*mathematics*

MDPI

*Article*

# Knowledge Distillation for Enhanced Age and Gender Prediction Accuracy

**Seunghyun Kim [1,†], Yeongje Park [1,†] and Eui Chul Lee [2,*]**

[1] Department of AI & Informatics, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea; 202134015@sangmyung.kr (S.K.); 202331043@sangmyung.kr (Y.P.)
[2] Department of Human-Centered Artificial Intelligence, Sangmyung University, Seoul 03016, Republic of Korea
* Correspondence: eclee@smu.ac.kr
† These authors contributed equally to this work.

**Abstract:** In recent years, the ability to accurately predict age and gender from facial images has gained significant traction across various fields such as personalized marketing, human–computer interaction, and security surveillance. However, the high computational cost of the current models limits their practicality for real-time applications on resource-constrained devices. This study addressed this challenge by leveraging knowledge distillation to develop lightweight age and gender prediction models that maintain a high accuracy. We propose a knowledge distillation method using teacher bounds for the efficient learning of small models for age and gender. This method allows the student model to selectively receive the teacher model's knowledge, preventing it from unconditionally learning from the teacher in challenging age/gender prediction tasks involving factors like illusions and makeup. Our experiments used MobileNetV3 and EfficientFormer as the student models and Vision Outlooker (VOLO)-D1 as the teacher model, resulting in substantial efficiency improvements. MobileNetV3-Small, one of the student models we experimented with, achieved a 94.27% reduction in parameters and a 99.17% reduction in Giga Floating Point Operations per Second (GFLOPs). Furthermore, the distilled MobileNetV3-Small model improved gender prediction accuracy from 88.11% to 90.78%. Our findings confirm that knowledge distillation can effectively enhance model performance across diverse demographic groups while ensuring efficiency for deployment on embedded devices. This research advances the development of practical, high-performance AI applications in resource-limited environments.

**Keywords:** knowledge distillation; age and gender prediction; MobileNet; EfficientFormer

**MSC:** 68T01; 68T07

check for updates

**Citation:** Kim, S.; Park, Y.; Lee, E.C. Knowledge Distillation for Enhanced Age and Gender Prediction Accuracy. *Mathematics* **2024**, *12*, 2647. https://doi.org/10.3390/math12172647

Academic Editor: Raymond Lee

Received: 16 July 2024
Revised: 20 August 2024
Accepted: 23 August 2024
Published: 26 August 2024

**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

In recent years, the ability to accurately predict age and gender from facial images has gained significant traction in various fields, such as personalized marketing, human–computer interaction, social media, security surveillance, advertising, and entertainment. Automatic age and gender estimation plays a crucial role in interpersonal communication and has been a key area of interest for computer vision researchers. It is vital for applications in demographic data collection, surveillance, human–computer interaction, marketing intelligence, and security. Advances in deep learning have notably improved the performance of age and gender prediction models. However, these models often have high computational costs, making them impractical for real-time applications on resource-constrained devices like mobile phones and edge devices. To address this challenge, knowledge distillation has emerged as a promising technique that reduces model size and computational requirements while maintaining prediction accuracy.

Knowledge distillation involves transferring knowledge from a large, complex teacher model to a smaller, more efficient student model [1]. The student model is trained to mimic

the behavior of the teacher model, inheriting its performance but significantly reducing its computational footprint. This process not only helps to achieve efficient models, but also helps to maintain or improve the accuracy of predictions. The main goal of our research is to leverage knowledge distillation to create lightweight age and gender prediction models that can be deployed on low-power devices without compromising performance. While general frameworks for age and gender prediction have been widely studied, there is a significant gap in adapting these models for different demographic groups, especially Asians. Facial features can vary significantly across ethnic groups, and models trained on predominantly non-Asian datasets may not perform optimally on Asian faces. This discrepancy creates a need to develop models that are not only efficient, but are also precisely tuned to specific demographics. Our research addresses this gap by focusing on the optimization of prediction models specific to Asian facial features.

Our research has three main goals: first, we aimed to create a model that can predict age and gender up to a target level. This ensures that the model we create performs up to a level that can be used in practice. Next, we want to make sure that the model is lightweight enough so that it can work in embedded environments. While various studies have produced models that perform very well, most of them are not suitable for porting to embedded environments. This has the disadvantage that they are not applicable to the wide variety of marketing or advertising and human–computer interaction applications that require such features. Therefore, we aimed to use a lightweight model that is so sufficiently diverse that it can be ported to a wide variety of domains. Finally, we ensured that the model was sufficiently tuned for the target race. Most model research works in human-related computer vision are based on Caucasian data, which tend not to perform well enough in other races, indicating that models are sensitive to the demographics of the data they are trained on. However, it is not feasible to create and use the same size of data for all races. Therefore, our third goal was to ensure that models trained on large enough Caucasian data can perform well even when tuned on smaller target races.

To achieve these goals and address the challenges inherent in age and gender prediction across diverse demographic groups, our research introduced a series of innovative strategies designed to optimize both model efficiency and accuracy. These strategies are particularly focused on ensuring that the models we develop are not only lightweight and capable of running on resource-constrained devices but are also finely tuned to perform well across different ethnic groups, particularly Asian populations, which are often underrepresented in existing datasets.

This research introduces three key contributions:

1. Application of Knowledge Distillation: We utilized knowledge distillation to create lightweight models that maintain the performance of complex teacher models, making them suitable for deployment on resource-constrained devices.
2. Demographic-Specific Model Tuning: Our approach focuses on tuning models specifically for Asian facial features, addressing the gap in existing research that primarily targets non-Asian populations.
3. Cross-Demographic Knowledge Transfer: We present a novel strategy to transfer knowledge from a teacher model trained on large-scale, predominantly Caucasian datasets to a smaller dataset representing a different demographic group, such as Asians, ensuring high performance across diverse racial backgrounds.

These contributions collectively enhance the practical applicability of age and gender prediction models, enabling them to be more inclusive, efficient, and versatile in various real-world applications.

## 2. Related Works

### *2.1. Age and Gender Prediction Research*

The task of recognizing age from faces has been studied extensively [2]. H. Zhang et al. [3] proposed an enhanced label distribution learning method for age estimation using Convolutional Neural Networks (CNNs), restricting age-label distribution to cover a limited

range of adjacent ages. Their approach addressed data scarcity issues and demonstrated improved performance, with Mean Absolute Errors (MAEs) of 3.14 on FG-NET and 2.15 on MORPH2 datasets. Y. Deng et al. [4] introduced a multi-feature learning and fusion technique for age estimation, employing subnetworks to gather age, ethnicity, and gender information. Their compact model, requiring only 20 MB, achieved competitive performance with MAEs of 2.47 on MORPH2 and 2.59 on FG-NET, making it suitable for mobile devices. A. Akbari et al. [5] developed a loss function for LDL-based age prediction based on optimal transport theory, which maximizes the impact of closely related ages. Their method yielded MAEs of 1.79 on MORPH2 and 3.41 on FG-NET, effectively utilizing age similarities to improve model performance.

There have also been different kinds of studies that used faces to predict gender. A. Swaminathan et al. [6] proposed a gender categorization method using face embeddings. Their approach, which had not been used for predicting gender, ethnicity, and age, utilized embedding vectors produced by a pre-trained Facenet model. These embeddings were then processed by various machine learning models to predict gender, achieving 97% accuracy on the UTK Faces dataset using K Nearest Neighbors (KNN). M. Alghaili et al. [7] introduced a network combining an NN4 architecture with a Variational Feature Learning (VFL) loss function, focusing on the central portion of faces. This network achieved state-of-the-art performance with 98.23% accuracy on the Adience dataset and performed well on a new dataset of covered or camouflaged faces. M.M. Islam et al. [8] employed Pareto frontier pre-trained CNN networks with transfer learning to develop a gender categorization framework. Using an unconstrained internet image dataset, they demonstrated that GoogLeNet, SqueezeNet, and ResNet50 networks achieved classification accuracies higher than 90% on the WIKI dataset, highlighting the efficiency of Pareto pre-trained models for this task.

There exists a wide range of prior research on gender and age recognition [9], and numerous works exist to predict each attribute separately. However, research has shown that gender and age are interrelated [10,11], and predicting them together can improve performance. Therefore, in this study, we focused on developing a single model that simultaneously predicts gender and age to improve the overall performance of the model. In addition to the expectation of improved performance, we also expect a significant weight saving by using a single model compared to using two models to handle the two tasks.

N. Shin et al. [12] proposed a moving window regression (MWR) algorithm for ordinal regression, which introduced the concept of relative rank (*rho*-rank) to quantify ordinal relations among ranks of input and reference instances. This method iteratively refines rank estimates by selecting reference instances to form a search window, achieving state-of-the-art performance on various benchmark datasets for facial age estimation. Moreover, J. Paplham et al. [13] conducted a comprehensive evaluation of state-of-the-art age estimation methods, highlighting the inconsistencies in the benchmarking process. Their analysis revealed that the performance differences between methods are negligible compared to the effects of other factors such as facial alignment, image resolution, and the amount of pre-training data. They proposed using FaRL as the backbone model and demonstrated its effectiveness across all public datasets. In a more recent study, Maksim Kuprashevich and Irina Tolstykh [14] presented MiVOLO, a multi-input transformer model for age and gender estimation that integrates facial and body information. Their model achieved state-of-the-art performance on five popular benchmarks, demonstrating superior generalization and real-time processing capabilities. Table 1 presents a summary of previous research on recognition of age and gender.

**Table 1.** Summary of recent studies on age and gender recognition.

| Target | Study | Dataset | MAE/Accuracy | Key Findings |
|---|---|---|---|---|
| Age | H. Zhang et al. [3] | FG-NET, MORPH2 | 3.14, 2.15 | Enhanced label distribution learning method for age estimation |
| | Y. Deng et al. [4] | MORPH2, FG-NET | 2.47, 2.59 | Multi-feature learning for compact mobile device models |
| | A. Akbari et al. [5] | MORPH2, FG-NET | 1.79, 3.41 | LDL with optimal transport theory to utilize age similarities |
| Gender | A. Swaminathan et al. [6] | UTK Faces | 97% | Employed face embeddings and KNN for gender prediction |
| | M. Alghaili et al. [7] | Adience | 98.23% | NN4 with VFL; high accuracy on covered faces |
| | M.M. Islam et al. [8] | WIKI | >90% | Used pre-trained CNNs with transfer learning for gender categorization |
| Age and Gender | N. Shin et al. [12] | Various | State-of-the-art | Introduced Moving Window Regression for refining age ranks |
| | J. Paplhám et al. [13] | Various | N/A | Highlighted benchmark inconsistencies; proposed FaRL model |
| | M. Kuprashevich et al. [14] | Various | State-of-the-art | MiVOLO, a multi-input transformer model for age and gender |

### *2.2. Lightweight Models*

To maximize the utility of the two models, we selected lightweight models as the backbone. Convolution Neural Network (CNN) models and transformer-based models are commonly used for processing image-based data. Therefore, in this study, we employed both types of models to conduct training and testing.

MobileNet V3 [15] is a lightweight deep learning model developed by Google and designed for efficient performance on mobile and embedded devices. It builds on the architecture of MobileNet V2 and further improves performance with automatic architecture discovery using Neural Architecture Search (NAS) and platform-aware optimizations. Inspired by EfficientNet, MobileNet V3 is designed to achieve high performance with fewer parameters, and it introduces a Squeeze-and-Excitation (SE) module to improve feature extraction by learning the relationship between channels. It also uses the h-swish function instead of the traditional ReLU activation function to enhance the nonlinearity of the model and improve performance. By maximizing efficiency on various hardware platforms through AutoML-based optimization, MobileNet V3 is able to minimize computational cost and memory usage while maintaining high accuracy, making it effective in mobile environments and systems with limited resources. MobileNet V3 provides large and small models depending on the size of the model, and both models were used in this study.

EfficientFormer [16] is a network proposed to address speed issues on devices limited by the high computation of traditional vision transformer-based models. The model consists of patch embedding and meta-transformer blocks (MBs), which generate output values from input images. The MB is divided into MB 3D, which processes three-dimensional data, and MB 4D, which processes four-dimensional data. MB 3D is based on transformers, while MB 4D is based on CNNs. EfficientFormer also provides models of different sizes depending on the number of MB blocks. In this study, we utilized the EfficientFormer-s2 model.

MiVOLO [14] is a new model that integrates face and body image data to estimate age and gender. Based on the VOLO [17] model, it utilizes a dual input/output architecture to improve generalization and accuracy. The model demonstrates state-of-the-art performance on five popular benchmarks. It also includes a new benchmark dataset annotated by human annotators. The model and code are openly available for further validation and inference. While MiVOLO was originally introduced as a model with both face and body images as input, since the dataset used in this study provides only face data, we utilized the VOLO-D1-based age/gender prediction model based on a single face image as a teacher model. For knowledge distillation on the UTKFace dataset [18], we used a model trained on the same UTKFace dataset as a teacher model, and for the Asian dataset, we used a

model trained on the Internet Movie Database-Clean (IMDB-Clean) dataset. This was due to the different age distributions of the two datasets.

## 3. Materials and Methods

In this section, we describe the overall progression of our research. Initially, we opted for the knowledge distillation technique to develop a highly usable and lightweight model for age and gender prediction. Consequently, we employed a high-performance model from prior research as the teacher model, while MobileNetV3 (small/large) and EfficientFormer were used as student models. The data utilized in this phase came from the UTKFace dataset, which predominantly comprises Western individuals. We implemented the models in a way that, despite their significantly smaller size, they exhibited only minimal performance degradation compared to the heavier teacher models, making them suitable for deployment in lightweight embedded environments.

Subsequently, we conducted distillation on another dataset, the Korean dataset from AIHub. There is a recognized challenge that age and gender prediction models trained on datasets from one ethnicity do not perform adequately on datasets from other ethnicities. Therefore, we utilized the model trained extensively on Western individuals as the teacher model for the relatively smaller-scale Asian dataset. This approach was aimed at ensuring that the distilled knowledge would yield sufficient performance on the Asian dataset tests as well. Detailed descriptions of each step follow below.

### *3.1. Dataset*

#### 3.1.1. UTKFace

UTKFace [18] provides face images for different ethnicities, ages, and genders, and it was used as a training dataset for the teacher and student models in the proposed method. However, the authors of [14] reported that the original UTKFace dataset does not provide a predefined training, validation, and test split, so they split the data by themselves, and the resulting split data have 13,144 training images and 3287 test images with age ranges (from 21 to 60). In this study, the VOLO-D1 model pre-trained on this split was used as the teacher model and distilled with the same split. The age and gender distribution of the UTKFace data split provided by Kuprashevich et al.'s study [14] is shown in Figure 1.

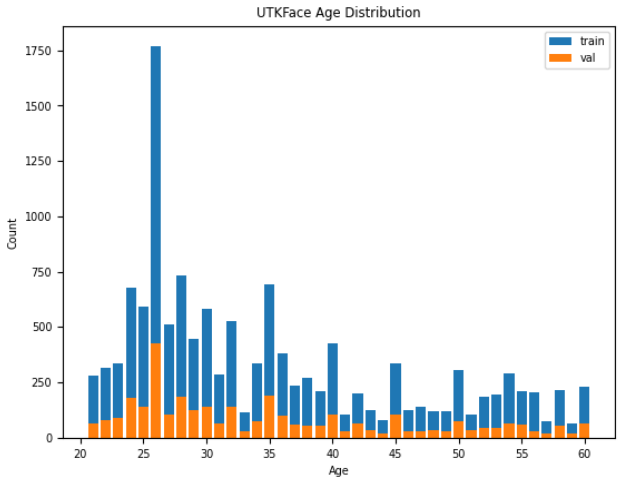



**Figure 1.** UTKFace dataset age distribution.

#### 3.1.2. Asian Dataset

The Asian dataset used in this study is the facial recognition aging image dataset from 'The Open AI Dataset Project (AI-Hub, S. Korea)'. All the data information can be accessed through 'AI-Hub, accessed on 19 August 2024. (www.aihub.or.kr)'. This comprehensive dataset contains facial images of people at different stages of their lives, from infancy to the time the photo was taken. It also provides detailed information about the gender and age

of each person. This dataset is specific to South Koreans, making it ideal for exploring the impact of knowledge distillation using teacher models trained primarily on Westerners. This allows us to scrutinize the performance of these pre-trained models when applied to datasets with significantly different ethnic representations. In this study, we performed training and testing using the data split provided by the dataset as is. The age and gender distributions provided by this dataset are shown in Figure 2.

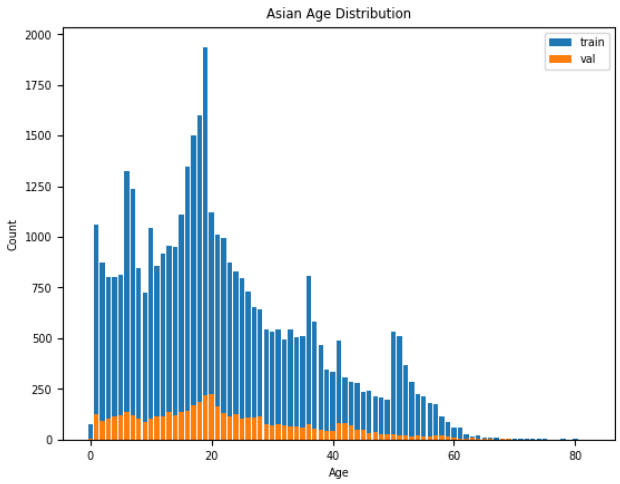



**Figure 2.** Asian dataset age distribution.

### *3.2. Knowledge Distillation*

In order to effectively distill the knowledge of the teacher model into the student model, we used the teacher bounded loss for age regression and gender classification together with the loss for the ground truth label. The teacher bounded loss is a mechanism where the distillation process is applied selectively based on the teacher model's performance relative to the student model. Specifically, when the teacher model's predictions are less accurate than the student model's predictions compared to the ground truth labels, the distillation is not performed for those data points. This approach prevents the student model from learning inaccurate information from the teacher model. By utilizing this method, it becomes possible to achieve more precise learning in challenging situations such as recognition when makeup alters facial features. In these cases, where the teacher model may struggle to make accurate predictions, the student model is not forced to adopt the teacher model's potentially incorrect knowledge, leading to a more robust and clear learning outcome.

The overall flow chart for learning the student model by calculating the teacher bounded loss used in this study is shown in Figure 3. For one dataset, the prediction results of the student model and the teacher model are compared with the label, and if the teacher model performs the prediction better than the student model, the distillation loss using the prediction of the teacher model is added and used in the final loss calculation. If the student model performs the prediction better, only the loss between the prediction of the student model and the label is used to calculate the loss for the corresponding data.

Through this method, it is possible to learn only the useful information of the teacher model in the process of distilling the knowledge of the teacher model into the student model. The overall process of the knowledge distillation method used in this study is as shown in Figure 4, and the total loss equation is as shown in Equation (1). Here, the proposed $\lambda$ value is 0.5.

$$\begin{aligned} L_{\text{total}}(R_s, R_t, C_s, C_t, y) = {} & \lambda L_{\text{MSE},b}(R_s, R_t, y) + \lambda L_{\text{CE},b}(C_s, C_t, y) \\ & + (1-\lambda) L_{\text{MSE},s}(R_s, y) + (1-\lambda) L_{\text{CE},s}(C_s, y) \end{aligned} \tag{1}$$

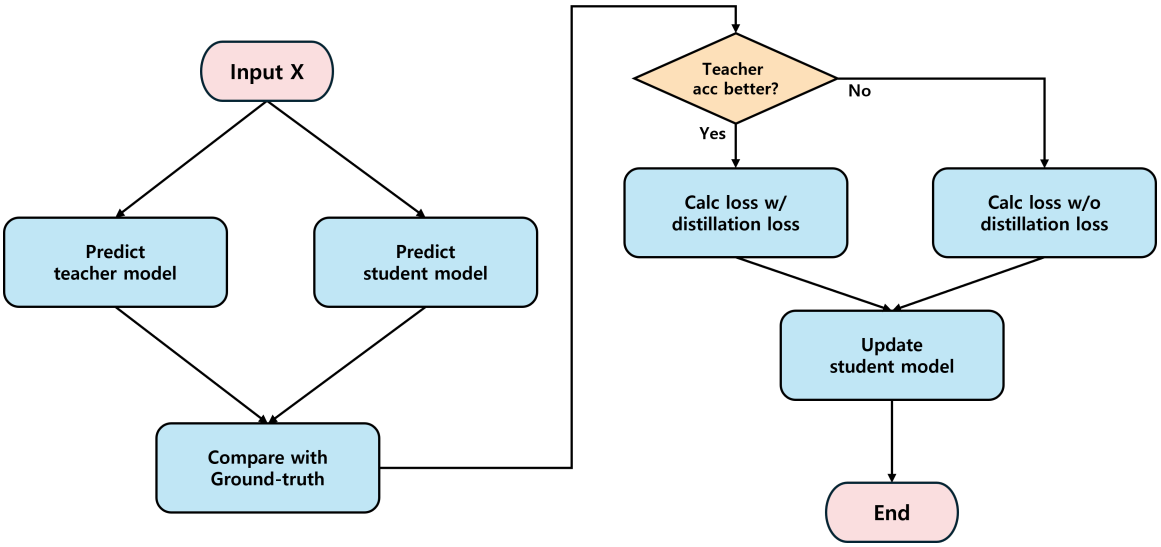



**Figure 3.** Flowchart of the student model weight update process using teacher bounded loss.

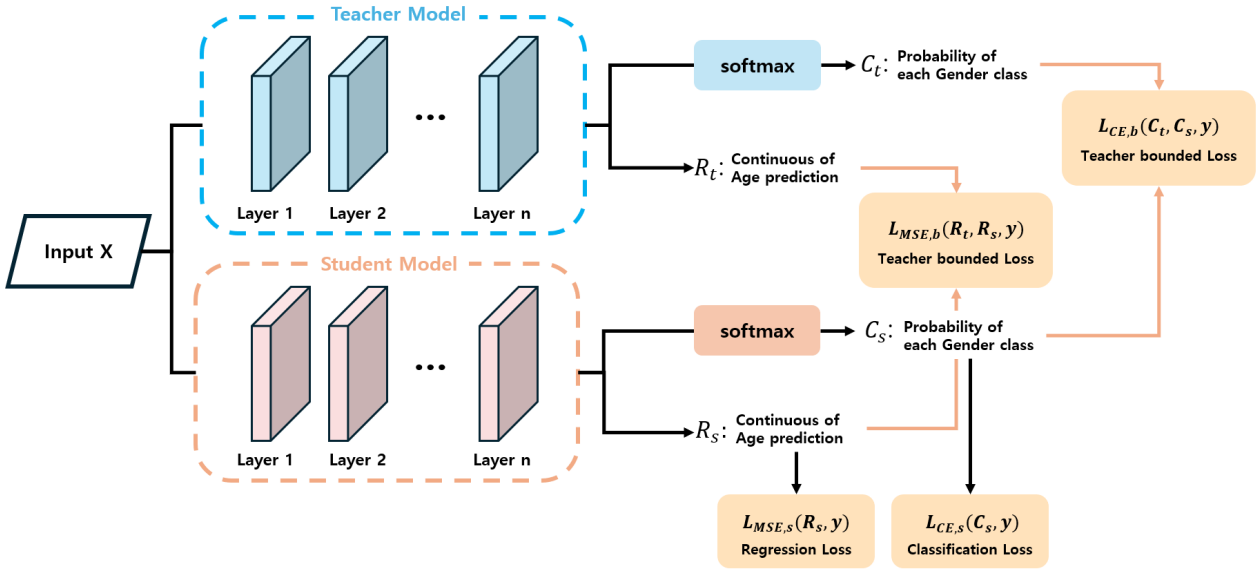



**Figure 4.** Proposed age and gender knowledge distillation method.

The teacher-bounded loss term of total loss has the following equation: $L_{\mathrm{MSE},b}$, $L_{\mathrm{CE},b}$, $L_{\mathrm{MSE},c}$, and $L_{\mathrm{CE},s}$. $L_{\mathrm{MSE},b}$ and $L_{\mathrm{CE},b}$ are the teacher-bounded loss equations for age and gender. Both equations take the predicted value of the teacher model, the predicted value of the student model, and the label as inputs, and for data that the student model predicts better, the corresponding loss term is set to 0 so that the knowledge of the teacher model is not transferred. Equations (2) and (3) represent teacher-bounded Mean Squared Error (MSE) loss equation and cross entropy loss equation, respectively.

$$L_{\mathrm{MSE},b}(R_s, R_t, y) \equiv \begin{cases} \|R_s - R_t\|_2^2, & \text{if } \|R_s - y\|_2^2 > \|R_t - y\|_2^2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$L_{\mathrm{CE},b}(C_s, C_t, y) \equiv \begin{cases} -\sum_{i=1}^{N} C_t^{(i)} \log(C_s^{(i)}), & \text{if } -\sum_{i=1}^{N} y^{(i)} \log(C_s^{(i)}) > -\sum_{i=1}^{N} y^{(i)} \log(C_t^{(i)}) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$L_{\mathrm{MSE},s}$ and $L_{\mathrm{CE},s}$ represent the loss equations between the predicted values of the student model and the label. The MSE loss term and CE term represent the Mean Squared

Error (MSE) and cross-entropy between the predicted value and the label, respectively. The detailed equation is as shown in Equations (4) and (5).

$$L_{\mathrm{MSE},s}(R_s, y) = \frac{1}{N}\sum_{i=1}^{N}(R_s^{(i)} - y^{(i)})^2 \tag{4}$$

$$L_{\mathrm{CE},s}(C_s, y) = -\frac{1}{N}\sum_{i=1}^{N} y_c^{(i)} \log(C_s^{(i)}) \tag{5}$$

## 4. Results

In this study, model learning was performed using MobileNetV3 and EfficientFormer models. The experiment was conducted in two stages for each dataset. The first method was to train each student model using only ground truth labels without a teacher model. This allows us to check the performance of the pure model without knowledge distillation. The second method was to train knowledge distillation using the predictions of the teacher model and ground truth labels together, which allows us to check the performance comparison before and after applying knowledge distillation. Each student model was configured with an individual predictor that predicted age and gender from the input image. The age predictor was configured as a fully connected layer for age regression, and the gender predictor was configured as a fully connected layer for gender classification. MSE Loss was used for age prediction, and Cross Entropy Loss was used for gender prediction. AdamW with a learning rate of $1 \times 10^{-4}$ was used as the optimizer, and the learning rate was set to decrease by 0.1 times every 10 epochs. Training was conducted for a total of 30 epochs, and the student model was trained using the age and gender predicted from the teacher model at each epoch. After training, the performance of the model was evaluated using the validation data, and the MAE for age prediction and the accuracy for gender prediction were recorded.

### *4.1. Comparison of Model Specifications and Efficiency*

In this subsection, we provide a detailed comparison of various models in terms of their number of parameters and Giga Floating Point Operations per Second (GFLOPs). The models under comparison include VOLO-D1, MobileNetV3-Small, MobileNetV3-Large, and EfficientFormer-l1.

The VOLO-D1 model serves as the teacher model in this analysis. We compare each of the other models against this benchmark to highlight the efficiency improvements. The parameters and GFLOPs for each model are listed, along with the percentage reduction in both metrics relative to the VOLO-D1-based age, gender estimation model. This information is critical for understanding the trade-offs between model complexity and computational efficiency, which are particularly important in resource-constrained environments such as mobile and edge computing. Table 2 provides a summary of the comparison. We can see that both MobileNetV3-Small and MobileNetV3-Large show significant reductions in parameters and GFLOPs. EfficientFormer-l1 also shows significant reductions, providing significant reductions in both parameters and computational cost.

**Table 2.** Comparison of model specs and reduction.

| Model | Params | GFLOPs | Param Red. (%) | GFLOPs Red. (%) |
|---|---|---|---|---|
| VOLO-D1(Teacher) | 25.8 M | 6.856 | - | - |
| MobileNetV3-S | 1.48 M | 0.057 | 94.27 | 99.17 |
| MobileNetV3-L | 3.89 M | 0.220 | 84.94 | 96.79 |
| EfficientFormer-l1 | 11.3 M | 1.299 | 55.97 | 81.05 |

The percentage reduction in the number of parameters and GFLOPs is calculated using Equations (6) and (7):

$$\text{Parameter Reduction (\%)} = \left(1 - \frac{\text{Number of Parameters of Target Student Model}}{\text{Number of Parameters of Teacher Model}}\right) \times 100 \quad (6)$$

$$\text{GFLOPs Reduction (\%)} = \left(1 - \frac{\text{GFLOPs of Target Student Model}}{\text{GFLOPs of Teacher Model}}\right) \times 100 \quad (7)$$

### *4.2. UTKFace Distillation*

In the distillation experiment conducted using the UTKFace dataset, the performance of various student models was evaluated. The student models based on MobileNetV3 and EfficientFormer were trained by transferring knowledge from the teacher model based on VOLO-D1, which was pre-trained on the same dataset. Table 3 summarizes the average MAE of age prediction before and after distillation and the accuracy of gender prediction for each model. The Age Cumulative Score at 5 (CS@5) metric represents the proportion of prediction results where the age prediction MAE is less than 5. As a result of the training, it can be confirmed that the overall performance increased after applying the method proposed in this study in all models used as student models. Figure 5 illustrates the impact of the proposed distillation method on the age prediction and gender classification tasks. The model for comparison was EfficientFormer-l1, one of the student models. As shown, the distilled models consistently outperform their non-distilled counterparts in both age prediction MAE and gender classification accuracy across the training epochs.

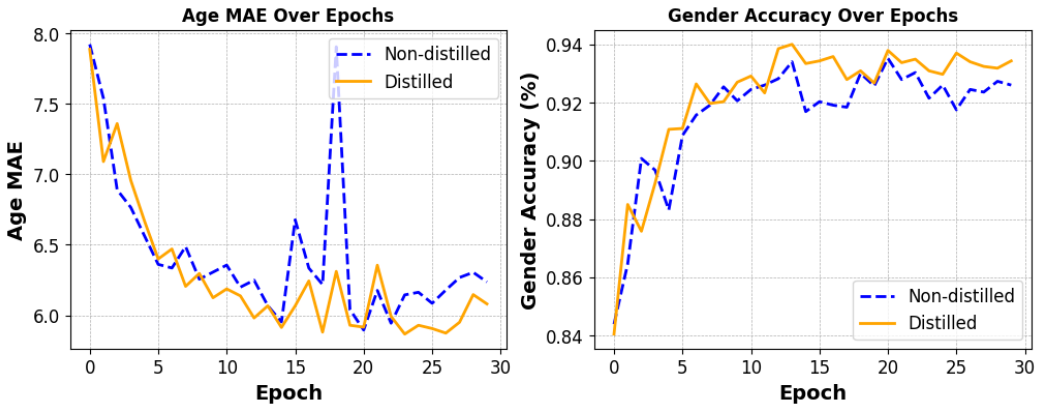



**Figure 5.** Comparison of age prediction MAE (**left**) and gender classification accuracy (**right**) over training epochs for distilled and non-distilled student models.

**Table 3.** Performance comparison of MobileNetV3, EfficientFormer, and teacher models trained on UTKFace dataset.

| | **Teacher Model** | **MobileNetV3-Large** | | **MobileNetV3-Small** | | **EfficientFormer-l1** | |
|---|---|---|---|---|---|---|---|
| | | **Non-Distilled** | **Distilled** | **Non-Distilled** | **Distilled** | **Non-Distilled** | **Distilled** |
| Age MAE | 4.2333 | 5.1943 | 5.0033 | 5.1342 | 5.0693 | 6.2365 | 6.0807 |
| Gender Acc. | 97.69% | 94.64% | 96.23% | 95.37% | 95.44% | 92.60% | 93.43% |
| Age CS@5 | 69.78% | 59.04% | 62.05% | 59.46% | 61.84% | 49.24% | 50.24% |

### *4.3. Qualitative Analysis of Distillation Results*

In addition to the quantitative results presented in Table 3, a qualitative analysis was conducted to further assess the performance of the distilled and non-distilled models. Figure 6 illustrates a comparison of the age and gender predictions made by both models on a selection of images from the UTKFace dataset.

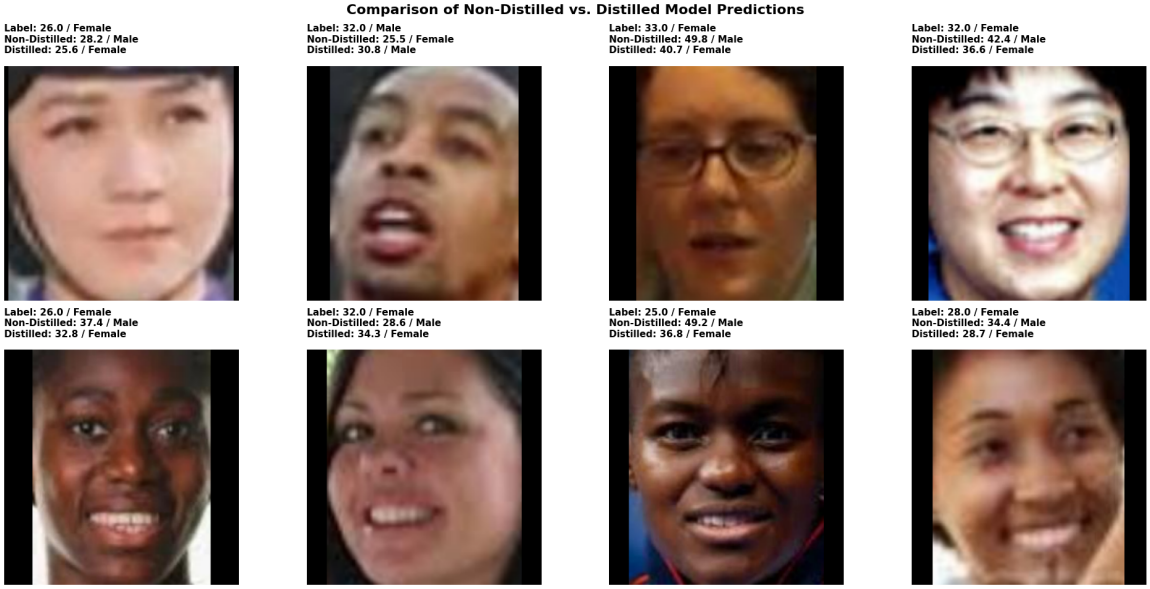



**Figure 6.** Comparison of non-distilled vs. distilled model predictions on UTKFace dataset.

As shown in the figure, the distilled model consistently provides predictions that are closer to the actual labels compared to the non-distilled model. For instance, in several cases, the non-distilled model incorrectly predicts the gender, whereas the distilled model successfully predicts the correct gender. Additionally, the age predictions made by the distilled model are often closer to the true age, demonstrating its superior performance in both age and gender prediction tasks.

Overall, this qualitative analysis further supports the effectiveness of the distillation approach used in this study, highlighting the distilled model's ability to generalize better and produce more accurate predictions compared to the non-distilled model.

### *4.4. Asian Dataset Distillation*

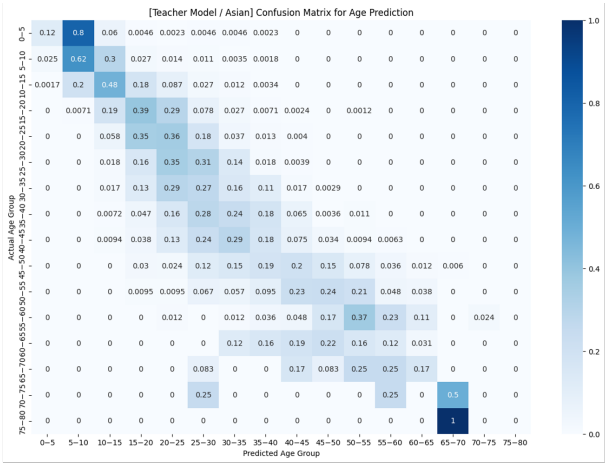
The performance of student models was also evaluated in distillation experiments using Asian datasets. These experiment were conducted with the MobileNetV3 model, which showed good performance in the experiments on the UTKFace dataset, and the teacher model was trained on the IMDB-Clean dataset, which mainly consists of Westerners. Through this comparison, we can confirm the effectiveness of our method of selectively distilling knowledge from teacher models trained on other datasets. The table below summarizes the average L1 loss of age prediction and gender prediction accuracy before and after distillation for each model. As shown in Table 4, we can confirm that the knowledge distillation method for teacher models on other datasets showed better performance when distillation was applied.

**Table 4.** Performance comparison of MobileNetV3-Small trained with Asian dataset.

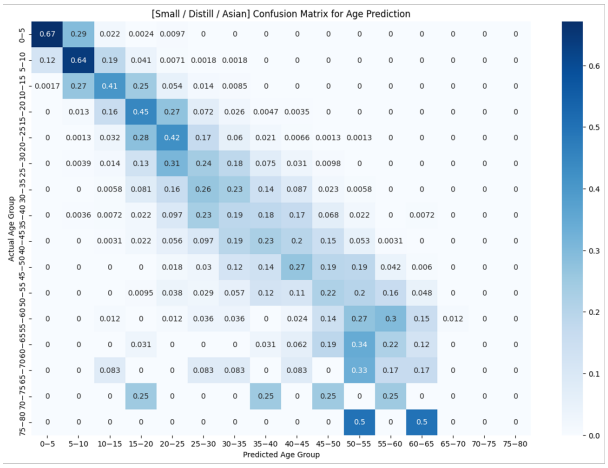
| | **MobileNetV3-Small** | | **Teacher Model** |
|---|---|---|---|
| | **Non-Distilled** | **Distilled** | |
| Age MAE | 4.5860 | 4.5449 | 5.2815 |
| Gender Acc. | 90.68% | 90.78% | 88.11% |
| Age CS@5 | 67.39% | 67.81% | 60.87% |

### *4.5. Age Confusion Matrix*

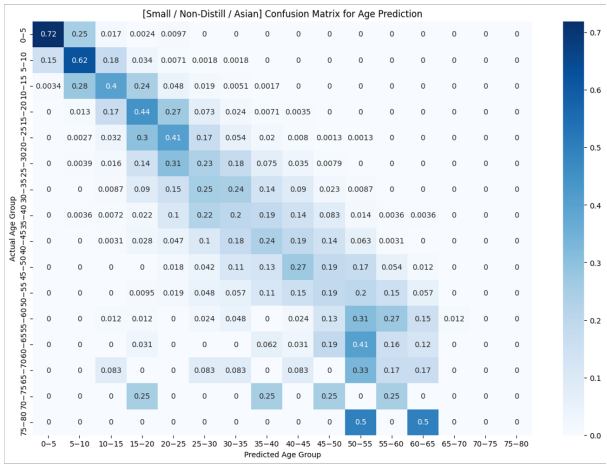
The confusion matrix in Figures 7 and 8 shows the comparison between the model's prediction and the actual label by dividing the age group into 5-year units. In the case of the UTKFace dataset, the confusion matrix was created by dividing it into eight age groups because the age range of the dataset used for learning was from 20 to 60. In the case of the Asian dataset, the comparison was performed with a total of 16 age groups for the same reason. A high accuracy is expressed in dark blue, and a low accuracy is expressed in a lighter color. Through this comparison, we can more intuitively compare the performance before and after knowledge distillation and further confirm whether the knowledge of the teacher model was normally transferred to the student model.




(a)




(b)

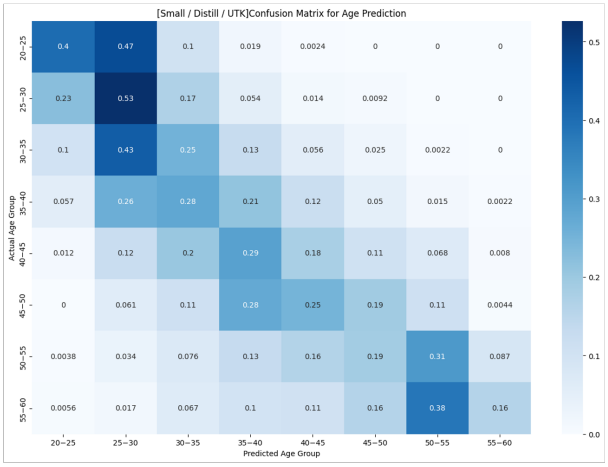



(c)

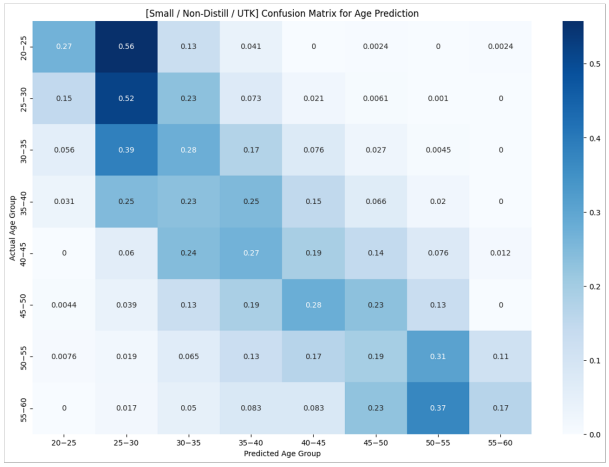
**Figure 7.** Confusion matrices for age estimation using models trained on the Asian dataset combined. (**a**) Results from the teacher model; (**b**,**c**) Results for MobileNetV3-Small. (**b**) No use of knowledge distillation; (**c**) Trained using knowledge distillation.



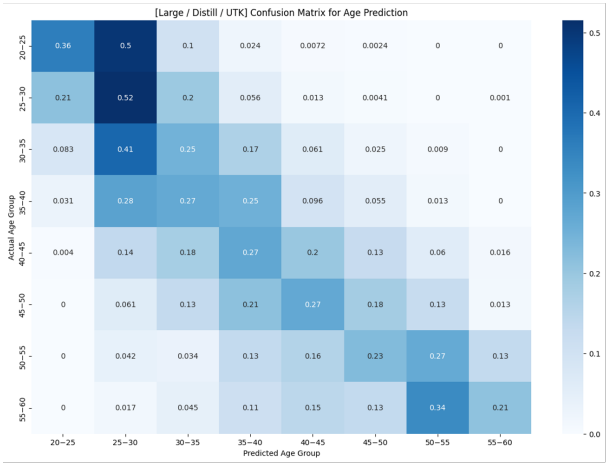

(a)




(b)




(c)




(d)




(e)

**Figure 8.** Confusion matrices for age estimation using models trained on the UTKFace dataset combined. (**a**) Results from the teacher model; (**b**,**c**) Results for MobileNetV3-Small; and (**d**,**e**) Results for MobileNetV3-Large. (**c**,**e**) No use of knowledge distillation; (**b**,**d**) Trained using knowledge distillation.

## 5. Discussion

The primary objective of this research was to develop lightweight models for age and gender prediction that can be deployed on resource-constrained devices while maintaining a high accuracy. Our study focused on leveraging knowledge distillation to achieve these goals, evaluating performance across different demographic datasets to ensure generalizability and robustness. Our experiments demonstrated that knowledge distillation significantly enhances the efficiency of student models without a substantial loss in accuracy. The comparison of various models, including MobileNetV3-Small, MobileNetV3-Large, and EfficientFormer-l1, against the VOLO-D1 teacher model highlighted remarkable reductions in the number of parameters and GFLOPs. For instance, MobileNetV3-Small exhibited a 94.27% reduction in parameters and a 99.17% reduction in GFLOPs compared to the teacher model, making it highly suitable for deployment on embedded systems. These reductions are crucial for real-time applications on devices with limited computational resources, such as mobile phones and edge devices.

A notable aspect of our research is the focus on optimizing models for specific demographic groups, particularly Asian facial features. This addresses a critical gap in the existing literature, where models predominantly trained on Western datasets perform suboptimally on other ethnic groups. Our experiments with the Asian dataset confirmed that knowledge distillation can effectively transfer knowledge from a teacher model trained on a different demographic, enhancing performance. For instance, the distilled MobileNetV3-Small model showed an improvement in gender prediction accuracy from 88.11% (teacher model) to 90.78%, demonstrating the method's effectiveness. The research also explored the broader applicability of knowledge distillation in transfer learning scenarios. By training student models with the predictions of a pre-trained teacher model, we ensured that the distilled models inherited the performance characteristics of the teacher model while being significantly more efficient. This approach is particularly advantageous in scenarios where training data for specific demographics is limited. By utilizing a robust teacher model trained on a larger, diverse dataset, we can achieve a high performance on smaller, demographic-specific datasets. The confusion matrices for age estimation on both the UTKFace and Asian datasets further validated the effectiveness of our approach. The models trained with knowledge distillation showed an improved accuracy across different age groups, with darker blue regions indicating a higher prediction accuracy. This visual representation confirms that knowledge distillation not only enhances overall performance but also improves consistency across various age ranges.

## 6. Conclusions

In conclusion, this study presents a significant advancement in the development of lightweight age and gender prediction models for resource-constrained environments. Our approach, centered on knowledge distillation, demonstrates the capability to reduce model size and computational complexity while maintaining a high accuracy. The research outcomes indicate that these models can be effectively deployed on embedded devices, facilitating real-time applications in fields such as personalized marketing, human–computer interaction, and security surveillance. The substantial reduction in parameters and GFLOPs for models like MobileNetV3-Small and EfficientFormer-l1 highlights the efficiency gains achieved through knowledge distillation. Our focus on optimizing models for Asian facial features addresses a significant gap in existing research, ensuring a high performance across diverse demographic groups. The successful application of knowledge distillation for transfer learning underscores its potential in scenarios with limited demographic-specific training data. The high accuracy rates achieved for gender prediction and low MAE values for age estimation confirm the robustness of the distilled models. To further enhance model's robustness and applicability, future research will focus on incorporating a broader range of ethnic diversity in training datasets and extending the approach to include more diverse groups. By testing and optimizing lightweight models with real-time faces from different countries, we aim to reduce potential biases and ensure reliable performance

across various demographic groups in dynamic, real-world environments. Additionally, future research could explore several avenues to further enhance the applicability and performance of lightweight models. Extending the approach to include more diverse ethnic groups and verifying performance across these groups, testing and optimizing these models in real-world scenarios to assess their performance and reliability in dynamic environments, combining knowledge distillation with other model compression techniques to achieve even greater efficiency, and investigating the potential of newer model architectures that may offer improved performance or efficiency benefits when combined with knowledge distillation are all promising directions. By addressing these future directions, we can continue to push the boundaries of what is achievable with lightweight, efficient deep learning models, making sophisticated AI accessible on even the most resource-constrained devices.

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

**Funding:** This work was supported by the NRF (National Research Foundation) of Korea funded by the Korea government (Ministry of Science and ICT) (RS-2024-00340935).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author due to (Restrictions apply to the availability of these data).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Lee, K.; Kim, S.; Lee, E.C. Fast and accurate facial expression image classification and regression method based on knowledge distillation. *Appl. Sci.* **2023**, *13*, 6409. [CrossRef]
2. Angulu, R.; Tapamo, J.R.; Adewumi, A.O. Age estimation via face images: A survey. *EURASIP J. Image Video Process.* **2018**, *2018*, 42. [CrossRef]
3. Zhang, H.; Zhang, Y.; Geng, X. Practical age estimation using deep label distribution learning. *Front. Comput. Sci.* **2021**, *15*, 153318. [CrossRef]
4. Deng, Y.; Teng, S.; Fei, L.; Zhang, W.; Rida, I. A multifeature learning and fusion network for facial age estimation. *Sensors* **2021**, *21*, 4597. [CrossRef] [PubMed]
5. Akbari, A.; Awais, M.; Fatemifar, S.; Khalid, S.S.; Kittler, J. A novel ground metric for optimal transport-based chronological age estimation. *IEEE Trans. Cybern.* **2021**, *52*, 9986–9999. [CrossRef] [PubMed]
6. Swaminathan, A.; Chaba, M.; Sharma, D.K.; Chaba, Y. Gender classification using facial embeddings: A novel approach. *Procedia Comput. Sci.* **2020**, *167*, 2634–2642. [CrossRef]
7. Alghaili, M.; Li, Z.; Ali, H.A. Deep feature learning for gender classification with covered/camouflaged faces. *IET Image Process.* **2020**, *14*, 3957–3964. [CrossRef]
8. Islam, M.M.; Tasnim, N.; Baek, J.H. Human gender classification using transfer learning via Pareto frontier CNN networks. *Inventions* **2020**, *5*, 16. [CrossRef]
9. Ghrban, Z.; Abbadi, N.K.E. Gender and Age Estimation from Human Faces Based on Deep Learning Techniques: A Review. *Int. J. Comput. Digit. Syst.* **2023**, *14*, 201–220. [CrossRef] [PubMed]
10. Grd, P.; Barčić, E.; Tomičić, I.; Okreša Đurić, B. Analysing the Impact of Gender Classification on Age Estimation. In Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference, Stavanger, Norway, 14–15 June 2023; pp. 134–137.
11. Di Mascio, T.; Fantozzi, P.; Laura, L.; Rughetti, V. Age and gender (face) recognition: A brief survey. In *Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference*; Springer: Cham, Switzerland, 2022; pp. 105–113.
12. Shin, N.H.; Lee, S.H.; Kim, C.S. Moving window regression: A novel approach to ordinal regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 18760–18769.
13. Paplhám, J.; Franc, V. A Call to Reflect on Evaluation Practices for Age Estimation: Comparative Analysis of the State-of-the-Art and a Unified Benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 1196–1205.
14. Kuprashevich, M.; Tolstykh, I. Mivolo: Multi-input transformer for age and gender estimation. In *International Conference on Analysis of Images, Social Networks and Texts*; Springer: Cham, Switzerland, 2023, pp. 212–226.

15. Koonce, B.; Koonce, B. MobileNetV3. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Apress: Berkeley, CA, USA, 2021; pp. 125–144.
16. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 12934–12949.
17. Yuan, L.; Hou, Q.; Jiang, Z.; Feng, J.; Yan, S. Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6575–6586. [CrossRef] [PubMed]
18. Zhang, Z.; Song, Y.; Qi, H. Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5810–5818.

**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.