applied sciences

MDPI

Article

EffShuffNet: An Efficient Neural Architecture for Adopting a Multi-Model

Jong-In Kim [1,†], Gwang-Hyun Yu [2,†], Jin Lee [2], Dang Thanh Vu [2], Jung-Hyun Kim [1], Hyun-Sun Park [1], Jin-Young Kim [2,*] and Sung-Hoon Hong [2,*]

[1] MicroLED Display Research Center, Korea Photonics Technology Institute (KOPTI), 9, Cheomdan Venture-ro 108beon-gil, Buk-gu, Gwangju 61007, Republic of Korea; jonginkim@kopti.re.kr (J.-I.K.)
[2] Department of ICT Convergence System Engineering, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju 61186, Republic of Korea
* Correspondence: beyondi@jnu.ac.kr (J.-Y.K.); hsh@jnu.ac.kr (S.-H.H.)
† These authors contributed equally to this work.

Abstract: This work discusses the challenges of multi-label image classification and presents a novel Efficient Shuffle Net (EffShuffNet) based on a convolutional neural network (CNN) architecture to address these challenges. Multi-label classification is difficult as the complexity of prediction increases with the number of labels and classes, and current multi-model approaches require optimized deep learning models which increase computational costs. The EffShuff block divides the input feature map into two parts and processes them differently, with one half undergoing a lightweight convolution and the other half undergoing average pooling. The EffShuff transition component shuffles the feature maps after lightweight convolution, resulting in a 57.9% reduction in computational cost compared to ShuffleNetv2. Furthermore, we propose EffShuff-Dense architecture, which incorporates Dense connection to further emphasize low-level features. In experiments, the EffShuffNet achieved 96.975% accuracy in age and gender classification, which is 5.83% higher than the state-of-the-art, while EffShuffDenseNet was even better with 97.63% accuracy. Additionally, the proposed models were found to have better classification performance with smaller model sizes in fine-grained image classification experiments.

Keywords: deep-learning; CNN; lightweight model; multi-label classification

Citation: Kim, J.-I.; Yu, G.-H.; Lee, J.; Vu, D.T.; Kim, J.-H.; Park, H.-S.; Kim, J.-Y.; Hong, S.-H. EffShuffNet: An Efficient Neural Architecture for Adopting a Multi-Model. *Appl. Sci.* **2023**, *13*, 3505. https://doi.org/10.3390/app13063505

Academic Editor: Yu-Dong Zhang

Received: 2 February 2023
Revised: 6 March 2023
Accepted: 7 March 2023
Published: 9 March 2023

1. Introduction

In recent years, deep learning technology has seen significant advancements, leading to an increase in research on deep learning-based image classification models. The ImageNet Dataset [1], a benchmark in the field of image classification, has been widely used to evaluate these models. The evolution of convolutional neural networks (CNNs) can be traced back to AlexNet [2], and since then, numerous CNN models have been developed and have demonstrated impressive performance in image classification, including VGG [3], GoogleNet [4], ResNet [5], and DenseNet [6]. The breakthrough of Transformer [7] in the field of natural language processing (NLP) in 2017 has also influenced the development of image classification models. Attention mechanisms have been applied in models such as Residual Attention Network [8], Squeeze-and-Excitation [9], Non-local Network [10], BottleNeck Attention Network [11] and Convolutional Block Attention Network [12]. In addition, recent studies have directly applied Transformer in image classification tasks, with models such as Vision Transformer [13] and Data efficient Image Transformer [14] showing promising results. Despite the steady improvement in classification accuracy, image classification research faces a trade-off between accuracy and model size, as larger models tend to consume more computational resources.

The image classification studies conducted thus far have broadened the scope of image classification tasks. Image classification tasks can be divided into three main categories:

binary class classification, in which only one correct label exists; multi-class classification, in which there is only one correct label; and multi-label classification, in which multiple correct labels exist. Multi-label classification requires a different learning approach compared to general classification due to the presence of multiple correct labels. This classification problem is considered a challenge in the field of data science because prediction difficulty increases as the number of labels and classes increases. A possible solution to this problem is to utilize multiple deep-learning models, each capable of classifying the correct label from an input image. However, using a multi-model approach in deep learning has limitations, such as increased complexity as the number of multi-label classification types increases, and an increase in computation and calculation due to the use of predictions output by multiple models. Generally, state-of-the-art image classification models have large model sizes and high amounts of computation, which can increase even further when applying an ensemble method.

In this work, we propose the application of the ensemble method to the classification of age and gender in portrait images using the EffShuff and EffShuff-Dense networks. These networks are lightweight and demonstrate good classification performance. The EffShuff network operates by dividing the input feature map into two parts, keeping half of the feature map unchanged, and applying lightweight convolution to the other half. The results from each convolution layer are then combined while preserving the features from the previous feature map. The EffShuff-Dense block, on the other hand, accumulates the features from previous feature maps in a highly efficient manner. By employing the ensemble method on either the EffShuff or EffShuff-Dense, we aim to demonstrate the possibility of achieving better classification results in the multi-label classification task of age and gender in portraits, compared to existing single convolutional neural network-based models. Furthermore, the performance of our method will be verified on a fine-grained image classification dataset with over 100 labels and a balanced distribution of classes.

The present research introduces a novel architecture, referred to as EffShuffNet, which is designed to address the challenge of efficient and effective deep learning model development for practical deployment in edge devices. The novelty of the proposed architecture can be summarized as follows.

- We propose the EffShuff and EffShuff-Dense blocks, which are modifications of the ShuffleNet-v2 unit architecture. Specifically, the "depth-wise convolution" in the transition stage of the ShuffleNet-v2 unit is replaced with "average pooling". This modification enables our EffShuffNet to effectively employ dense connections, thereby enhancing the representation capability of the model by incorporating context throughout all bottom layers.
- We evaluate the proposed architecture using age-and-gender prediction as a representative task requiring multiple light-weight models. Additionally, we test the EffShuff-Dense unit with fine-grained classification tasks, which are challenging tasks, to demonstrate that dense connections are a primary factor in enhancing the performance of the model, which is missing in recent lightweight models.

The structure of this article is as follows: Section 2 introduces the characteristics of representative models for image classification; Section 3 describes multi-models for age and gender classification based on the proposed EffShuff model and EffShuff-Dense model; and Section 4 compares the experimental results with convolutional neural network-based models among representative models of image classification, and Section 5 summarizes the conclusions. The code of all experiments is shared on https://github.com/GWANGHYUNYU/EffShuff_Block (accessed on 6 March 2023).

2. Related Work

This section is divided into five paragraphs. The first paragraph introduces the foundational models for image classification, which sets the context for subsequent discussion on recent trends and advancements in the field. The second paragraph highlights current trends in image classification, which is a rapidly evolving field with new developments

frequently emerging. The third paragraph discusses previous research aimed at achieving lightweight, efficient models, which is an important area of focus for practical deployment in edge devices. The fourth and fifth paragraphs briefly present the potential of light-weight models in practical applications, underscoring the importance of developing models that are both accurate and efficient. Finally, the sixth paragraph provides a brief overview of the proposed architecture and the purpose for introducing it, positioning it as a novel contribution that addresses the limitations of existing lightweight models.

In the field of deep learning-based image classification, convolutional neural network (CNN) models have demonstrated remarkable performance. AlexNet [2], which is considered the origin of CNN, established the standard of CNN models in image classification by incorporating diverse-sized convolutional filters, Rectified Linear Unit (ReLU) activation function, and max pooling in its design, drawing inspiration from LeNet [15]. Subsequently, the VGGNet [3] model was introduced with the unique characteristic of having varying receptive fields from the input, achieved through the use of 3×3 convolution filters. GoogleNet [4], on the other hand, was designed with an Inception block that concatenates the outputs of convolutional filters of different sizes and was constructed with a complex model using the bottleneck technique and actively incorporating 1×1 convolutional filters. However, building deep layers remained a challenge. ResNet [5] resolved this challenge by introducing residual learning that added the input value as is after performing convolution operation through a short-cut connection, enabling the construction of deep CNN models. Since then, various studies have been conducted based on ResNet, among which DenseNet [6] adopted a structure that concatenated all previous inputs through dense connection, instead of simply adding them as is. However, these CNN models that enhance image classification performance have the disadvantage of increased model size with increased classification accuracy

In the realm of deep learning-based image classification models, the attention mechanism of the Transformer [7,13], a novel approach in the field of natural language processing, has been incorporated. The Residual Attention Network [8] proposed a structure that utilized an independent attention module to incorporate both the spatial and channel information of previous feature maps, while the SE Net [9] incorporated an attention mechanism based on channel information of previous feature maps. In the BAM [11] model, both spatial and channel information of previous feature maps were considered as attention in parallel, and the CBAM [12] model proposed a method of applying spatial attention after channel attention. The Non-local Network employed self-attention generation and was applied based on spatial information, as in the Transformer method. Although models incorporating the attention mechanism have demonstrated improved classification accuracy compared to conventional convolutional neural network-based models, they entail additional computation. ViT and DeiT [13,14], which directly apply the Transformer to image classification, exhibit promising performance, but have a significant time and space complexity.

On the other hand, as research on image classification models based on convolutional neural networks has been around for a long time, there are studies on models focused on lightweight so that they can work well in limited hardware, such as smartphones or embedded boards. In the field of deep learning-based image classification, Xception [16] introduced a simplified version of the Inception block by implementing independent 3×3 convolutional filters for each channel of the output feature map from a 1×1 convolutional filter, reducing computational cost while preserving high performance. SqueezeNet [17] reduced the model size by 50 times compared to Alexnet through the use of a bottleneck 1×1 filter and a fire block of 1×1 filter for half and 3×3 filter for the rest. MobileNet [18] introduced the concept of 3×3 depthwise convolution and 1×1 pointwise convolution by combining 3×3 convolutional filters from Xception with 1×1 convolutional filters. ShuffleNet [19] proposed a lighter model by grouping and processing the MobileNet structure based on channels and later concatenating the results. MobileNetV2 [20] further reduced its size through the implementation of Bottleneck Resid-

ual Block, and ShuffleNetV2 [21] also achieved lighter weight through a thorough analysis of lightweight models.

In recent years, there has been a trend in deep learning towards training generable big models with unsupervised pre-text tasks on redundant datasets, followed by fine-tuning on a supervised downstream task with a smaller model. For instance, the Contrastive Captioner (CoCa) [22] pretrain image-text encoder-decoder achieves state-of-the-art accuracy of 91.0% on ImageNet, while Contrastive Language-Image Pre-training (CLIP) [23] enables zero-shot transfer for generalization to unseen object categories with 76.2% accuracy on ImageNet. Model distillation [24–26], where the predictions of large, complex teacher models are distilled into smaller models, is the most widely known form of distillation. Interestingly, Noisy Student Training [27] is one of a counter example of light-weight distillation, which combines self-training with distillation using equal-or-larger student models and adding noise during learning. When applied to labeled and pseudo-labeled images with noisy data and stochastic depth, Noisy Student Training using the Efficient Net student model achieved 88.4% top-1 accuracy on ImageNet.

Practitioners seeking to implement lightweight neural models in practical applications can refer to prior research. For instance, ref [28] presents various shallow neural networks for fall detection in the elderly, integrated with wearable devices that capture accelerometer data as input signals. Moreover, ref [29] introduces WearNet, a lightweight convolutional neural network that detects scratches resulting from cylinder flat sliding. These models represent only a small selection of the numerous endeavors aimed at incorporating lightweight neural network models into low-computational products. Despite their lightweight structure, these models have limitations in terms of low classification accuracy compared to general convolutional neural network models and Transformer series models.

Therefore, to effectively address multi-classification problems, such as the classification of both age and gender from a single input, it is crucial to evaluate the strengths and limitations of various image classification models. Merely optimizing for classification performance in a single model for multiple tasks may result in a model that is too complex, and a focus on minimizing model size could negatively impact classification accuracy. In light of this, we propose EffShuff and EffShuff-Dense models, which leverage the attributes of each convolutional output feature map to enhance classification performance while still being relatively lightweight. These models will be incorporated into a multi-model approach aimed at solving the multi-classification problem of age and gender, and the efficacy of this approach will be demonstrated.

3. Methodology

This section describes a multi-model method for age and gender classification. To this end, the detailed configuration of the EffShuff and EffShuff-Dense models and their differences from the existing models are analyzed.

3.1. Age and Gender Classification

The multi-model configuration for the classification of age and gender is proposed and depicted in Figure 1. Deep learning models are established to predict each label of age and gender from the same portrait input. Hence, the overall size, computational complexity, and classification performance of the multi-model are dependent on the performance of the individual deep-learning models.

The generalized configuration of multi-models for multi-label classification is depicted in Figure 2. The configuration consists of individual deep learning models, each of which is tasked with classifying a specific label from a single input. The number of deep learning models required is directly proportional to the number of classifications to be made from the input.

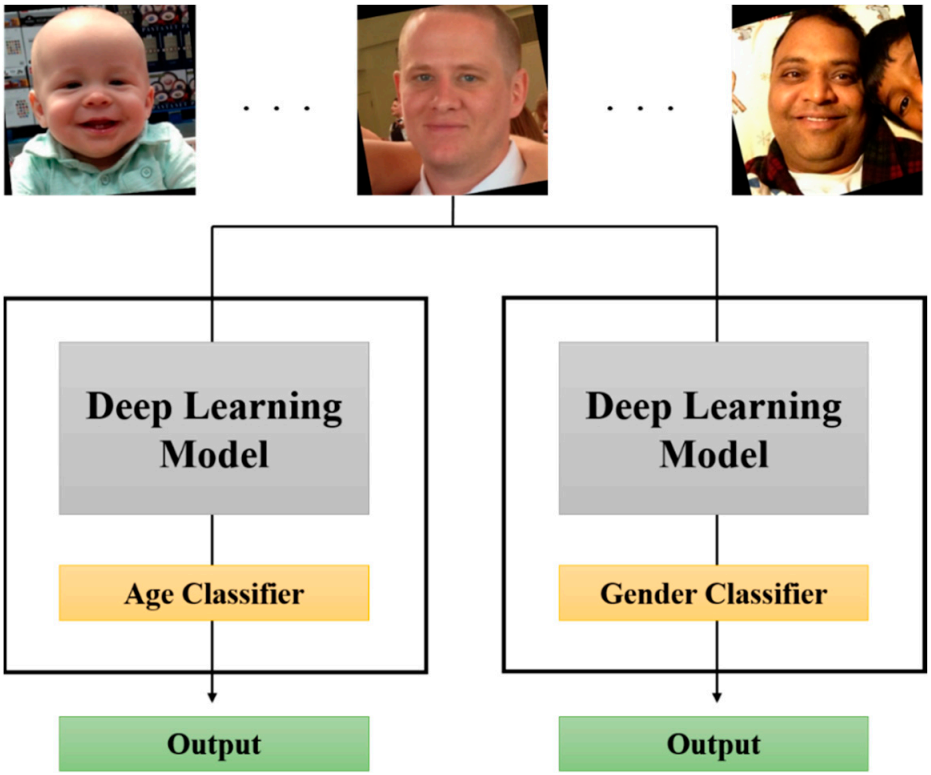

Figure 1. Multi model architecture of age and gender classification.

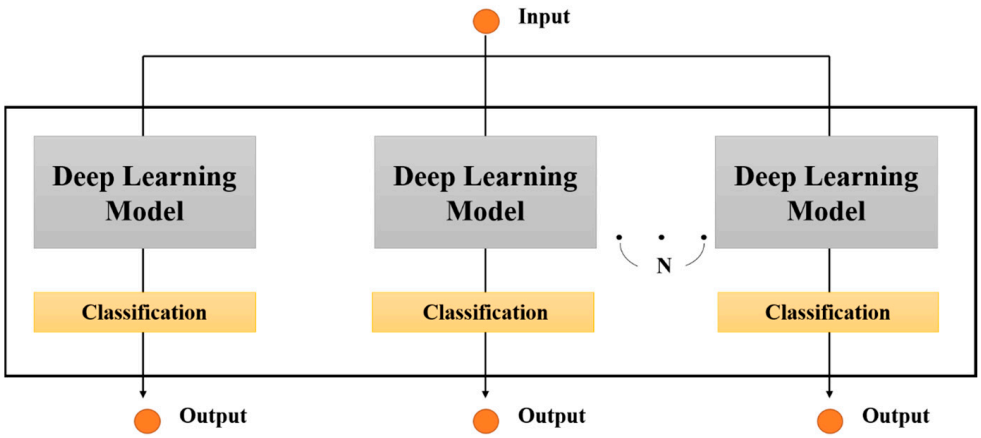

Figure 2. Multi model architecture of multi-label classification.

3.2. EffShuffNet

The configuration of the proposed EffShuff and EffShuff transition blocks is depicted in Figure 3. The EffShuffNet is a lightweight convolutional composed of three main components: the general stem, the EffShuff block, and the EffShuff transition. The general stem is a sequence of convolutional layers that extract low-level features. The EffShuff block and EffShuff transition operate at each stage and within the same dimension feature space or changed dimension feature space at a deeper level. Together, these components form the architecture of EffShuffNet. Inspired by ShuffleNet-V2, the EffShuff configuration balances the channels of the input and output feature maps, avoids group convolution, reduces branching, and reduces computation.

As illustrated in Figure 4, the EffShuff block has a similar configuration to the ShuffleNet Unit; however, the EffShuff transitions utilizes 3×3 average pooling and 1×1 convolution instead of processing 3×3 depthwise convolution and 1×1 pointwise convolution with stride two, thus simplifying the compression process while reducing

computation. As shown in Table 1, the overall structure of the EffShuffNet model takes an input of 224 × 224 × 3 and extracts low-level features through convolution and max pooling operations three times in total using the stem area and EffShuff transition, with EffShuff blocks three, seven, and three inserted between EffShuff transitions.

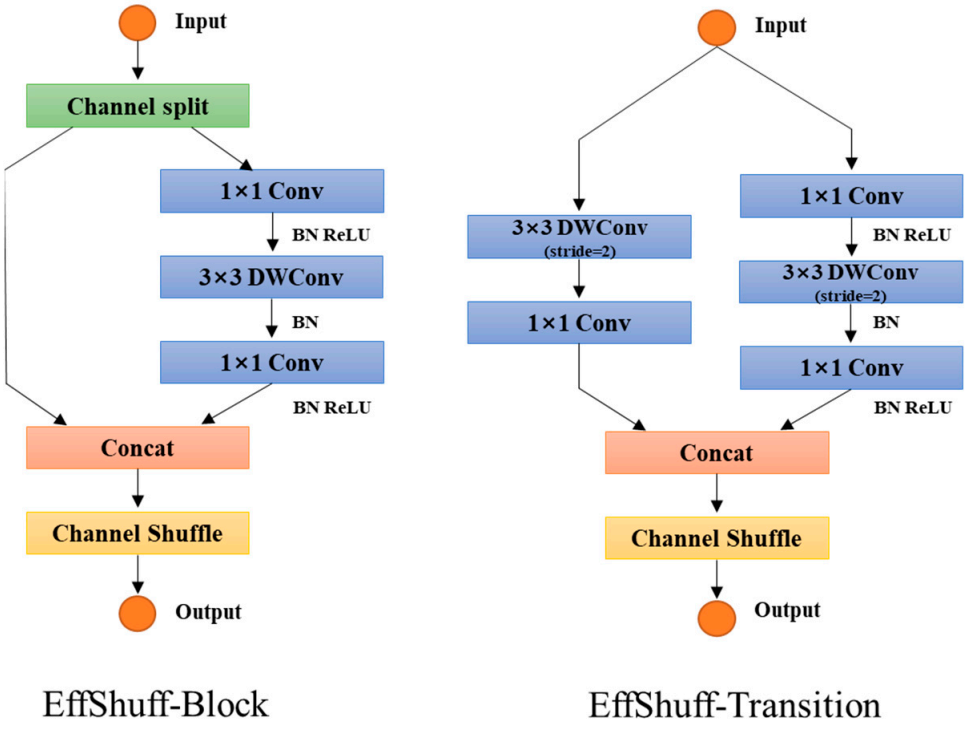

Figure 3. Architecture of EffShuffNet building blocks including EffShuff block and EffShuff transition.

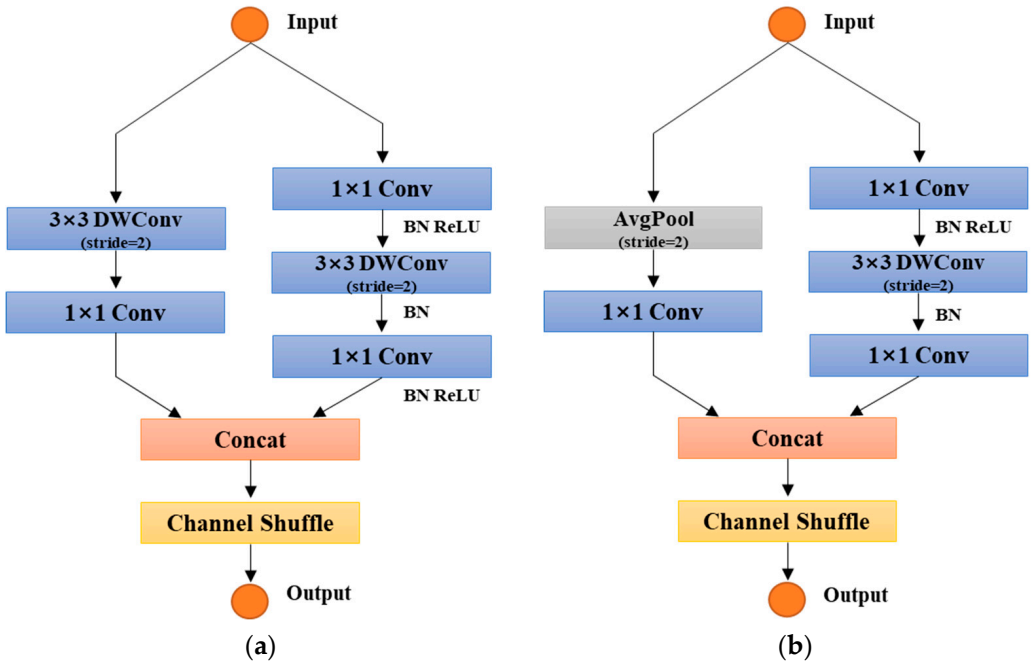

Figure 4. Differences between an EffShuff transition block and a ShuffleNet down sampling unit. The 3 × 3 DWConv layer is replaced by AvgPool with stride two. (**a**) ShuffleNet v2 down sampling unit; (**b**) EffShuff transition block.

Table 1. EffShuffNet architecture.

Layers	Stage	Output Size	Output Filters	Repeat
Input	General Stem	224 × 224	3	1
Convolution		112 × 112	24	1
MaxPool		56 × 56	24	1
EffShuff transition	Stage 1	28 × 28	116	1
EffShuff block		28 × 28	116	1
EffShuff block		28 × 28	116	2
EffShuff transition	Stage 2	14 × 14	232	1
EffShuff block		14 × 14	232	1
EffShuff block		14 × 14	232	6
EffShuff transition	Stage 3	7 × 7	464	1
EffShuff block		7 × 7	464	1
EffShuff block		7 × 7	464	2
GAP	Classifier	1 × 1	464	1
Softmax		1 × 1	No of classes	1

3.3. EffShuffDenseNet

The EffShuffNet successfully achieves both lightweight and performance, leveraging its ability to retain the low-level features generated from the stem area to the classifier. Building on this, the EffShuffDenseNet model is proposed, incorporating the concept of Dense connection from DenseNet, which optimizes the preservation of low-level features.

The structure of the EffShuff-Dense block is depicted in Figure 5. The input feature map is divided into two parts in the channel direction, with one part remaining unchanged and the other part undergoing lightened convolution operations consisting of 1 × 1 convolution, 3 × 3 depthwise convolution, and 1 × 1 pointwise convolution. The outputs of these two parts are then concatenated. Unlike the previously introduced EffShuff block, the concatenated output from EffShuff-Dense block is not mixed, but instead utilized as input for the next EffShuff-Dense block. Such a connection is necessary since the EffShuff-Dense transition can be utilized as is to reduce both the horizontal and vertical size, while doubling the number of channels.

The concept of EffShuff-Dense block is illustrated in Figure 6. The channel of the input feature map utilized in EffShuff-Dense block is represented by the variable C, which is divided into two parts, C_{front} and C_{back}. C_{front} is utilized as is, while C_{back} undergoes a lightweight convolution process that involves a 1 × 1 convolution, followed by a 3 × 3 depthwise convolution and finally, a 1 × 1 pointwise convolution. This results in an increase in the channel size, denoted by K, referred to as the growth rate, with a larger K representing a larger model as a hyperparameter. The concatenated output results in an increase in the channel size to $C_{front} + C_{back} + K = C + K$. As the EffShuff-Dense block is repeated, the outputs accumulate as much as $K/2$, calculated for the previous feature map C_{front} and each block. The use of EffShuff transition in pooling helps to suppress overfitting of low-level features by combining the entire feature map. The EffShuffDenseNet is lightweight while still considering all features calculated from the stem area to the previous classifier stage. As shown in Table 2, the overall EffShuffDenseNet model receives inputs of 224 × 224 × 3 and utilizes the stem area and Y-Transition a total of three times for the extraction of low-level features through convolution and max pooling. The structure employs EffShuff-Dense block three, seven, and three times between transitions, with a growth rate of 16 for EffShuff-Dense block.

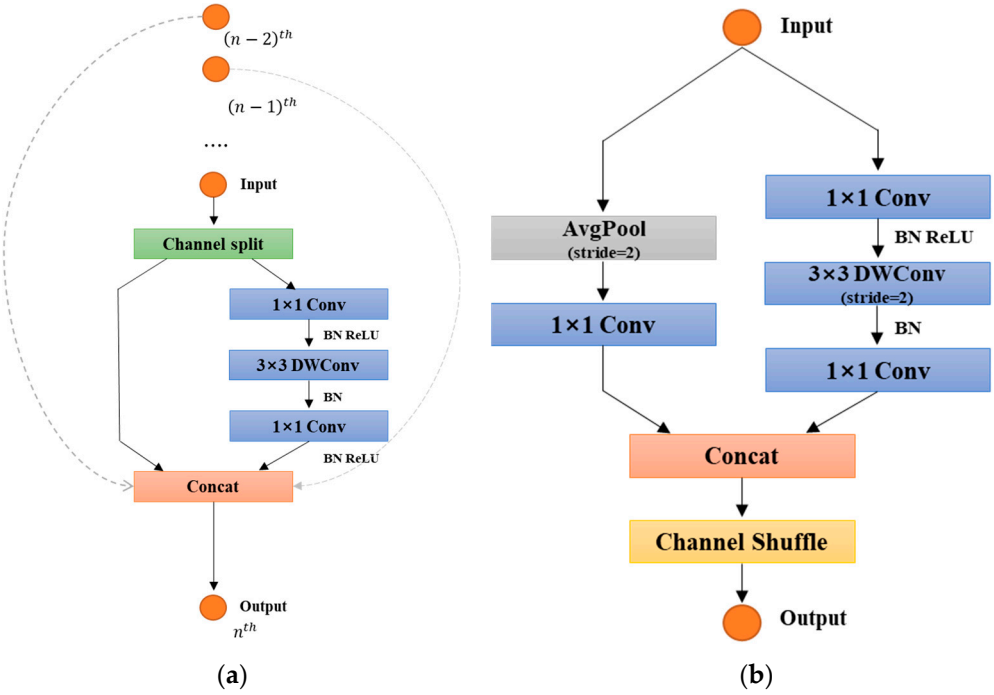

Figure 5. Architecture of EffShuff-Dense block and EffShuff-Dense transition. The main difference between EffShuff block and EffShuff-Dense block is that the Concat layer retrieves dense connection from previous layers. (**a**) EffShuff-Dense block; (**b**) EffShuff-Dense transition.

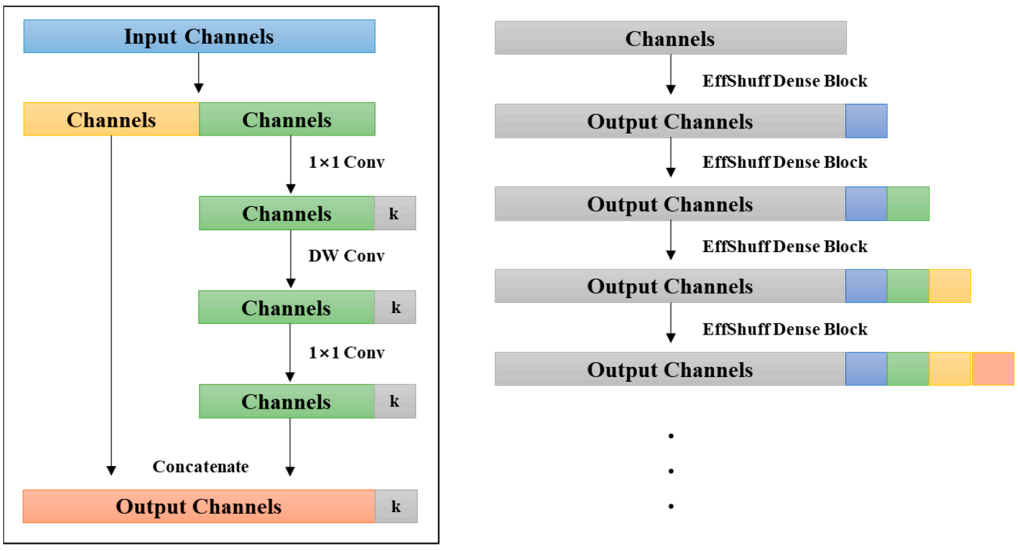

Figure 6. Details of EffShuff-Dense-Block.

The primary difference between the EffShuff unit and the ShuffleNet-v2 unit is the replacement of the "depth-wise convolution" with "average pooling" during the transition, which results in a lighter version of ShuffleNet. Empirical findings demonstrate that the accuracy of EffShuffNet is comparable to that of ShuffleNet-v2 across all tasks. Notably, the utilization of "average pooling" as a read-out operation facilitates dense connections in the network, as the introduction of additional channels does not entail extra parameters for the "average pooling" process, unlike the linearly increasing parameters in the case of "depth-wise convolution".

Table 2. EffShuff-Dense-Block model architecture.

Layers	Stage	Output Size	Output Filters	Growth Rate (K)
Input	General Stem	224 × 224	3	
Convolution		112 × 112	64	
MaxPool		56 × 56	64	
EffShuff transition	Stage 1	28 × 28	92	
EffShuff-Dense block		28 × 28	108	16
EffShuff-Dense block		28 × 28	124	16
EffShuff-Dense block		28 × 28	140	16
EffShuff transition	Stage 2	14 × 14	176	
EffShuff-Dense block		14 × 14	192	16
EffShuff-Dense block		14 × 14	208	16
EffShuff-Dense block		14 × 14	224	16
EffShuff-Dense block		14 × 14	224	16
EffShuff-Dense block		14 × 14	224	16
EffShuff-Dense block		14 × 14	272	16
EffShuff-Dense block		14 × 14	288	16
EffShuff transition	Stage 3	7 × 7	440	
EffShuff-Dense block		7 × 7	456	16
EffShuff-Dense block		7 × 7	472	16
EffShuff-Dense block		7 × 7	488	16
GAP	Classifier	1 × 1	488	
Softmax		1 × 1	No of classes	

4. Experimental Results

In this section, we conduct experiments to evaluate the performance of the EffShuffNet and EffShuffDenseNet, which are lightweight convolutional neural network-based multi-models for age and gender classification.

4.1. Environmental Settings

The experimental evaluation of the EffShuffNet and EffShuffDenseNet was carried out on a deep-learning server equipped with Intel(R) Xeon(R) W-2123 CPU and NVIDIA RTX 3060 VRAM 12 GB, with 32 GB RAM memory. The experiments were implemented using the TensorFlow 2.10 framework, programmed in Python, with official models of ResNet50, DenseNet201, MobileNet_V2, and EfficientNetB0 being directly utilized. Both scratch models and models pre-trained with ImageNet were employed in the experiments.

4.2. Dataset

The evaluation of the performance of the EffShuffNet and EffShuffDenseNet was conducted using several datasets in the proposed multi-model for age and gender classification. The datasets used were the Audience dataset [30], the Butterfly & Moths dataset [31], and the NA Birds dataset [32].

- The Audience dataset [30] is a well-known dataset for age and gender classification that was published in 2014. The images used in this dataset were captured under various real-world conditions, including different appearances, poses, lighting conditions, and image quality. The dataset consisted of 15,163 images for age classification, which was

divided into 10 classes, and 15,300 images for gender classification, which was divided into four classes.

- The Butterfly & Moths dataset [31], as depicted in Figure 7b, is a dataset for butterfly and moth classification. This dataset was published on Kaggle in 2019 and consisted of 12,639 images with 100 classes and an input size of 224 × 224 × 3.

(**a**) (**b**) (**c**)

Figure 7. Examples from three datasets. (**a**) The Audience dataset; (**b**) Butterfly & Moths dataset; (**c**) NA Birds dataset.

- The NA Birds dataset [32], illustrated in Figure 7c, is a representative dataset for fine-grain image classification. This dataset comprises photos of 400 species of birds commonly found in North America and contains 48,000 annotations. The dataset consists of 48,562 images and 555 classes.

All datasets were resized to the input size of 224 × 224 × 3. The NA Birds dataset, in particular, was zero-padded to create a square image, and then resized to the desired size. For each dataset, 90% of the images were used for training and 10% for testing. The same ratio was used for all datasets in the experiment.

4.3. Results

In the multi-model proposed for age and gender classification, ResNet50, DenseNet201, EfficientNetB0, MobileNetV2, ShuffleNetV2, the EffShuffNet, and the EffShuffDenseNet, were trained and evaluated from scratch. The datasets for age and gender classification were divided into the age dataset and gender dataset, respectively, and the learning process was executed. Additionally, transfer learning was performed on the models, ResNet50, DenseNet201, EfficientNetB0, and MobileNetV2, which were provided as TensorFlow2 formulas and had their weights pre-trained with ImageNet. The hyperparameters used for age and gender classification included the learning rate and Adam optimizer, with batch sizes of 16 or 64 and 100 epochs. The validation accuracy was monitored using early stopping, and if no change was observed within a set period, the learning process was terminated.

Table 3 presents the results of a comparative experiment on image classification models for age classification. The EffShuffNet exhibits the smallest model size at 1.13 M and the lowest computational cost at 4.85 G compared to other scratch models, while attaining the highest classification accuracy of 96.37%, rivaling that of ResNet50. Although the EffShuffDenseNet showed an improvement in classification accuracy to 97.42% relative to the EffShuffNet, it was observed that the model size and computation cost slightly increased. However, both the EffShuffNet and EffShuffDenseNet were confirmed to have smaller model sizes compared to MobileNetV2 and ShuffleNetV2, which are well-known lightweight convolutional neural network models. Table 4 presents the results of a comparative experiment on image classification models for gender classification. Similar to the results of the age classification, the EffShuffNet has the smallest model size and lowest computational cost relative to other scratch models, while achieving the highest classification accuracy at 97.58%. The EffShuffDenseNet was observed to have improved classification accuracy compared to the EffShuffNet.

Table 3. Experimental results of age classification.

Model	Top-1 Accuracy	Params	Flops
ResNet50 (PreTrained)	98.06	24.59 M	124.33 G
ResNet50 (Scratch)	96.3	24.59 M	124.33 G
DenseNet201 (PreTrained)	98.94	19.26 M	138.39 G
DenseNet201 (Scratch)	95.91	19.26 M	138.39 G
EfficientNetB0 (PreTrained)	96.63	4.67 M	13.01 G
EfficientNetB0 (Scratch)	93.27	4.67 M	13.01 G
MobileNetV2 (PreTrained)	98.15	2.88 M	10.00 G
MobileNetV2 (Scratch)	95.31	2.88 M	10.00 G
ShuffleNetV2 (Scratch)	95.97	1.95 M	5.86 G
EffShuffNet	96.37	1.13 M	4.85 G
EffShuffDenseNet	97.42	1.43 M	7.44 G

Table 4. Experimental results of gender classification.

Model	Top-1 Accuracy	Params	Flops
ResNet50 (PreTrained)	98.75	24.59 M	124.33 G
ResNet50 (Scratch)	97.18	24.59 M	124.33 G
DenseNet201 (PreTrained)	98.15	19.26 M	138.39 G
DenseNet201 (Scratch)	96.73	19.26 M	138.39 G
EfficientNetB0 (PreTrained)	92.94	4.67 M	13.01 G
EfficientNetB0 (Scratch)	95.09	4.67 M	13.01 G
MobileNetV2 (PreTrained)	98.43	2.88 M	10.00 G
MobileNetV2 (Scratch)	96.92	2.88 M	10.00 G
ShuffleNetV2 (Scratch)	96.53	1.95 M	5.86 G
EffShuffNet	97.58	1.13 M	4.85 G
EffShuffDenseNet	97.84	1.43 M	7.44 G

The results of this experiment indicate that the multi-model of EffShuffNet surpasses ResNet50, which previously demonstrated the highest accuracy in scratch-based age and gender classification, with a classification accuracy of 96.975%. Furthermore, the multi-

model of EffShuffNet exhibits a significant reduction in both model size and computation amount compared to the multi-model of ResNet50. Specifically, the model size of the multi-model of EffShuffNet is 2.26 M, a reduction of 21.7 times in comparison to the 49.18 M size of the multi-model of ResNet50. Additionally, the computational amount of the multi-model of EffShuffNet is 9.7 G, a reduction of 25.6 times compared to the 248.66 G computational amount of the multi-model of ResNet50.

In the experiments for butterfly and moth classification and NA bird classification, the hyperparameter of learning rate was utilized and Adam was selected as the optimizer. The batch size was set to 64 and the number of epochs was 100, with the validation accuracy being monitored using early stopping. Training was ceased if the validation accuracy remained unchanged for 15 iterations.

The results of the comparison experiment for the image classification model for butterfly and moth classification are presented in Table 5. The EffShuffNet displayed the smallest model size of 1.21 M and the lowest computation amount of 4.85 G compared to other scratch models, while exhibiting the highest classification accuracy of 97.70%. Although the EffShuffDenseNet showed an improvement in classification accuracy to 98.25% compared to the EffShuffNet, there was a slight increase in both the model size and computation volume. Nonetheless, it was demonstrated that the EffShuffNet and EffShuffDenseNet possess smaller model sizes in comparison to MobileNetV2 and ShuffleNetV2, which are well-known lightweight convolutional neural network models.

Table 5. Experimental results of butterfly and moth classification.

Model	Top-1 Accuracy	Params	Flops
ResNet50 (PreTrained)	99.68	33.62 M	124.33 G
ResNet50 (Scratch)	94.29	33.62 M	124.33 G
DenseNet201 (PreTrained)	99.60	27.73 M	138.39 G
DenseNet201 (Scratch)	96.91	27.73 M	138.39 G
EfficientNetB0 (PreTrained)	99.20	10.32 M	13.01 G
EfficientNetB0 (Scratch)	94.06	10.32 M	13.01 G
MobileNetV2 (PreTrained)	99.76	8.53 M	10.00 G
MobileNetV2 (Scratch)	95.64	8.53 M	10.00 G
ShuffleNetV2 (Scratch)	95.56	2.05 M	5.86 G
EffShuff-Block	97.70	1.21 M	4.85 G
EffShuff-Dense-Block	98.25	1.47 M	7.44 G

The results of the comparison experiment for NA Birds classification are presented in Table 6. Similar to the results for butterfly and moth classification, the EffShuffNet had the smallest model size, 1.64 M, and the lowest computational amount, 9.70 G, among the scratch models. However, its classification accuracy was relatively low at 88.26%. Upon examining the configuration of the EffShuffNet, it was found that the output layer before the Softmax activation had only 464 units, which was less than the number of classes, 555, in the NA Birds dataset. To address this issue, the EffShuffNet-Plus was introduced, which added an output layer of 512 units while retaining the model size and

computational amount. As a result, the model size was reduced to the smallest at 1.92 M and the computational amount was reduced to the lowest at 10.53 G, with an improved classification accuracy of 91.68%, which was the highest among the models except for the EffShuffDenseNet. The EffShuffDenseNet improved the accuracy to 91.08% compared to the EffShuffNet. Nevertheless, the EffShuffNet and EffShuffDenseNet still had smaller model sizes than the MobileNetV2 and ShuffleNetV2 models, which are representative of lightweight convolutional neural networks.

Table 6. Experimental results of NA Birds classification.

Model	Top-1 Accuracy	Params	Flops
ResNet50 (PreTrained)	93.18	79.28 M	248.67 G
ResNet50 (Scratch)	90.42	79.28 M	248.67 G
DenseNet201 (PreTrained)	95.22	70.53 M	276.79 G
DenseNet201 (Scratch)	91.47	70.53 M	276.79 G
EfficientNetB0 (PreTrained)	87.41	38.85 M	26.01 G
EfficientNetB0 (Scratch)	90.63	38.85 M	26.01 G
MobileNetV2 (PreTrained)	94.17	37.06 M	20.00 G
MobileNetV2 (Scratch)	90.60	37.06 M	20.00 G
ShuffleNetV2 (Scratch)	87.41	2.51 M	11.73 G
EffShuffNet	88.26	1.64 M	9.70 G
EffShuffNet-Plus	91.68	1.92 M	10.53 G
EffShuffDenseNet	91.08	1.69 M	14.89 G

In all of the image classification experiments conducted, the EffShuffNet, EffShuffDenseNet, and EffShuffNet-Plus were not found to outperform other pre-trained image classification models. It is expected that higher classification accuracy will be achieved through transfer learning, by pre-training the EffShuffNet and EffShuffDenseNet with the ImageNet dataset. Figures 8 and 9 depict the experimental results for the EffShuffNet and EffShuffDenseNet, which are lightweight and exhibit good classification performance in the proposed multi-model for age and gender classification.

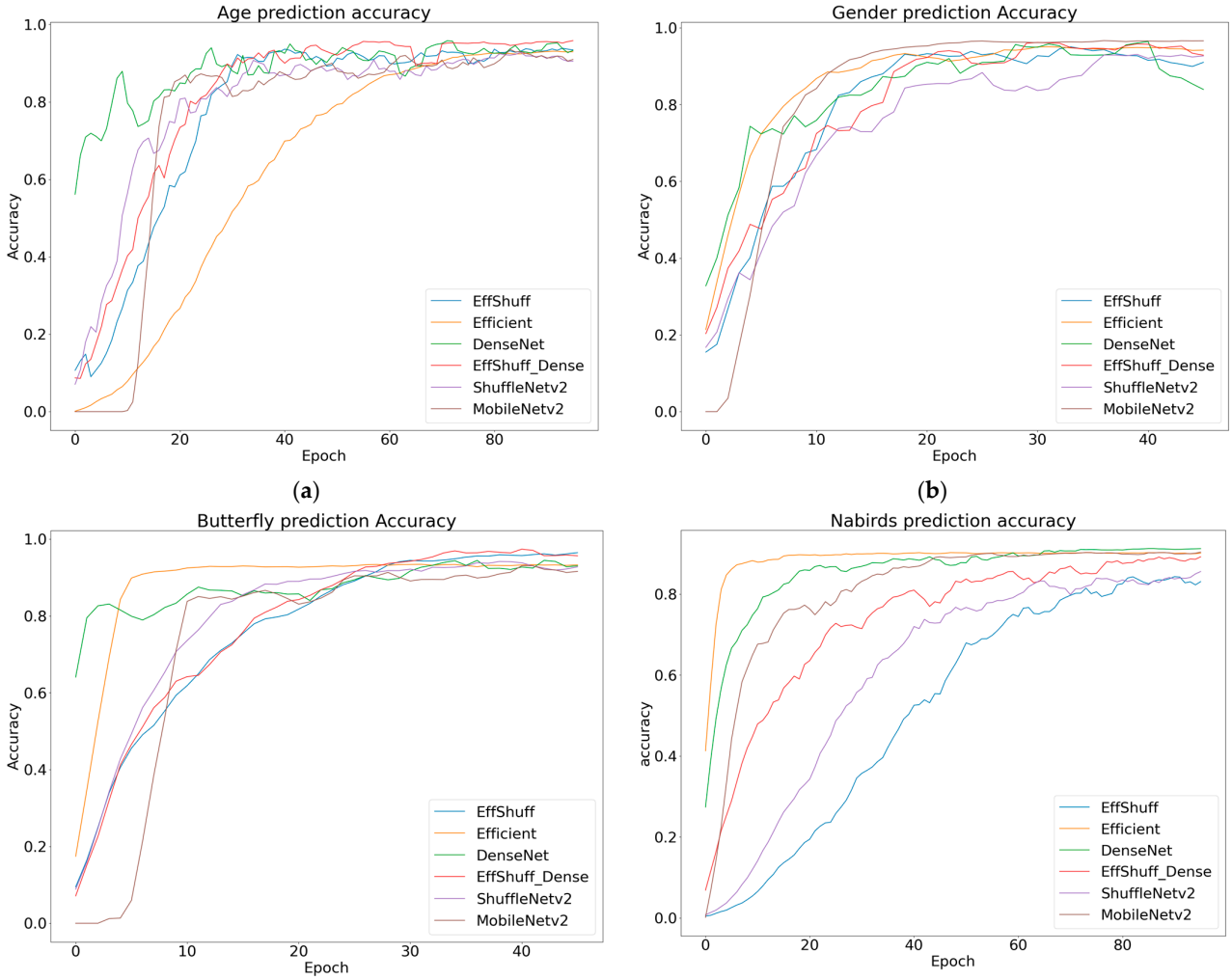

Figure 8. Accuracy comparison of EffShuffNet, EffShuffDenseNet and other models in validation scheme. (**a**) The Audience dataset: age prediction; (**b**) the Audience dataset: gender prediction; (**c**) the Butterfly & Moths dataset: class prediction; (**d**) the NA Birds dataset: class prediction.

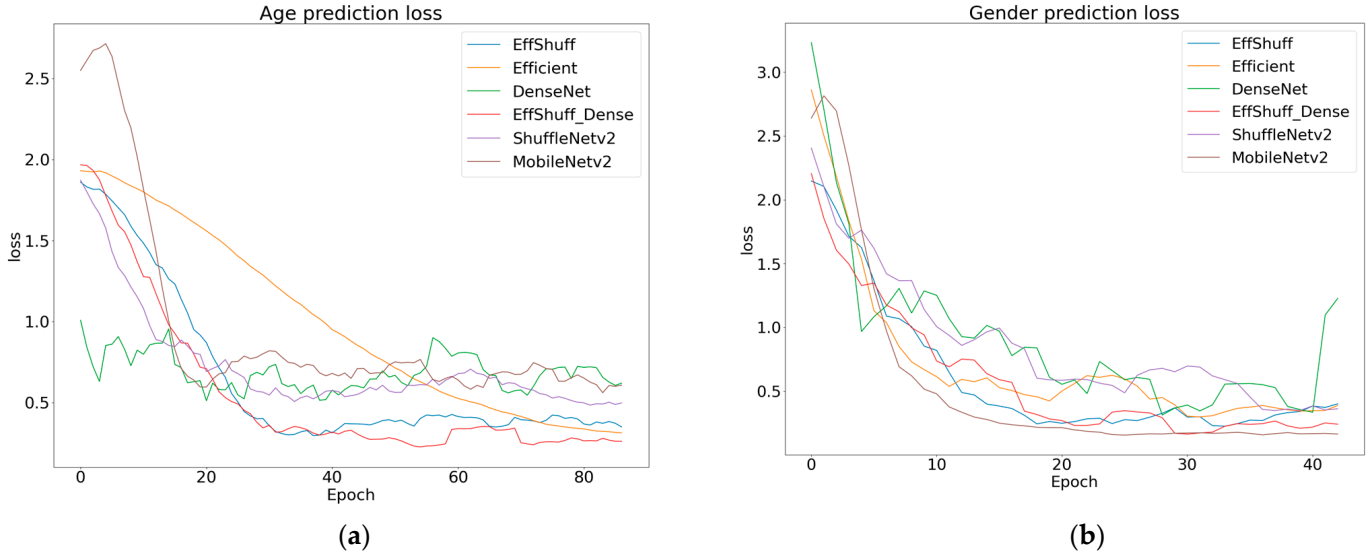

Figure 9. *Cont.*

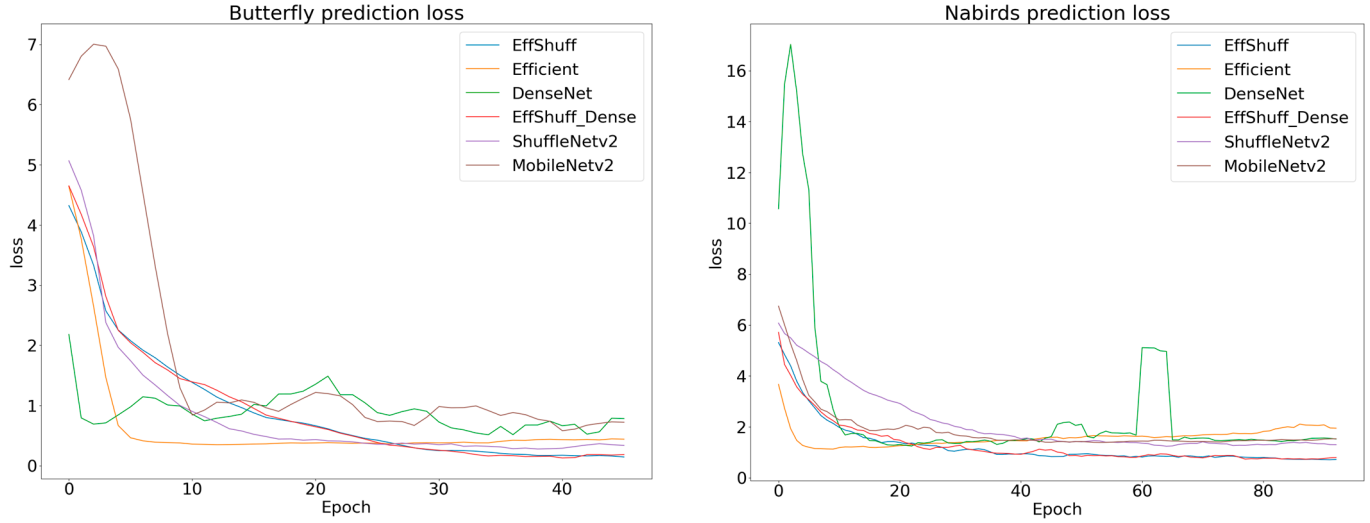

Figure 9. Loss comparison of EffShuffNet, EffShuffDenseNet and other models in validation scheme. (**a**) The Audience dataset: age prediction; (**b**) the Audience dataset: gender prediction; (**c**) the Butterfly & Moths dataset: class prediction; (**d**) the NA Birds dataset: class prediction.

4.4. Ablation Study

An ablation study was conducted to evaluate the effect of the scale of the model on its performance (Figure 10). The scale of the model is determined by the number of parameters, where a scale of one denotes the original model complexity. To reduce the complexity, we scale down the model to a 0.5 scale, while a scale of $x(x > 1)$ represents an increase in the number of parameters by a factor of x. The scale of the model is proportional to the number of channels in each layer, with the width of the model being scaled while maintaining the depth constant. It should be noted that the complexity of our proposed EffShuffNet at all scales is simpler than that of ShuffleNetv2, as we follow the architecture of ShuffleNetv2.

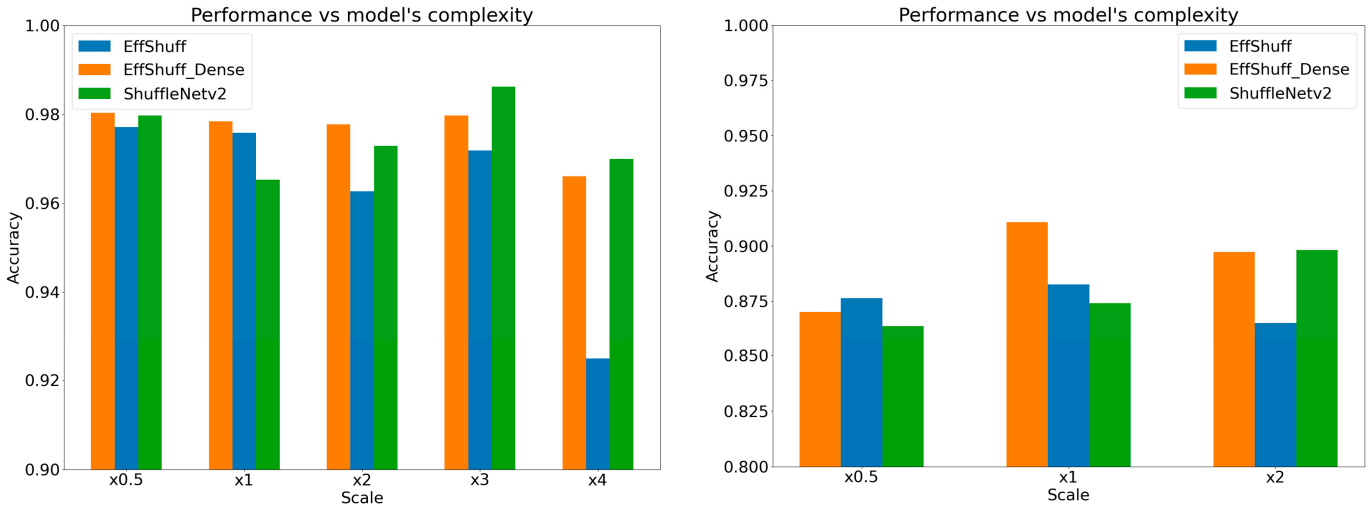

Figure 10. Performance (accuracy) comparison of our proposed EffShuffNet and ShuffleNetv2 with respect to the model's complexity. (**a**) The Audience dataset: age prediction; (**b**) the NA Birds dataset: class prediction.

To compare the performance of our proposed models, EffShuff and EffShuffDense Net, with the corresponding ShuffleNetv2, we conducted experiments on two tasks: age

prediction and NA Birds classification. For the age prediction task, we evaluated the performance of the models at five levels of scale. The experimental results showed that the EffShuffDenseNet outperformed the ShuffleNet at small scales (from $\times 0.5$ to $\times 2$), while the ShuffleNetv2 performed better at larger scales ($\times 3$ and $\times 4$). The results also demonstrated that the EffShuffDenseNet consistently outperformed the EffShuffNet, highlighting the importance of the Dense connection in our building block. Similar experiments were conducted on the NA-Birds dataset, and the results suggest that ShuffleNetv2 is better suited for scaling on large datasets, consistent with the rule that a "complex" dataset requires a "bigger" model. It should be noted that the EffShuffNet utilizes Pooling within a building block, which is a "free parameters" operator that compresses features. Therefore, increasing the number of channels may introduce difficulties for training due to the lack of adjusted weights.

5. Conclusions

The advancement in deep learning technology has led to an increased focus on studies utilizing convolutional neural networks and transformer-based models in image classification. Multi-model methods, which generate classification results from multiple deep learning models, are deemed an optimal solution for multi-label classification tasks where multiple correct labels may exist from a single input. However, it requires the optimization of deep learning models for multiple correct answer labels, and there is a challenge in terms of the exponential increase in computation and computational resources required.

In this study, we present EffShuffNet and EffShuffDenseNet based on convolutional neural networks as lightweight and high-performing multi-models for multi-classification. The EffShuffNet separates the input feature map into two parts, preserves one part as it is, and performs a lightweight convolution operation on the other part before blending the low-level features into the classifier evenly. The EffShuffDenseNet enhances the delivery of low-level features to the classifier without blending, by increasing the output through the growth rate and emphasizing the continuous enhancement of low-level features from the EffShuffNet. To evaluate the performance of these models for multi-classification, we conducted experiments on age and gender classification.

The results of the age and gender classification experiment showed that the EffShuffNet achieved a higher classification accuracy of 96.975%, compared to the 96.775% of the ResNet50, which was the best performer among scratch-based models. In terms of model size and computation, the EffShuffNet demonstrated a significant reduction, with a size of 2.26 M and computational amount of 9.7 G, compared to the 49.18 M and 248.66 G of the ResNet50. Additionally, the EffShuffDenseNet showed further improvement in classification performance with a classification accuracy of 97.63%, although there was a slight increase in model size and computational amount compared to the EffShuffNet model. Nevertheless, both the EffShuffNet and EffShuffDenseNet demonstrated superiority in terms of both model size and computational amount over the MobileNetV2 and ShuffleNetV2, which are representative lightweight convolutional neural network models. The proposed EffShuffNet and EffShuffDenseNet were further confirmed to have a combination of lightweight and high classification accuracy through a fine-grain image classification experiment, which is known to be challenging.

In future work, a fair comparison with the pre-trained models can be established by pre-training the EffShuffNet and EffShuffDenseNet with the ImageNet dataset. An investigation aimed at demonstrating the effectiveness of the DenseNet is also planned.

Author Contributions: Conceptualization, J.-I.K. and G.-H.Y.; methodology, J.-I.K., G.-H.Y. and J.L.; software, J.L. and D.T.V.; validation, J.-Y.K., G.-H.Y. and H.-S.P.; formal analysis, G.-H.Y. and D.T.V.; investigation, J.-I.K. and J.-H.K.; resources, J.-H.K.; data curation, J.-I.K. and S.-H.H.; writing—original draft preparation, J.-I.K.; writing—review and editing, S.-H.H. and J.-Y.K.; visualization, J.L. and H.-S.P.; supervision, S.-H.H. and J.-Y.K.; project administration, J.-H.K., S.-H.H. and J.-Y.K.; funding acquisition, J.-H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Technology Innovation Program (20018905, Development of windows and interior/exterior display technology for autonomous vehicles) funded by the Ministry of Trade, Industry and Energy (MOTIE, Republic of Korea).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study can be found publicly. The Audience dataset is available at https://talhassner.github.io/home/projects/Adience/Adience-data.html (accessed on 6 March 2023). The Butterfly & Moths dataset is available at https://www.kaggle.com/datasets/gpiosenka/butterfly-images40-species (accessed on 6 March 2023), and, the NABirds dataset is available at https://dl.allaboutbirds.org/nabirds (accessed on 6 March 2023).

Acknowledgments: This work was supported by the Technology Innovation Program (20018905, Development of windows and interior/exterior display technology for autonomous vehicles) funded by the Ministry of Trade, Industry and Energy (MOTIE, Republic of Korea).

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

References

1. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 18 August 2009; pp. 248–255.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
4. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
5. Kaiming, H.; Xiangyu, Z.; Shaoqing, R.; Jian, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
6. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. *Attention is All You Need. Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2017.
8. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
9. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
10. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
11. Jongchan, P.; Sanghyun, W.; Joon-Young, L.; So, K.I. BAM: Bottleneck Attention Module. *arXiv* **2018**, arXiv:1807.06514.
12. Sanghyun, W.; Jongchan, P.; Joon-Young, L.; So, K.I. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
13. Alexey, D.; Lucas, B.; Alexander, K.; Dirk, W.; Xiaohua, Z.; Thomas, U.; Mostafa, D.; Matthias, M.; Georg, H.; Sylvain, G.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
14. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021.
15. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
16. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
17. Iandola, I.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
18. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
19. Xiangyu, Z.; Xinyu, Z.; Mengxiao, L.; Jian, S. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.

20. Mark, S.; Andrew, H.; Menglong, Z.; Andrey, Z.; Liang-Chieh, C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
21. Ningning, M.; Xiangyu, Z.; Hai-Tao, Z.; Jian, S. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
22. Jiahui, Y.; ZiRui, W.; Vijay, V.; Legg, Y.; Mojtaba, S.; Yonghui, W. CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv* **2022**, arXiv:2205.01917.
23. Alec, R.; Wook, K.J.; Chris, H.; Aditya, R.; Gabriel, G.; Sandhini, A.; Girish, S.; Amanda, A.; Pamela, M.; Jack, C.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; Volume 139, pp. 8748–8763.
24. Kang, J.; Jeonghwan, G. Ensemble learning of lightweight deep learning models using knowledge distillation for image classification. *Mathematics* **2020**, *8*, 1652. [CrossRef]
25. Hinton, G.; Oriol, V.; Jeff, D. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
26. Raphael, T.; Yao, L.; Linqing, L.; Lili, M.; Olga, V.; Jimmy, L. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *arXiv* **2019**, arXiv:1903.12136.
27. Qizhe, X.; Minh-Thang, L.; Eduard, H.; Quoc, V.L. Self-training with Noisy Student improves ImageNet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10687–10698.
28. Gaojing, W.; Qingquan, L.; Lei, W.; Yuanshi, Z.; Zheng, L. Elderly Fall Detection with an Accelerometer Using Lightweight Neural Networks. *Electronics* **2019**, *8*, 1354. [CrossRef]
29. Wei, L.; Liangchi, Z.; Chuhan, W.; Zhenxiang, C.; Chao, N. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1995–2015.
30. Eran, E.; Roee, E.; Tal, H. Age and Gender Estimation of Unfiltered Faces. *Trans. Inf. Forensics Secur.* **2014**, *9*, 2170–2179.
31. Gerry. Butterfly & Moths Image Classification 100 Species. Available online: https://www.kaggle.com/datasets/gpiosenka/butterfly-images40-species (accessed on 6 March 2023).
32. Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; Belongie, S. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 595–604.