting results. First, all the ensemble configurations on the validation set recorded higher performance than the baseline (0.50), which shows the feasibility of the proposed multi-task network-based ensemble approach for the LSD task. Second, the third ensemble configuration (i.e.,"MA Strong & MA Best Weak") that only included Mix-aug-based models yielded the best F1-score of 71.82, which shows the superior performance of the Mix-augment method for the LSD task in particular. Finally, all ensemble configurations containing Mix-aug-based models showed relatively high performances (i.e., over 71% F1 score). Moreover, these models even outperformed the best-performing individual strong model (i.e., DAN with DINO(ResNet50) which resulted in the F1-score of 70.57). In contrast, ensemble configurations with Mix-up-based models did not work well. Specifically, the first and second ensemble configurations composed of Mix-up-based models only even performed worse than some of individual models.

Evaluation on Test Set. In this section, we discuss the performance of our framework on the test set of the LSD challenge. Table 5 summarizes the final leader-board for the Learning from Synthetic Data (LSD) challenge of the 4th ABAW 2022 competition. As described in the first row in Table 5, the baseline performance is an F1 score of 30%, and only the participating teams with a test score higher than the baseline were listed on the final leader-board. It can be found that the performance of the baseline decreased drastically when it was evaluated on the test set (from 50.0 on the val-set to 30.0 on the test-set),

Table 5. Highest F1 scores on the test set

Method	F1(%)
Baseline	30
IMLAB [31]	30.84
USTC-AC [36]	30.92
STAR-2022 [45]	32.4
SSSIHL-DMACS [14]	33.64
SZTU-CVGroup [34]	34.32
HUST-ANT [33]	34.83
ICT-VIPL	34.83
PPAA	36.51
HSE-NN [41]	37.18
Ours (MTL Ensemble)	**35.87**

which reveals how challenging the LSD task is. The winner of the LSD challenge, team HSE-NN [41], brought out the mean F1 score of 37.18%. The runner-up, team PPAA, achieved an F1 score of 36.51%. Our method ranked third with an F1 score of 35.87%, following team HSE-NN and team PPAA. Our best model configuration was "MA Strong & MA Best Weak" configuration, as listed in the 3rd row of Table 4. It worked the best not only on the validation set (71.82) but also test set (35.87). Similar to the baseline, we also experienced a significant amount of performance drop when evaluating on the test set. However, we could demonstrate that our approach is promising and has more potential to address the challenging problem called "Learning from Synthetic Data".

More details on the test result of our ensemble configurations can be found from Table 4. Similar to the validation result, all the proposed ensemble configurations outperformed the baseline method (30.0) on the test set. Most of the configurations showed similar patterns on the test set as well, except that the ensemble configuration including only Mix-up-based models (i.e., the second configuration) worked better on the test set.

4 Conclusion

In this paper, we proposed a multi-task learning-based architecture for facial expression recognition in-the-wild and presented the results for the LSD challenge of the 4th ABAW competition. Furthermore, we designed a multi-task learning pipeline to solve facial expression recognition and face landmark detection tasks for learning more robust and representative facial features. Also, we augmented the training data with image mixing-based data augmentation methods, specifically, Mix-up and Mix-augment algorithms. Finally, our method produced the mean F1 score of 71.82% and 35.87% on the validation set and test set, respectively, resulting in 3rd place in the competition. Our future work will include the use of generative approaches and the development of a more efficient multi-task learning pipeline to achieve better classification performance.

Acknowledgement. This work was supported by the NRF grant funded by the Korea government (MSIT) (No.2021R1F1A1059665), by the Basic Research Program through the NRF grant funded by the Korea Government (MSIT) (No.2020R1A4A1017775), and by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (P0017123, The Competency Development Program for Industry Specialist).

References

1. Abbasnejad, I., Sridharan, S., Nguyen, D., Denman, S., Fookes, C., Lucey, S.: Using synthetic data to improve facial expression analysis with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1609–1618 (2017)
2. AI-Hub: Video dataset for korean facial expression recognition. Available at https://bit.ly/3ODKQNj. Accessed 21 Jul 2022
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
5. Canedo, D., Neves, A.J.: Facial expression recognition using computer vision: a systematic review. Appl. Sci. **9**(21), 4678 (2019)
6. Cao, J., Cholakkal, H., Anwer, R.M., Khan, F.S., Pang, Y., Shao, L.: D2Det: towards high quality object detection and instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11485–11494 (2020)
7. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
8. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
9. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: RandAugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
10. Deng, L., Platt, J.: Ensemble deep learning for speech recognition. In: Proceedings of Interspeech (2014)
11. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. ACM Trans. Graph. (ToG) **40**(4), 1–13 (2021)
12. Fu, J., Liu, J., Jiang, J., Li, Y., Bao, Y., Lu, H.: Scene segmentation with dual relation-aware attention network. IEEE Trans. Neural Netw. Learn. Syst. **32**(6), 2547–2560 (2020)
13. Gao, H., Ogawara, K.: Face alignment using a GAN-based photorealistic synthetic dataset. In: 2022 7th International Conference on Control and Robotics Engineering (ICCRE), pp. 147–151. IEEE (2022)
14. Gera, D., Kumar, B.N.S., Kumar, B.V.R., Balasubramanian, S.: SS-MFAR : semi-supervised multi-task facial affect recognition. arXiv preprint arXiv:2207.09012 (2022)
15. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 87–102. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_6

16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
17. Hu, J., et al.: ISTR: end-to-end instance segmentation with transformers. arXiv preprint arXiv:2105.00637 (2021)
18. Huang, Y., Chen, F., Lv, S., Wang, X.: Facial expression recognition: a survey. Symmetry **11**(10), 1189 (2019)
19. Jeong, J.Y., Hong, Y.G., Kim, D., Jeong, J.W., Jung, Y., Kim, S.H.: Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2353–2358 (2022)
20. Kollias, D.: ABAW: learning from synthetic data & multi-task learning challenges. arXiv preprint arXiv:2207.01138 (2022)
21. Kollias, D.: Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2328–2336 (2022)
22. Kollias, D., Cheng, S., Pantic, M., Zafeiriou, S.: Photorealistic facial synthesis in the dimensional affect space. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
23. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.: Deep neural network augmentation: generating faces for affect analysis. Int. J. Comput. Vis. **128**(5), 1455–1484 (2020)
24. Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.: Recognition of affect in the wild using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 26–33 (2017)
25. Kollias, D., Sharmanska, V., Zafeiriou, S.: Distribution matching for heterogeneous multi-task learning: a large-scale face study. arXiv preprint arXiv:2105.03790 (2021)
26. Kollias, D., et al.: Deep affect prediction in-the-wild: aff-wild database and challenge, deep architectures, and beyond. Int. J. Comput. Vis. **127**(6), 907–929 (2019)
27. Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: aff-wild2, multi-task learning and arcface. arXiv preprint arXiv:1910.04855 (2019)
28. Kollias, D., Zafeiriou, S.: VA-StarGAN: continuous affect generation. In: Blanc-Talon, J., Delmas, P., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2020. LNCS, vol. 12002, pp. 227–238. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-40605-9_20
29. Kollias, D., Zafeiriou, S.: Affect analysis in-the-wild: valence-arousal, expressions, action units and a unified framework. arXiv preprint arXiv:2103.15792 (2021)
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems **25** (2012)
31. Lee, H., Lim, H., Lim, S.: BYEL : bootstrap on your emotion latent. arXiv preprint arXiv:2207.10003 (2022)
32. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **13**, 1195–1215 (2020)
33. Li, S., et al.: Facial affect analysis: Learning from synthetic data & multi-task learning challenges. arXiv preprint arXiv:2207.09748 (2022)
34. Mao, S., Li, X., Chen, J., Peng, X.: Au-supervised convolutional vision transformers for synthetic facial expression recognition. arXiv preprint arXiv:2207.09777 (2022)
35. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)

36. Miao, X., Wang, J., Chang, Y., Wu, Y., Wang, S.: Hand-assisted expression recognition method from synthetic images at the fourth ABAW challenge. arXiv preprint arXiv:2207.09661 (2022)
37. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)
38. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
39. Psaroudakis, A., Kollias, D.: Mixaugment & mixup: Augmentation methods for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 2367–2375 (June 2022)
40. Rossi, L., Karimi, A., Prati, A.: Recursively refined R-CNN: instance segmentation with self-RoI rebalancing. In: Tsapatsoulis, N., Panayides, A., Theocharides, T., Lanitis, A., Pattichis, C., Vento, M. (eds.) CAIP 2021. LNCS, vol. 13052, pp. 476–486. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-89128-2_46
41. Savchenko, A.V.: HSE-NN team at the 4th ABAW competition: Multi-task emotion recognition and learning from synthetic images. arXiv preprint arXiv:2207.09508 (2022)
42. Savchenko, A.V., Savchenko, L.V., Makarov, I.: Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Trans. Affect. Comput. **13**, 2132–2143 (2022)
43. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
44. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: improved calibration and predictive uncertainty for deep neural networks. In: Advances in Neural Information Processing Systems **32** (2019)
45. Wang, L., Li, H., Liu, C.: Hybrid CNN-transformer model for facial affect recognition in the ABAW4 challenge. arXiv preprint arXiv:2207.10201 (2022)
46. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract your attention: multi-head cross attention network for facial expression recognition. arXiv preprint arXiv:2109.07270 (2021)
47. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
48. Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-Wild: valence and arousal'in-the-wild'challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34–41 (2017)
49. Zeng, J., Shan, S., Chen, X.: Facial expression recognition with inconsistently annotated datasets. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
51. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: From facial expression recognition to interpersonal relation prediction. Int. J. Comput. Vis. **126**(5), 550–569 (2017). https://doi.org/10.1007/s11263-017-1055-1