applied sciences

MDPI

Article

Lightweight Micro-Expression Recognition on Composite Database

Nur Aishah Ab Razak [1,*] and Shahnorbanun Sahran [2,*]

[1] Faculty of Technology, Science and Information, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[2] Center for Artificial Intelligence Technology (CAIT), Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
* Correspondence: nuraishahrazak@yahoo.com (N.A.A.R.); shahnorbanun@ukm.edu.my (S.S.)

Abstract: The potential of leveraging micro-expression in various areas such as security, health care and education has intensified interests in this area. Unlike facial expression, micro-expression is subtle and occurs rapidly, making it imperceptible. Micro-expression recognition (MER) on composite dataset following Micro-Expression Grand Challenge 2019 protocol is an ongoing research area with challenges stemming from demographic variety of the samples as well as small and imbalanced dataset. However, most micro-expression recognition (MER) approaches today are complex and require computationally expensive pre-processing but result in average performance. This work will demonstrate how transfer learning from a larger and varied macro-expression database (FER 2013) in a lightweight deep learning network before fine-tuning on the composite dataset can achieve high MER performance using only static images as input. The imbalanced dataset problem is redefined as an algorithm tuning problem instead of data engineering and generation problem to lighten the pre-processing steps. The proposed MER model is developed from truncated EfficientNet-B0 model consisting of 15 layers with only 867k parameters. A simple algorithm tuning that manipulates the loss function to place more importance on minority classes is suggested to deal with the imbalanced dataset. Experimental results using Leave-One-Subject-Out cross-validation on the composite dataset show substantial performance increase compared to the state-of-the-art models.

Keywords: micro-expression recognition; composite dataset; EfficientNet; transfer learning; lightweight deep learning network

check for updates

Citation: Ab Razak, N.A.; Sahran, S. Lightweight Micro-Expression Recognition on Composite Database. *Appl. Sci.* **2023**, *13*, 1846. https://doi.org/10.3390/app13031846

Academic Editor: Evaggelos Spyrou

Received: 16 December 2022
Revised: 14 January 2023
Accepted: 21 January 2023
Published: 31 January 2023

1. Introduction

Micro-expressions are subtle, spontaneous and quick facial expressions with duration less than 0.5 seconds and involving only certain face muscles [1]. They occur when a person tries to hide or withhold their emotions from being expressed, which suggests genuine hidden emotions rather than normal facial expressions or verbal signals [2]. Micro-expression recognition (MER) aims to automatically identify the hidden emotions such as Happiness, Sadness or Disgust from the micro-expressions. MER has potential usage in a variety of fields such as security, education and public health. However, the quick and delicate nature of micro-expressions makes it challenging to recognize them. For instance, US coast guards who were trained to recognized micro-expressions using the Micro-Expression Training Tool (METT) pioneered by Ekman achieved no more than 50% accuracy [3].

Earlier works on Micro-Expression Recognition (MER) were more focused on crafting features that can best represent the subtle muscle movements in micro-expressions such as variants of Local Binary Patterns (LBP) [4,5] and Histogram of Gradients (HOG) [6]. To increase MER accuracy, many researchers then turn to optical flow and optical strain features with the aim of including temporal-based information from the quick facial movements in micro-expressions [7,8]. However, these features require pre-extraction either manually or automatically which increases MER complexity, and as such are not practical for a real-life recognition system.

Along with the success of deep learning methods' applications in Facial Expression Recognition (FER), trends in MER are now geared more towards deep learning approaches, yet most of the implementation in MER utilizes OF features [9–13] making them more complex. Similarly, while [14] proposed an end-to-end micro-expression spotting and recognition system, it requires an additional synthetic OF generation step to augment the training dataset. Other implementations using Dynamic Images (DI) [15–17] and facial graph [18] required pre-constructing the input features.

Due to the insufficient samples in the micro-expression databases for deep learning methods, many have employed the transfer learning technique [19,20], but the effect of transfer learning has not been explored and explained, while the MER accuracies are still low.

Additionally, works such as [12,13,17] do not generalize well as they are trained and validated using one or two micro-expression databases only having similar demographic and data collection conditions. A better evaluation method proposed in Micro-Expression Grand Challenge (MEGC) 2019 involves conducting Leave-One-Subject-Out (LOSO) cross-validation on the composite micro-expression datasets using Unweighted F1 (UF1) score metric designed to alleviate class imbalance and model generalization problems [21]. MER using this composite dataset is an ongoing research problem with the best performance so far by MTM-NET only having attained 0.864 (Unweighted F1) UF1 Score and 0.85 (Unweighted Average Recall) UAR score [15], due to the more challenging dataset with diverse demography and sample collection methods as well as an imbalanced dataset.

To understand how to improve MER on the composite dataset [22] has discovered the importance of lower model and input complexity when utilizing deep learning on the composite dataset, whereas deeper models are more suitable for single database evaluation. Model complexity refers to the number of parameters in the model that consists of all the learnable weights and biases, whereas input complexity is the input resolution [10]. However, research on reducing both the input and model complexity is still lacking so far.

In this work, a MER model is proposed that is based on deep learning with low input and low model complexity to overcome the small and imbalanced dataset as well as to learn salient features from the subtle and spontaneous micro-expressions with improved accuracy on the composite dataset. The contribution of this article is as follows:

- A lightweight and shallow deep learning model based on EfficientNet-B0 model is developed and proposed to demonstrate the limits of the deep learning model's depth to MER accuracy;
- Low input complexity is suggested by using static images of the apex frames without any pre-extracted features as input combined with transfer learning from FER2013 to learn generic facial expression features;
- Addressed imbalanced dataset problem from loss function manipulation's perspective instead of relying on data engineering or generation techniques resulting in a simpler end-to-end MER workflow;
- Finally, effectiveness of the proposed method is proven by the superior performance in MEGC 2019 evaluation protocol.

2. Related Work

2.1. Composite Datase

Currently, there are three widely used spontaneous micro-expression datasets which are SMIC [23], CASME II [24] and SAMM [25] containing video samples of induced micro-expressions. All three datasets have different demographic distributions, data collection methods and micro-expressions labels. In MEGC 2019, a composite database that combines video samples of the three datasets was proposed to produce a realistic database with a variety of subjects and samples [21]. It also standardizes the emotion labels by collapsing them into three classes which are *Positive*, *Negative* and *Surprise*. Samples with class '*Others*' are dropped. This composite dataset is more challenging due to the class imbalance problem with most samples classified into *Negative* class as shown in Table 1 in addition to the diverse subjects and video recording conditions. The proposed model in this work is developed using this composite dataset and evaluated according to MEGC 2019 Leave-One-Subject-Out (LOSO) cross-validation (CV) protocol to demonstrate its performance on a diverse dataset with the class imbalanced.

Table 1. Micro-expression datasets.

Dataset	Samples	Annotations [2]	Class	Emotion Classes [1]
SMIC	164	N/A	3	51 P, 70 N, 43 Su
CASME II	247	On, Apex, Off	5	32 H, 64 D, 25 Su, 27 R, 99 O
SAMM	159	On, Apex, Off	7	57 A, 12 C, 9 D, 8 F, 26 H, 6 S, 15 Su, 26 O
MEGC 2019	442	As per original	3	109 P, 250 N, 83 Su

[1] Emotion Classes—P: Positive. N: Negative. O: Others. H: Happiness. D: Disgust. Su: Surprised. R: Repression. T: Tense. F: Fear. C: Contempt. Sa: Sad. A: Anger. [2] Annotations—On: Onset frame, Off: Offset frame, Apex: Apex frame, N/A: Not available.

2.2. Apex Frame as Input

The work of [7] was the first to suggest that apex frames—which is the frame when the facial expression reached its peak in a micro-expression video—are sufficient for MER, but they encoded the apex frame information in Bi-WOOF feature. The advantages of using apex frame image only as input are low input complexity and allowing leveraging static facial image databases that are widely available. The downside is facial motion information will be lost and apex frame location must be located beforehand. So far, only two works utilized apex frames as input for MER on the composite dataset, which are [26] that required significant pre-processing to locate the apex frames, and ICE-GAN that involved additional data generation steps, cancelling the benefit of low input complexity [27]. In this work, the apex-locating method used is based on [28] that approximated apex frame positions as the middle frame in a video sample for its simplicity and proven performance in MEGC 2019.

2.3. Transfer Learning from FER Datasets

Transfer learning approaches are often used in MER to prevent overfitting the deep learning model [19]. In MER, this is mostly done by pre-training the deep learning model on facial expression datasets such as CK+ [19], Karolinska Directed Emotional Faces (KDEF) [29], Oulu-Casia and MMI [15] datasets to learn facial or expression features. In an extensive cross-evaluation of seven facial expression databases, FER2013 was shown to have the best generalization and transfers well to other facial expression databases as compared to CK+, Oulu-Casia and AffectNet, amongst others [30]. FER2013 is a publicly available dataset consisting of 35887, in-the-wild images of facial expressions labelled into seven emotion classes [31]. In this work, transfer learning from FER2013 to MER composite dataset is explored.

2.4. Lightweight MER

Several works proposed shallow and lightweight deep learning models to avoid overfitting [10] or to reduce heavy computation [32]. Ref. [22] compared performances of shallow model to deeper model and discovered that shallower networks perform better on the composite MER dataset as they are less sensitive to inter-dataset differences. To produce a lightweight MER model, [32] utilized Neural Architecture Search (NAS) to build a CNN model, but their approach used spatiotemporal features. EfficientNet model also utilized NAS to find the best combination of MBConv blocks— the building blocks for MobileNet models [33]. EfficientNet pre-trained on ImageNet dataset had achieved superior results on five out of eight transfer learning tasks on common datasets compared to state-of-the-art models such as ResNet-152, DenseNet-201 and Inception models despite its smaller size. Therefore, this work proposed a lightweight MER model based on the EfficientNet model.

2.5. Loss Functions for Imbalanced Dataset

Several studies utilized loss functions adapted for imbalanced datasets instead of cross-entropy (CE) loss functions, such as triplet loss function that combined losses from triplet inputs [15] and balanced multi-class focal loss (FL) function that penalizes dominant samples based on the sample volume for each class [16]. Ref. [16] has shown excellent performance on the composite dataset; however, the comparison when using FL compared to CE loss functions was not discussed. In this article, the approach of [16] is taken by using balanced focal loss (FL) while elucidating the difference when using CE loss.

3. Methodology

The proposed MER method, as illustrated in Figure 1, consists of data pre-processing steps to neutralize head movements and remove noises (non-expression related) for the composite dataset and relabeling samples in FER2013 dataset, followed by the model pre-training on FER2013, transferring the pre-trained model parameters and fine-tuning the model on the composite dataset before conducting LOSO CV evaluation. These steps are outlined in detail in the subsequent subsections.

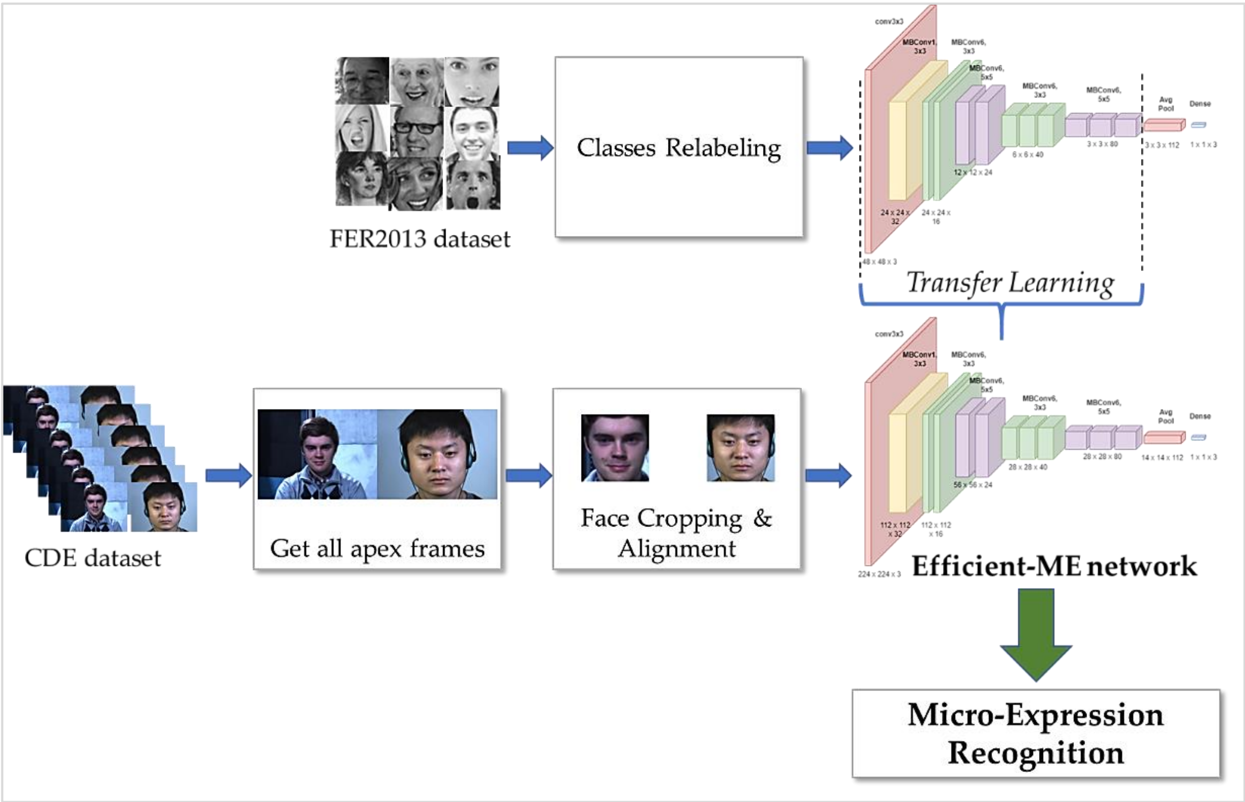

Figure 1. Proposed approach for Micro-Expression Recognition.

3.1. Data Pre-Processing

3.1.1. FER 2013 Dataset

Since the composite MER dataset is using 3-class emotion labels (Positive, Negative and Surprise), samples in FER 2013 are also relabeled accordingly as depicted in Table 2. Samples with label 'Neutral' are dropped as they are unneeded. To discount the effect of imbalanced dataset in FER2013, the same number of samples (4002 samples) is randomly chosen from the relabeled classes giving us a total of 12,006 samples. This dataset is then divided into training and validation sets with 80:20 splits.

Table 2. Labels and number of samples in FER 2013 database.

Original Class	New Class	Total
Angry: 4953, Disgust: 547, Fear: 5121, Sad: 6077	Negative	16,698
Happiness: 8989	Positive	8989
Surprise: 4002	Surprise	4002
Neutral: 6198	N/A	N/A

3.1.2. Composite Dataset

Get All Apex Frames

Samples from CASME II and SAMM datasets in the composite dataset both include the annotation for the apex frames. However, SMIC dataset does not. We take the approach of [28] to approximate the apex frame positions for SMIC dataset as the middle frame in a video sample. After this, apex frames for all the samples in the composite dataset are gathered to be the input to the proposed model. The calculation of the apex frame locations given a *frameList*, which is an array of sorted frames from onset to offset frames in the sample video, is shown below.

$$\begin{aligned} \text{apexLocation} &= \text{length(frameList)}/2 \\ \text{apexFrame} &= \text{frameList[apexLocation]} \end{aligned} \tag{1}$$

Face Cropping and Alignment

At first, face detection was performed on the apex frame images with the goal of retrieving key points that will be used in the face alignment and cropping step. The key points are the right eye and left eye center points as well as the face boundary locations in the image extracted using MTCNN [34] as shown in Figure 2.

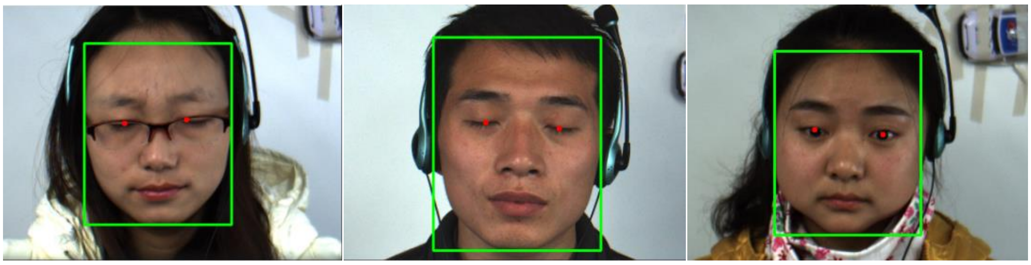

Figure 2. Face and eye detection using MTCNN. Green box: detected face boundary. Red dots: detected eyes' center points.

Next, face alignment is performed to minimize bias due to the head movements as shown in Figure 3. This is done by rotating the image around its center point. The rotation angle is calculated using the inverse of tangent mathematical formula on the difference between the x and y coordinates of the eyes' center locations. The calculation is shown in Equation (2).

$$\begin{aligned} \text{delta\_x} &= \text{right\_eye\_x} - \text{left\_eye\_x} \\ \text{delta\_y} &= \text{right\_eye\_y} - \text{left\_eye\_y} \\ \text{rotation\_angle} &= (\text{atan(delta\_y/delta\_x)} * 180)/\pi \end{aligned} \tag{2}$$

Afterwards, the image is cropped using the face boundary locations found in the initial step to eliminate unnecessary details unrelated to facial expression in the image. The cropped images are then resized to 224 × 224 pixels while ensuring that the image is scaled appropriately. The image dimension (224 × 224) was selected in accordance with the input requirement of the proposed model.

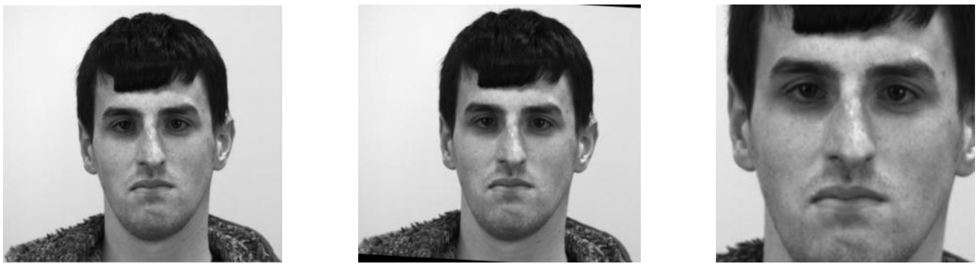

Figure 3. Face alignment and cropping. From left to right: original image, face aligned image, cropped image.

Data Augmentation

Simple image augmentations are applied to prevent overfitting which are horizontal flip and random brightness on random samples in the training dataset. The seed used for the transformation is a randomly generated number between 0 and 10.

3.2. Proposed Model

The proposed model (Figure 4) is based on EfficientNet-B0 model, the lightest variant of EfficientNet. However, EfficientNet-B0 is still comparatively bigger than the other shallower networks proposed for MER—bearing in mind that these networks require pre-extracted OF features [9,10,32]. Therefore, this work proposed a truncated version of EfficientNet-B0, called Efficient-ME with a balance of model complexity (number of parameters) and feature extraction capability. The goal is to demonstrate that a lightweight network under a million parameters with low input complexity can perform better than the other shallow MER models.

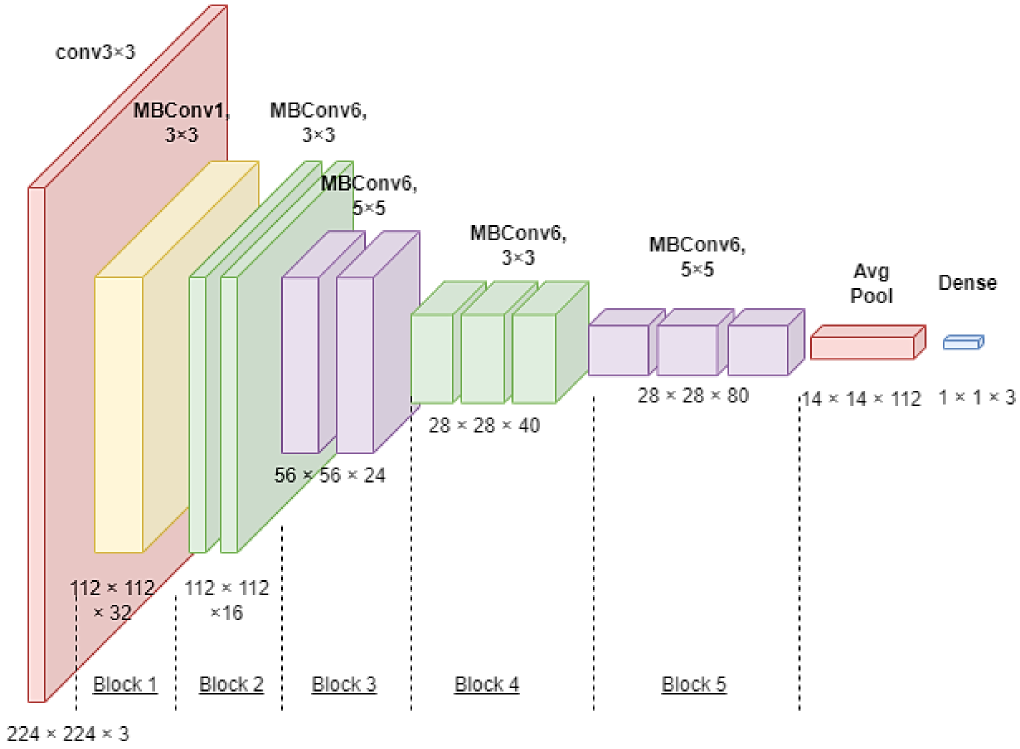

Figure 4. Efficient-ME network architecture.

This network keeps the first 5 blocks of EfficientNet-B0 and drops the last 2 blocks. After the 5th block of EfficientNet-B0, a global average pooling layer and a dropout layer is added before being fed to a dense layer with 3 nodes for classification. Global average pooling layer was added to generate the output feature map to be used in the dense classification layer. A dropout rate of 0.5 is utilized to prevent overfitting [16].

The input dimension is maintained at 224 × 224 with 3 channels. The input images were not transformed to grayscale nor normalized as the first few layers in the Keras's implementation of EfficientNet already take care of this. The dense classification layer uses Softmax activation function together with L1 and L2 kernel regularizers with both regularizing factors set to 0.01 to reduce the kernel weights and prevent overfitting [35]. With the proposed setup, the total number of parameters in Efficient-ME is 867k which is under the target of 1 million parameters.

The justification for truncating EfficientNet-B0 until the 5th block stems from the observation in [36] that a variety of MBConv layers in a network gives better results than using a single type of layers only. Therefore, truncating up to the 5th block will maintain the pair of MBConv6 with 3x3 convolution and MBConv6 with 5x5 convolution configuration (Block 2 and Block 3 pair as well as Block 4 and Block 5 pair, respectively). Another motivation for choosing Block 5 as the cut-off point is the target model size of under 1 million that is achievable by truncating at the 5th block. It is also important that the last convolution layer can extract sufficient features for accurate classification. Truncating the network shallower will result in fewer features and details to be captured by the model.

The optimizer used is Adam optimizer with multi-class focal loss (FL). The parameters' configuration for the FL loss function is $\gamma = 2$ [37] and α that represents the classes' weights as [0.27, 0.35 and 0.38] for class Negative, Positive and Surprise, respectively. Adam optimizer and the α values are selected by using grid search cross-validation.

3.3. Feature Learning from FER 2013

The input layer is first adjusted to (48 × 48) as per FER2013 image sizes before initializing the model's weights from ImageNet pre-trained weights. The model is then trained on FER2013 training dataset for 5 iterations of 100 epochs with early stopping after 10 epochs of no improvement to the validation loss. In each iteration, the base layers were frozen and the network is trained with initial learning rate 0.001 and batch size 16 before fine-tuning the whole network with lower initial learning rate of 0.0001. After the feature learning is completed, the model is saved as the pre-trained model.

3.4. Transfer Learning to Composite Dataset

To learn from FER 2013 database, the weight from the pre-trained model is transferred to the Efficient-ME model. Next, the head of the model is trained by freezing the base layers for 5 iterations, batch size 16 and 15 epochs with early stopping after 2 epochs of no improvement to the validation loss. Then, the whole network is fine-tuned by training for another 5 iterations, batch size 16 and 20 epochs with the same early stopping configuration. The initial learning rate for head training is 0.001 and 0.0001 for the fine-tuning.

The model is trained and evaluated using MEGC 2019 protocol [21] which means the training and test process is repeated 68 times (once for each subject in the composite dataset). At each cross-validation, all samples for a subject are set aside for testing while the rest are used for training the model. At the end of each LOSO CV cycle, there will be 68 testing results that will be used for evaluating the model's performance.

3.5. Experiments

3.5.1. Experimental Setup

The experiments were done on Windows 11 (x64) with a single NVIDIA GeForce RTX 3070 GPU having 8GB of dedicated memory. The model was developed using Keras library on Tensorflow 2.7 framework and Python 3.9 programming language.

3.5.2. Baselines

For consistent comparison, only deep learning methods that adhere to the MEGC 2019 protocol are selected as the baselines. STSTNet [10] which is the lightest MER architecture so far is the baseline for lightweight models. RNAS-MER is the benchmark for NAS-based models and video input [32]. GEME is included to evaluate how using Dynamic Images as input measures up against the proposed method of using apex frames only [16], while ICE-GAN [27] is the baseline for the method using apex frames only. MTM-NET [15] which has attained the highest UF1 score in CDE evaluation is the benchmark for the transfer learning method.

The UF1 and UAR scores for the baseline studies in Table 3 are per what was published in the respective publications as all of them use LOSO CV protocol and metrics on the same composite dataset. The models' parameter counts are also taken from the publications for STSTNet and RNAS-MER, calculated using the model parameters' memory size for ICE-GAN and from the layers' configuration in the GEME model. The parameters' count for MTM-NET could not be deduced as the published work does not specify the required details.

Table 3. Baseline methods on MEGC 2019 composite dataset. OF: Optical Flow. DI: Dynamic images.

Method	Year	Input	UF1	UAR	# Parameters
STSTNet [10]	2019	OF	0.735	0.760	1.67k
ICE-GAN [27]	2020	Apex	0.845	0.841	21.6 mil
MTM-Net [15]	2020	Onset-Apex	0.86	0.85	n/a
GEME [16]	2021	DI	0.7221	0.7303	53 mil
RNAS-MER [32]	2023	Video	0.8302	0.8511	1.91 mil

3.6. Evaluation Metrics

The metrics used to evaluate the model's performance are UF1 and UAR scores as outlined in MEGC 2019 [21]. The formulae for UF1 and UAR are presented below.

$$UF1 = \frac{\sum_{i=1}^{c} F1_i}{N_c} \tag{3}$$

where C is the class labels ('Positive', 'Negative', 'Surprise') and N_C is the number of classes (3).

$$UAR = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{N_i} \tag{4}$$

where C is the number of classes, TP_i is the True positive count for the class i and N_i is the number of samples in the class I.

4. Results and Discussions

Three experiments were carried out specifically: determining whether lightweight deep learning network with low input complexity can achieve high MER performance, investigating the effect of source database in transfer learning and verifying whether the imbalanced dataset problem can be addressed solely by manipulating the loss functions used. The performance metrics used for comparison are UF1 and UAR scores on the composite dataset as outlined in MEGC 2019.

4.1. Experiment 1: Lightweight Models with Low Input Complexity Can Achieve High MER Performance

In Table 4, it can be observed that the proposed method, Efficient-ME, achieved significant results with 14% UF1 score improvement over the highest score so far attained by MTM-Net. Although STSTNet is still much smaller than Efficient-ME, their method takes a pre-extracted Optical Flow input that allows their network to be compressed further. Besides, their UFI and UAR scores are average at best. In contrast, the proposed method does not need any prior feature extraction, accomplishing low input complexity.

Table 4. UF1 and UAR scores against baseline methods on composite dataset.

Method	UF1 Score	UAR Score	# Parameters
STSTNet [10]	0.735	0.760	1.67k
ICE-GAN [27]	0.845	0.841	21.6 mil
MTM-Net [15]	0.86	0.85	n/a
GEME [16]	0.7221	0.7303	53 mil
RNAS-MER [32]	0.8302	0.8511	1.91 mil
Efficient-ME	**0.987**	**1.178**	867 K
Efficient-ME (tr3)	0.76	0.89	68 K

A smaller variant of the proposed model, Efficient-ME (tr3), with only 68k parameters built by truncating the base EfficientNet-B0 model up to the third block, attained reasonable performance with UF1 score of 0.76, better than the results reported by [16] that use Dynamic Images as input and [10] using pre-extracted optical flow. This demonstrates that complex inputs requiring an involved feature engineering process are not necessary to improve MER as smaller models like Efficient-ME (tr3) can achieve acceptable performance by consuming just apex frame images as input. This also proves that apex frame images contain sufficient and relevant micro-expression information for MER.

However, Efficient-ME (tr3) performs worse than the proposed model despite claims by [9–11] that a smaller and lighter network is better for the MER task. To understand this phenomenon, we examine the output feature maps for both models. Efficient-ME (tr3) generates 40 output features of size 28 × 28 at the last MBConv block (Block 3) as can be seen in Figure 5. The features extracted are still extremely low-level consisting of edges' information from the shapes in the image but also from the textures and contours present. This makes the feature map noisier and harder to classify in the dense layer.

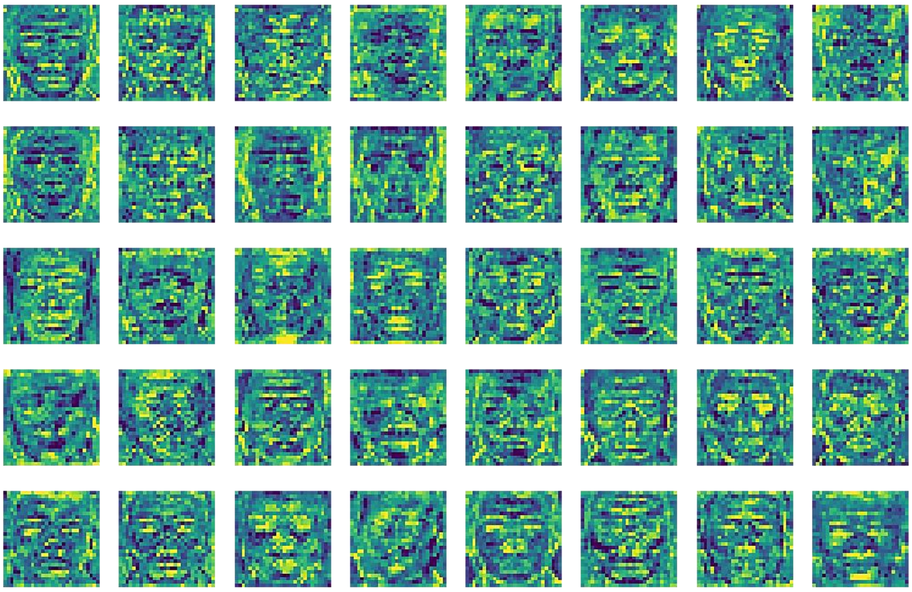

Figure 5. Output feature map of Efficient-ME (tr3).

On the other hand, the proposed model, Efficient-ME, generates 112 features at the last MBConv block (Block 5) with size 14 × 14. The output feature map consists of higher-level details aggregated from previous layers like shapes and edges. Figure 6 shows that the model can tune into the general facial movements such as the pulled up upper lip and downturned mouth that indicates the emotion '*Disgust*'. Hence, the output feature map from this deeper network has the advantage of being immune to unnecessary details like the face texture.

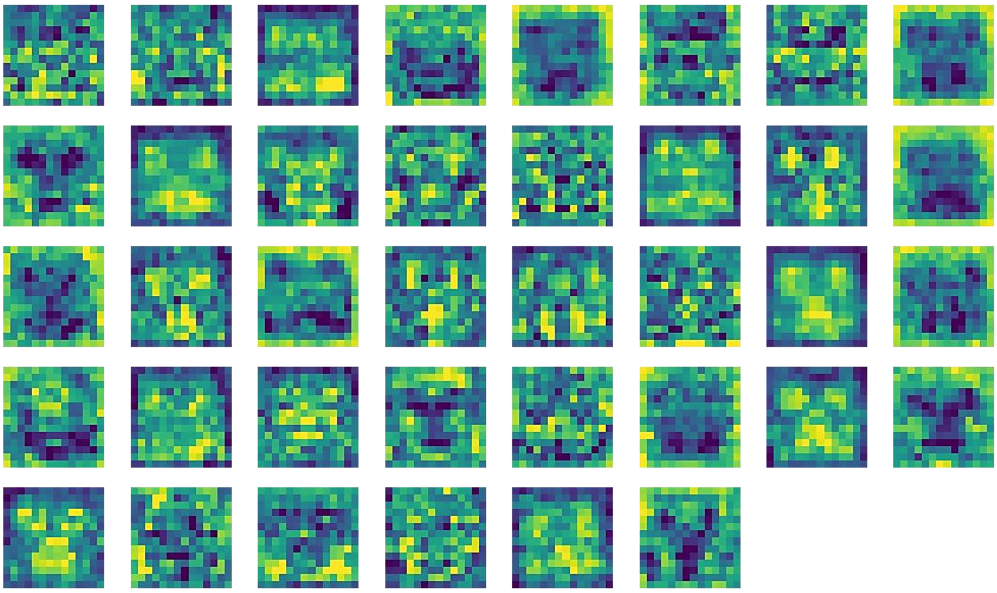

Figure 6. Output feature map of Efficient-ME. The feature maps shown are a sample of 37 features from the total 112 features.

From the comparison of the output feature maps of Efficient-ME and Efficient-ME (tr3), it can be inferred that there is a limit to how small a deep learning model can be before it loses its efficacy in features extractions. Since CNN-based models rely on the stacks of convolution layers to learn features sequentially and build knowledge from the ground up, it is important to consider the balance between a smaller number of parameters and the level of features extraction at the final convolution layer when designing lightweight models.

4.2. Experiment 2: Source Database Used in Transfer Learning Plays a Big Role in MER

MER performances of 2 identical Efficient-ME models were compared: one loaded with pre-trained ImageNet weights and the other with the pre-trained FER 2013 weights. The results show a substantial 50% increase in performance on the model pre-trained with FER 2013 database compared to using ImageNet pre-trained weights. To see whether increasing the model size can improve the MER performance using transfer learning from ImageNet, similar validation was also conducted on the original EfficientNet-B0 architecture with pre-trained ImageNet weights. However, based on Table 5, no noticeable improvements were observed. A further comparative test was conducted by pre-training Efficient-ME model with CK+ dataset to see if the type of facial expression dataset used for transfer learning is important. The result showed considerable performance improvement compared to using ImageNet pre-trained weights but was still inferior to using FER2013 pre-trained weights.

Table 5. Transfer learning evaluation using ImageNet, FER 2013 and CK+ source databases.

Method	UF1 Score	UAR Score
Efficient-ME with pre-trained ImageNet weights	0.46	0.5376
EfficientNet-B0 with pre-trained ImageNet weights	0.44	0.52
Efficient-ME with pre-trained CK+ weights	0.949	1.12
Efficient-ME with pre-trained FER 2013 weights	**0.985**	**1.178**

Therefore, it is prudent to conclude that transfer learning from facial expression databases, such as FER 2013 and CK+, greatly improves MER. However, the type of facial expression dataset used is also important. For this instance, FER 2013 performed better than CK+ due to its diverse and spontaneous samples as well as its larger size.

4.3. Experiment 3: Imbalanced Dataset Problem Can Be Circumvented by Algorithm Tuning

The effect of using multi-class focal loss (FL) using the proposed class weightage is compared to using standard cross-entropy loss (CE). Both Efficient-ME models used were pre-trained using a balanced FER2013 dataset. From the results in Table 6, it can be noted that using multi-class focal loss with class weightage improves the UF1and UAR scores on the imbalanced composite dataset by a considerable margin with UFI score of 0.98 when using FL compared to UFI score 0.916 when using CE, respectively. Drilling into the per-class F1 scores, Efficient-ME with CE performed worst in recognizing class *Surprise*, which has the least number of samples in the composite dataset. On the contrary, the per-class F1 scores for Efficient-ME with FL are more balanced and even. This shows the efficacy of simple algorithm tuning such as loss function manipulation to deal with an imbalanced dataset by emphasizing the importance of minority class without requiring additional samples generation.

Table 6. Loss function evaluation for Efficient-ME pre-trained on balanced FER2013.

Method	UF1 Score	UAR Score	F1-Neg	F1-Pos	F1-Surp
Efficient-ME with FL	**0.98**	**1.178**	**0.99**	**0.98**	**0.98**
Efficient-ME with CE	0.916	1.109	0.94	0.93	0.87

FL: Focal Loss. CE: Cross-entropy Loss. Neg: Class Negative. Pos: Class Positive. Surp: Class Surprise.

Further experiment was done to uncover the effect of FL and CE when the model is pre-trained on an imbalanced source database. This is done by taking the whole FER2013 samples without re-sampling, supplying 8989 samples for class Positive, 16,698 samples for class Negative and 4002 samples for class Surprise.

As can be observed in Table 7, the specific loss function used has less effect when the model is pre-trained on a source database with matching class distribution. The full UFI score and the per-class scores when using FL do not significantly differ to CE although the score for CE is slightly higher. This is because the original imbalanced FER2013 also has the least samples for class *Surprise* followed by class *Positive* like the composite dataset.

Table 7. Loss function evaluation for Efficient-ME pre-trained on imbalanced FER2013.

Method	UF1 Score	UAR Score	F1-Neg	F1-Pos	F1-Surp
Efficient-ME with FL	0.97	1.165	0.98	0.97	0.97
Efficient-ME with CE	**0.99**	**1.191**	**0.99**	**0.99**	**0.99**

FL: Focal Loss. CE: Cross-entropy Loss. Neg: Class Negative. Pos: Class Positive. Surp: Class Surprise.

Hence, it can be summarized that manipulating loss functions for an imbalanced dataset is useful when the class distribution of the source database does not match the target database during transfer learning. This discovery opens the possibility of exploiting the deep learning algorithms to solve imbalanced dataset problems during transfer learning.

5. Conclusions and Future Work

MER has huge potential applications in various domains but is challenging due to the intricate and low-intensity facial movements as well as the small datasets available. In this work, a novel lightweight deep learning model, Efficient-ME, with just 867k parameters, is proposed by truncating the baseline EfficientNet-B0 model after the 5th MBConv blocks pre-trained on a macro-expression database, FER2013, before fine-tuned and evaluated on the composite dataset.

This work has demonstrated that a lightweight deep learning model is able to achieve superior MER performance with the consideration that the model must still be able to extract high-level features. Future comparative study of the lightweight limit of the deep learning model will reveal the deep learning network constraints for MER. It was also confirmed that utilizing apex frames as input with no other features' pre-extraction steps is sufficient for capturing details pertaining to micro-expressions. The combination of the lightweight model with low input complexity utilizing a static apex frame image is proven to achieve state-of-the-art performance compared to other shallow and lightweight MER models.

Furthermore, in the transfer learning's source databases experiment, FER2013 is revealed to be more effective than by using a massive general image database such as ImageNet or lab-controlled and smaller facial expression dataset such as CK+. This proves the importance of selecting a source database that is closely related to the target task as well as diverse enough from which to learn general features. Future work that evaluates the different macro-expression databases using the same baseline model and evaluation protocol will add understanding of the influence of source data of transfer learning for MER.

In dealing with an imbalanced dataset problem, it can be established that using weighted multi-class focal loss can yield high MER performance when utilizing transfer learning if the source database's class distribution does not match the target, rendering complex data augmentation strategies redundant.

In the future, the proposed MER framework can be extended for evaluating 5-class MER. This will necessitate a standard emotion class labelling across the different databases and increase the available micro-expression data.

Author Contributions: Conceptualization, N.A.A.R.; methodology, N.A.A.R.; software, N.A.A.R.; validation, N.A.A.R.; formal analysis, N.A.A.R.; investigation, N.A.A.R.; resources, N.A.A.R.; data curation, N.A.A.R.; writing—original draft preparation, N.A.A.R.; writing—review and editing, N.A.A.R. and S.S.; visualization, N.A.A.R.; supervision, S.S.; project administration, N.A.A.R.; funding acquisition, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Kebangsaan Malaysia, grant number GP2020, GP2021, FTM1 and FTM3.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of public databases, which were conducted according to the guidelines of the Declaration of Helsinki and approved by the relevant review boards. We complied with the terms of use of the databases regarding the publication of data.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Restrictions apply to the availability of these data: CASME II dataset is obtained from ©Xiaolan Fu and is available from http://casme.psych.ac.cn/casme/e2, accessed on 13 January 2023, SMIC dataset is obtained from Xiaobai.Li@oulu.fi and SAMM dataset is obtained from M.Yap@mmu.ac.uk and is available from http://www2.docm.mmu.ac.uk/STAFF/m.yap/SAMM/index.php, accessed on 13 January 2023. These datasets are publicly available: FER2013 dataset can be obtained from https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data, accessed on 13 January 2023 and CK+ dataset can be obtained from https://www.kaggle.com/datasets/shawon10/ckplus, accessed on 13 January 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yan, W.J.; Wu, Q.; Liang, J.; Chen, Y.H.; Fu, X. How fast are the leaked facial expressions: The duration of micro-expressions. *J. Nonverbal Behav.* **2013**, *37*, 217–230. [CrossRef]
2. Ekman, P.; Friesen, W.V. Nonverbal leakage and clues to deception. *Psychiatry* **1969**, *32*, 88–106. [CrossRef] [PubMed]
3. Frank, M.G.; Maccario, C.J.; Govindaraju, V. Behavior and security. In *Protecting Airline Passengers in the Age of Terrorism*; Seidenstat, P., Splane, F.X., Eds.; ABC-CLIO: Santa Barbara, CA, USA, 2009; pp. 86–106.
4. Pfister, T.; Li, X.; Zhao, G.; Pietikäinen, M. Recognizing spontaneous facial micro-expressions. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011. [CrossRef]
5. Wang, Y.; See, J.; Phan, R.C.W.; Oh, Y.H. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014. [CrossRef]

6. Chen, M.; Ma, H.T.; Li, J.; Wang, H. Emotion recognition using fixed length micro-expressions sequence and weighting method. In Proceedings of the 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR), Angkor Wat, Cambodia, 6–10 June 2016. [CrossRef]
7. Liong, S.T.; See, J.; Wong, K.S.; Phan, R. Less is more: Micro-expression recognition from video using apex frame. *Signal Process. Image Commun.* **2016**, *62*, 82–92. [CrossRef]
8. Liong, S.-T.; Wong, K. Micro-expression recognition using apex frame with phase information. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017. [CrossRef]
9. Khor, H.Q.; See, J.; Liong, S.T.; Phan, R.C.W.; Lin, W. Dual-Stream Shallow Networks for Facial Micro-Expression Recognition. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019. [CrossRef]
10. Liong, S.T.; Gan, Y.S.; See, J.; Khor, H.Q.; Huang, Y.C. Shallow triple stream three-dimensional CNN (STSTNet). In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019. [CrossRef]
11. Xia, Z.; Hong, X.; Gao, X.; Feng, X.G. Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-Expressions. *IEEE Trans. Multimed.* **2020**, *22*, 626–640. [CrossRef]
12. Wang, S.; Guan, S.; Lin, H.; Huang, J.; Long, F.; Yao, J. Micro-Expression Recognition Based on Optical Flow and PCANet+. *Sensors* **2022**, *22*, 4296. [CrossRef] [PubMed]
13. Liu, Y.; Li, Y.; Yi, X.; Hu, Z.; Zhang, H.; Liu, Y. Micro-expression recognition model based on TV-L1 optical flow method and improved ShuffleNet. *Sci. Rep.* **2022**, *12*, 17522. [CrossRef] [PubMed]
14. Sie-Min, K.; Zulkifley, M.A.; Kamari, N.A.M. Optimal Compact Network for Micro-Expression Analysis System. *Sensors* **2022**, *22*, 4011. [CrossRef] [PubMed]
15. Xia, B.; Wang, W.; Wang, S.; Chen, E. Learning from Macro-expression: A micro-expression recognition framework. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 14–16 October 2020. [CrossRef]
16. Nie, X.; Takalkar, M.A.; Duan, M.; Zhang, H.; Xu, M. GEME: Dual-stream multi-task Gender-based Micro-Expression recognition. *Neurocomputing* **2021**, *427*, 13–28. [CrossRef]
17. Verm, M.; Vipparthi, S.K.; Malaviya, G.S. AffectiveNet: Affective-Motion Feature Learning for Micro-Expression Recognition. *IEEE Multimed.* **2020**, *28*, 17–27. [CrossRef]
18. Lei, L.; Chen, T.; Li, S.; Li, J. Micro-expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021. [CrossRef]
19. Peng, M.; Wu, Z.; Zhang, Z.; Chen, T. From macro to micro expression recognition: Deep learning on small datasets using transfer learning. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018. [CrossRef]
20. Sun, N.; Cao, S.; Li, D.; He, J.; Yu, L. Dynamic Micro-Expression Recognition Using Knowledge Distillation. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1037–1043. [CrossRef]
21. See, J.; Yap, M.H.; Li, J.; Hong, X.; Wang, S.-J. MEGC 2019—The Second Facial Micro-Expressions Grand Challenge. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019. [CrossRef]
22. Xia, Z.; Peng, W.; Khor, H.Q.; Feng, X.; Zhao, G. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 8590–8605. [CrossRef] [PubMed]
23. Li, X.; Pfister, T.; Huang, X.; Zhao, G.; Pietikäinen, M. A Spontaneous Micro-expression Database: Inducement, Collection and Baseline. In Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai, China, 22–26 April 2013. [CrossRef]
24. Yan, W.J.; Li, X.; Wang, S.J.; Zhao, G.Y.; Liu, Y.J.; Chen, Y.H.; Fu, X.L. CASME II: An improved spontaneous micro expression database and the baseline evaluation. *PLoS ONE* **2013**, *9*, e86041. [CrossRef] [PubMed]
25. Davison, A.K.; Lansley, C.; Coste, N.; Tan, K.; Yap, M.H. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Trans. Affect. Comput.* **2018**, *9*, 116–129. [CrossRef]
26. Quang, N.V.; Chun, J.; Tokuyama, T. CapsuleNet for Micro-Expression Recognition. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019. [CrossRef]
27. Yu, J.; Zhang, C.; Song, Y.; Cai, W. ICE-GAN: Identity-aware and capsule-enhanced GAN for micro-expression recognition and synthesis. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzen, China, 18–22 July 2021. [CrossRef]
28. Zhou, L.; Mao, Q.; Xue, L. Dual-Inception Network for Cross-Database Micro-Expression Recognition. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019. [CrossRef]
29. Wang, S.J.; Li, B.-J.; Liu, Y.-J.; Yan, W.-J.; Ou, X.; Huang, X.; Xu, F.; Fu, X. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing* **2018**, *312*, 251–262. [CrossRef]
30. Li, S.; Deng, W. A Deeper Look at Facial Expression Dataset Bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [CrossRef]

31. Goodfellow, I.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Lee, D.H.; Zhou, Y.; et al. Challenges in Representation Learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Republic of Korea, 3–7 November 2013. [CrossRef]
32. Verma, M.; Lubal, P.; Vipparthi, S.K.; Abdel-Mottaleb, M. RNAS-MER: A Refined Neural Architecture Search with Hybrid Spatiotemporal Operations for Micro-Expression Recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023.
33. Tan, M.; Quoc, V.L. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Available online: http://proceedings.mlr.press/v97/tan19a.html (accessed on 11 September 2020).
34. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
35. Hashmi, M.F.; Ashish, B.; Sharma, V.; Keskar, A.G.; Bokde, N.D.; Yoon, J.H.; Geem, Z.W. LARNet: Real-time detection of facial micro expression using lossless attention residual network. *Sensors* **2021**, *21*, 1098. [CrossRef] [PubMed]
36. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.