2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS) | 978-1-6654-6773-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPICS55264.2022.9873621

# Age Estimation with Synthetic Mask Generation Based on MobileNet and Facial Keypoint Detection

Yequan Bie
School of Automation Science and Engineering
South China University of Technology
Guangzhou, China
yequanbie@gmail.com

Lingjun Liao
School of Animation and Digital Arts
Communication University of China
Beijing, China
ryleeling@gmail.com

Xinyi Ling
School of Computer Science
Wuhan University
Wuhan, China
xinyilingd@gmail.com

Xiaolong Zhu
School of Computer Science
Wuhan University
Wuhan, China
xiaolongzhu620@gmail.com

Xiaxue Ouyang
School of Mathematics
Wuhan University
Wuhan, China
xiaxueouyang3@gmail.com

Junhui He
School of Computer Science
Wuhan University
Wuhan, China
zeonfaiho@gmail.com

Yize Tong
School of Electronic and Electrical Engineering
Southwest Jiaotong University
Sichuan, China
yizelouis@gmail.com

Teoh Teik Toe
College of Business
Nanyang Technological University
Singapore
ttteoh@ntu.edu.sg

***Abstract*—Facial age estimation is one of the most important tasks in the field of face recognition and recommendation system. Since the COVID-19 pandemic, people have been required to wear masks, which can be a challenge for traditional recognition methods. In this paper, an improved convolutional neural network architecture based on MobileNet is proposed to perform age estimation. For the challenge of masked faces, an innovative mask generation method using face keypoint detection is adopted, extracting the key points of the faces in order to add synthetic masks to simulate the real situations. Then we compare the estimation results of the original images and the synthetic images. Our method is applied to the WIKI Face dataset containing more than 150,000 images, and achieves MAE of 3.79 and 6.54 on unmasked and masked faces, respectively, which demonstrates the effectiveness of the proposed model.**

***Keywords*—Age Estimation, Keypoint Detection, MobileNet, Mask Generation, Deep Learning**

## I. Introduction

Facial age recognition is a task that extracts facial features from facial images to estimate the approximate age of a person. In recent years, the number of practical cases related to age analysis has increased and its application value has been further highlighted by many fields focusing on human aging research, where automatic facial age estimation plays an important role in the huge potential applications of human-computer interaction.

Since people of different age groups have significant differences in consumption habits, entertainment styles, aesthetic needs, etc., if the user's facial image can be recognized and thus the approximate age of the user can be estimated, it is possible to provide information suitable for people of specific age groups and avoid inappropriate services. For example, it can be used to prevent teenagers from accessing undesirable information on the Internet and minors from buying cigarettes, alcohol or adult products from vending machines. Besides, better results can be obtained by placing targeted advertisements on the Internet.

Most existing methods for facial age estimation typically use hand-crafted feature descriptors such as LBP and AAM to represent faces, which requires strong prior knowledge to design them. Jiwen Lu proposed a cost-sensitive local binary multi-feature learning (CS-LBFL) [8] to learn multiple sets of hash functions of facial patches to exploit complementary information and improved performance by learning the distinguished local facial descriptors directly from the original pixel values of the facial representation. Ningning Yu proposed an integrated learning strategy that fuses weak classifiers and gains a strong one to perform age classification [9]. Experimental results show that the algorithm can improve an exact match (AEM) and an error of one age category (AEO) by 4.75% and 6.75%, respectively, compared to the best of the three weak classifiers. Fares Alnajar found that expression dynamics can be used to better estimate a person's age and proposed a fully automated age estimation framework that outperformed the generic method [10]. Lou proposed a deep adversarial metric learning (DAML) method to address this problem, which learns to synthesize semi-hard negatives based on existing training samples [11].

Since the COVID-19 pandemic, people have been required to wear masks in most scenarios. Age estimation for face images with masks has become necessary, however, most previous researches on various face recognition techniques have focused on face images without masks. Therefore, in order to investigate whether deep learning method can still make relatively accurate estimates of facial age even when recognizing images of faces wearing masks, a mask generation method is proposed in this paper. Our model is applied to both unmasked and mask-generated face image dataset.

This paper aims to provide some unique insights and innovative approaches to the study of age estimation both on clear and obscured facial images. For age estimation, a new convolutional neural network architecture based on improved MobileNet is proposed. For mask generation, a face key point detection method using ensemble learning is applied to add synthetic masks to images of human faces. Our model is applied to both the original dataset and the synthetic dataset. In addition, in order to optimize our model, innovative training stratigies are adopted, and our model successfully achieves high performance.

## II. DATASET & DATA PRE-PROCESSING

### A. Description

In this paper, Facial Age dataset which was prepared from WIKI Face dataset is applied to evaluate our model, and each of the faces is from images used in Wikipedia [18].

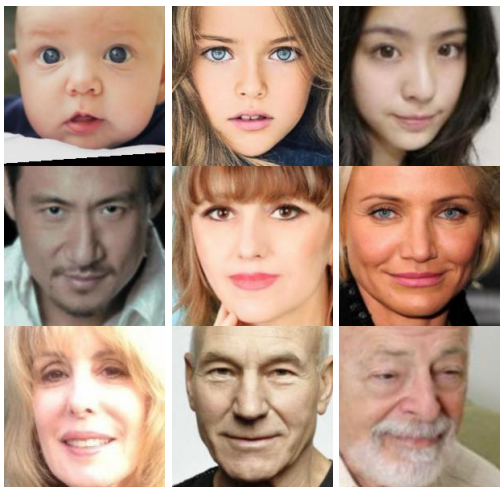


Fig. 1. Examples of Facial Age dataset. These examples correspond to face images of 1, 10, 20, 30, ..., 100 years old respectively.

The dataset contains 100 categories, corresponding to 1-100 years old face images, with a total of 158,208 images. Some examples of the dataset are shown in Fig. 1.

### B. Data Augmentation & Cross Validation

In the field of computer vision based on deep learning, we often need to use models with very deep network layers for image recognition, which results in large number of parameters. When the model is complex and data is not sufficient, there will be an overfitting problem, that is, the model over-learns the data in the training set, resulting in the model performing well on the training set, but losing the ability to generalize, i.e. the model performs poorly on the test set. To this end, this paper uses data augmentation to alleviate overfitting. The adopted data augmentation methods include the following.

- Center crop
- Horizontal flip
- Hue saturation value adjustment
- Resize

For dataset spliting, we adopt 4-fold cross-validation. That is to say, the dataset is randomly divided into four parts, and each time one part (i.e. 25%) is taken out as the validation set, and the remaining data is used as the training set. A total of four complete training sessions are performed and tested separately using the corresponding validation set. Finally, the average of the four evaluation results is regarded as the performance of the model. The benefit of using K-fold cross validation is that it can take advantage of data, alleviate overfitting to a certain extent, and reduce generalization errors as well.

### C. Data Pre-processing for Mask Generation

With increasing use of masks during the prevailing of the pandemic, we are interested in the accuracy of our model while predicting the age of an unmasked biological face. To reduce the difficulty of data acquisition and improve model robustness, we chose the WIKI Face dataset and proposed a procedure based on [5] to digitally generate masks on our selected face images. An overall of the proposed workflow is shown in Fig. 2. and the overall synthesis process is shown in Fig. 3.

*1) Basic Facial Key Points Detection:* In general, an unmasked face contains many biological facial features. Firstly, we use the face detector [6] to crop the face and secondly apply a basic facial key points detector to extract such basic facial features. The basic facial key points detector uses a three-stage convolutional network to determine the location of keypoints [13]. The first stage of convolutional neural network is used for the initial detection, and it makes an approximate estimation even if there are regions that are unclear or blurred in the image. The CNN of the first stage takes the whole face image as input and considers the texture of the whole picture so as to extract the features of the whole face. The second stage and the third stage, on the other hand, serves to refine the extracted key points and precise the results. Networks of these two stages are shallower than that of first stage. Both the second and third stage employ a separately trained convolutional network to pinpoint its local regions of the face. The same regressor is applied at different cascade stages. In this model, the key points are detected by refining the left eye and the right eye center, the nose tip, the left and the right mouth corner. In the first stage, there are four convolutional layers followed by max pooling, and two fully connected layers. The local features extracted by the first stage are input to the second and third stage. The prediction results of the first stage can only be improved in a limited way during training, and the accuracy increases gradually as the cascade increases. The

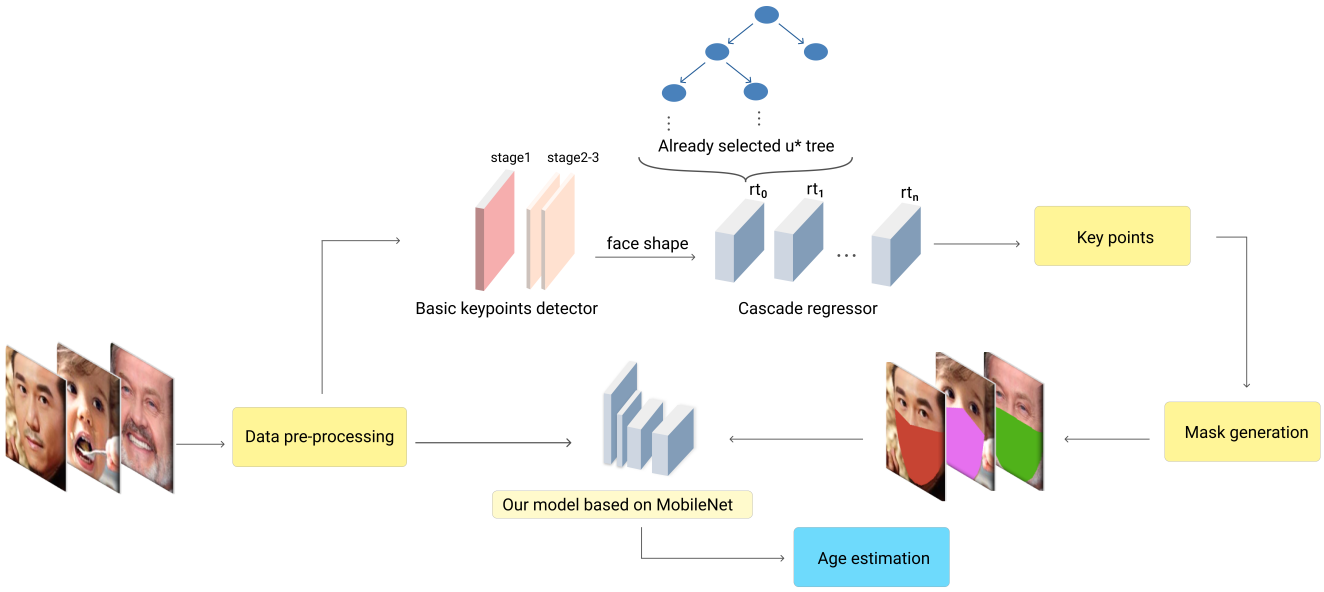



Fig. 2. Overview of the proposed workflow

sizes of patches and search ranges keep reducing along the cascade. In our whole keypoint detector, the basic keypoints will be used as face shape input, further increasing the number of keypoints to 68, such as the nose bridge and facial contour points. The details will be mentioned in the Method part.

*2) Digitally Synthetizing Masks:* For wide masks, specific facial key points we have detected are selected to be sequentially concatenated to generate a closed geometric polygon, which will have the shape of a mask. If the keypoints are too far apart, additional points are inserted between them in order to smooth out the lines.

Based on this proposed workflow, 68 facial key points around the eyes, mouth, nose, face contour etc. are detected. In this paper, we select keypoints around the face contour and the nose as the base points for our mask generation, and the selected points are shown in Fig. 3(b).

## III. Method

### A. Key point detection

The features extracted by basic facial keypoints detector are used as the initial face shape input and then employ a cascaded regressor to accurately estimate the feature keypoints of the face. In this section we refine and extract the previously available face shape based on the idea of an ensemble of regression trees [17]. Suppose the vector of facial shape $S = (x_1^T, x_2^T, \ldots, x_p^T)^T \in \mathbb{R}^{2p}$ denotes the coordinates of all $p$ facial landmarks in images. $\widehat{S}^{(t)}$ denotes our current estimate of $S$. Each regressor $r_t$ in the cascade predicts an update vector, and is added to $\widehat{S}^{(t)}$ to improve estimation iteratively.

$$\widehat{S}^{(t+1)} = \widehat{S}^{(t)} + r_t(I, \widehat{S}^{(t)}) \tag{1}$$

In order to train each regressor $r_t$ , the gradient tree boosting algorithm is adopted. $I_i$ is a face image and $S_i$ mentioned earlier is its shape. For the purpose to learn the first regression function in the cascade $r_0$, three elements were constructed from our data including $(I_{\pi_i}, \widehat{S}^{(0)}, \Delta S_i^{(0)})_{i=1}^N$ as a set, which means an image, an initial shape estimation and an target update step, respectively (i = 1, …, N).

$$\pi_i \in \{1, \ldots, n\} \tag{2}$$

$$\widehat{S}_i^{(0)} \in \{S_1, \ldots, S_n\} \setminus S_{\pi_i} \tag{3}$$

$$\Delta S_i^{(0)} = S_{\pi_i} - \widehat{S}_i^{(0)} \tag{4}$$

In order to learn the regression function $r_t$, we first initialize it using the following equation.

$$f_0(I, \widehat{S}^{(t)}) = \underset{\gamma \in \mathbb{R}^{2p}}{arg\ min} \sum_{i=1}^{N} ||\Delta S_i^{(t)} - \gamma|| \tag{5}$$

Then for k = 1, …, K, (6) and (7) are performed iteratively.

$$r_{ik} = \Delta S_i^{(t)} - f_{k-1}(I_{\pi_i}, \widehat{S}_i^{(t)}) \tag{6}$$

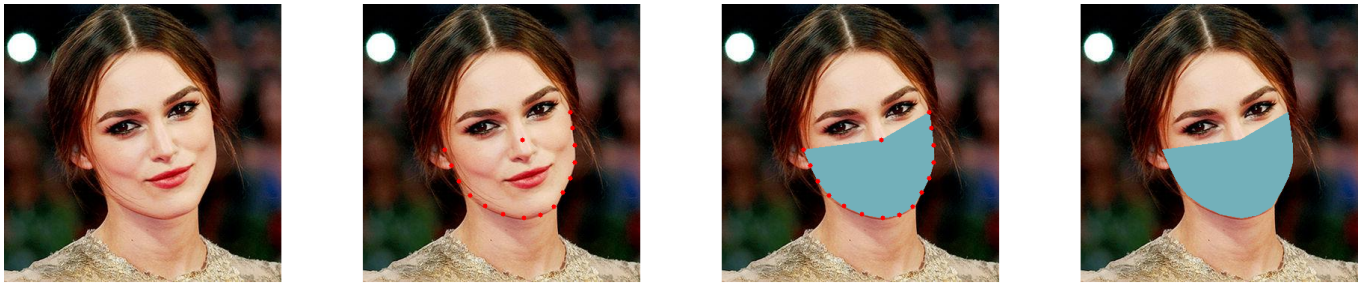
$$f_k(I, \widehat{S}^{(t)}) = f_{k-1}(I, \widehat{S}^{(t)}) + v g_k(I, \widehat{S}^{(t)}) \tag{7}$$

where $g_k(I, \widehat{S}^{(t)})$ is a weak regression function by fitting a regression tree to the target $r_{ik}$. $v$ is the learning rate, setting $v < 1$ helps combat overfitting.

The output $r_t$ after the training iteration is as follow.

$$r_t(I, \widehat{S}^{(t)}) = f_K(I, \widehat{S}^{(t)}) \tag{8}$$

During the node splitting process of the tree in gradient tree boosting alorithm, each node generates some sets of random candidates $\mu$. The decision to generate a split node is determined based on the intensity difference between two pixels. A split threshold is randomly generated after the pixel difference between the pixel values at these two points of each



(a) Original image (b) Facial keypoints (c) Digital mask added (d) Masked face image

Fig. 3. Mask generation

image is calculated. All images are judged by the threshold and segmented into two parts itratively. Finally the candidate group $\mu$ is generated. Then the next step is to select the best $\mu^*$ from the candidates $\mu$. The criterion to judge whether the split is good or not is the magnitude of the variance. That is, we choose the node $\mu^*$ with the smallest squared error sum among the candidates, since they are more likely to belong to the same class. The subsequent splitting of each node follows this step until the splitting reaches a leaf node.

### B. MobileNet

As one of the most popular convolutional neural networks in recent years, MobileNet [1] has the features of small size and high efficiency. Compared with traditional convolutional neural networks, MobileNet replaces the standard convolutional filters with the depthwise separable convolutional filters [4], which are used to decompose and recombine features in an efficient way.

The depthwise separable convolution would split the standard convolution into a depthwise convolution, which is to convolve each channel separately without changing the depth of the input feature image, and a pointwise convolution, namely, a 1×1 convolution is adopted to ascend or descend input feature's dimension as shown in Fig. 4.

For the same $D_F \times D_F \times M$ input feature map and producing $D_F \times D_F \times N$ output map, the standard convolutional layers with a size of $D_K \times D_K \times M \times N$ have the computational cost of $D_K \times D_K \times M \times N \times D_F \times D_F$. However, the depthwise separable convolution can have the cost only of $D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F$.

By expressing convolution as a two step process of filtering and combining we get a reduction in computation of

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_k^2} \tag{9}$$

Under the condition of obtaining similar results, depthwise separable convolution can reach 9 times less parameters and computation than standard convolution, which greatly shortens the latency caused by traditional networks. In addition, MobileNet V2 also uses linear bottleneck as activation function. This allows neural networks to be more robust in the case of low precision computations. Similarly, the residual structure brought by shortcut allows MobileNet to reuse image features.

### C. Improved Architecture

In this paper, we treat age estimation as a regression task. In order to make mobilenet perform better on age estimation prediction, a new network architecture based on MobileNet is proposed in this paper. We innovatively modify some network structures of MobileNet, the body architecture of our model is shown in I, including the following items:

- Batch normalization
  We apply batch normalization both before the input and the output layer. In order to address the effects of internal covariate shift issues ,we normalize inter-layer input values which are performed on small batches of training data. It can greatly imcrease the convergence speed of training and also allows less careful initialization of parameters.

- Global average pooling
  A global average pooling [14] layer is applied after the last point-wise convolution layer. GAP obtains a single value by averaging all the pixel values of the feature map. It can be regarded as a replacement for the fully connected layer in order to reduce the number of parameters and allievates overfitting.

- Fully connected layer and dropout
  We only use a fully connected layer as the output layer of the network without using softmax, which is more suitable for regression task. As an important regularization method, dropout can effectively allievate the problem of overfitting. It can be regarded as bagging a large number of neural networks with shared parameters. Thus, a dropout strategy is applied to the output fully connected layer.

## IV. EXPERIMENTS

### A. Evaluation Metrics

The performance of models is measured by Mean Absolute Error (MAE), which is calculated using the average absolute

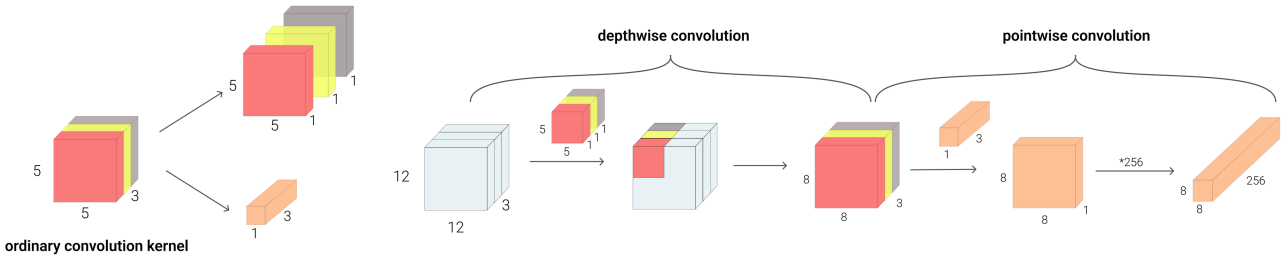



Fig. 4. Depthwise and pointwise convolution of MobileNet

TABLE I
MODEL BODY ARCHITECTURE

| Layer (type) | Output Shape | Param |
|---|---|---|
| batch_normalization | (100, 100, 3) | 12 |
| input | (100, 100, 3) | 0 |
| conv1 | (50, 50, 32) | 864 |
| conv1_bn | (50, 50, 32) | 128 |
| conv1_relu | (50, 50, 32) | 0 |
| conv_dw | (50, 50, 32) | 288 |
| conv_pw | (50, 50, 64) | 2048 |
| conv_pad (ZeroPadding2D) | (51, 51, 64) | 0 |
| conv_dw | (25, 25, 64) | 576 |
| conv_pw | (25, 25, 128) | 8192 |
| conv_dw | (25, 25, 128) | 1152 |
| conv_pw | (25, 25, 128) | 16384 |
| conv_pad (ZeroPadding2D) | (26, 26, 128) | 0 |
| conv_dw | (12, 12, 128) | 1152 |
| conv_pw | (12, 12, 256) | 32768 |
| conv_dw | (12, 12, 256) | 2304 |
| conv_pw | (12, 12, 256) | 65536 |
| conv_pad (ZeroPadding2D) | (13, 13, 256) | 0 |
| conv_dw | (6, 6, 256) | 2304 |
| conv_pw | (6, 6, 512) | 131072 |
| $5\times$ conv_dw | (6, 6, 512) | 4608 |
| $5\times$ conv_pw | (6, 6, 512) | 262144 |
| conv_pad (ZeroPadding2D) | (7, 7, 512) | 0 |
| conv_dw | (3, 3, 512) | 4608 |
| conv_pw | (3, 3, 1024) | 524288 |
| conv_dw | (3, 3, 1024) | 9216 |
| conv_pw | (3, 3, 1024) | 1048576 |
| batch_normalization | (3, 3, 1024) | 4096 |
| global_average_pooling2d | (1024) | 0 |
| dropout | (1024) | 0 |
| FC | (1) | 1025 |

errors between model's age predictions and the ground truth on the validation set. The benefit of using MAE is that it can clearly reflect the performance of the model in age estimation as a regression task.

$$MAE(y_{predict}, y_{gt}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{10}$$

where $y_{predict}$ and $y_{gt}$ represent the age predictions and ground truth, $N$ is the number of samples, $y_i$ and $\hat{y}_i$ denote the $i_{th}$ prediction sample and groudtruth label, respectively.

### B. Implementation Details

Our model is implemented based on Tensorflow and Keras. In the experiments, some effective training stratigies are used in order to attain a better result. During the training process, we adopt strategies such as early stopping and learning rate decay to increase the speed of convergence as well as K-fold cross-validation to tune the hyperparameters that optimize the model generalization performance. Specifically, the criterion of early stopping is validation loss with patience equals 5, that is, the training ends when there is no improvement in performance for five consecutive epochs. The decay factor of learning rate is set to 0.15 with patience equals 3. We adopt Adam [2] as the optimizer of our model and the $\beta_1$ and $\beta_2$ are set to 0.9 and 0.999, respectively.

### C. Result Analysis

TABLE II
COMPARISON BETWEEN DIFFERENT METHODS INCLUDING SOTA.

| Methods | MAE | MAE (masked face) |
|---|---|---|
| Multi-output ConvNet [16] | 3.85 | 7.35 |
| DLPPA [15] | 4.02 | 6.98 |
| **Ours** | **3.79** | **6.54** |

As shown in II, our model outperforms other facical age estimation method and achieves state-of-the-art. The MAE of the age estimation on unmasked face dataset is 3.79 and achieves MAE of 6.54 on masked face dataset. The excellent performance of our model is due to sufficient data preprocessing, accurate keypoint detection and improved model structure based on MobileNet. In addition, modified training strategies also contributed. Some examples of the results are shown in Fig. 5.

## V. CONCLUSION

Facial information has become popular and valuable resource for diverse research. In this paper, an improved convolutional neural network architecture based on MobileNet is proposed and used to estimate age on both unmasked and synthetic masked face dataset. This paper also proposes an innovative mask generation method using face keypoint detection in order to address the problem of insufficient data. Our method is applied to WIKI Face dataset with a total of 158,208 images and reaches MAE of 3.79 and 6.54 on

(a)GT: 3 Predict: 3.04 (b)GT: 7 Predict: 7.22 (c)GT: 31 Predict: 32.09 (d)GT: 45 Predict: 47.74 (e)GT: 59 Predict: 57.19

Fig. 5. Examples of mask generation and age estimation results

unmasked and masked faces, respectively, achieving state-of-the-art. We believe our model can also be applied to the applications of future medical service or recommendation systems.

## REFERENCES


[1] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[2] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[3] Gu, Jiatao, et al. "Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis." arXiv preprint arXiv:2110.08985 (2021).

[4] Mamalet, Franck, and Christophe Garcia. "Simplifying convnets for fast learning." International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2012.

[5] L Ngan Mei, J Grother Patrick and K Hanaoka Kayee, Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms, 2020.

[6] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. Ieee, 2001, 1:I-I.

[7] S. Seneviratne, N. Kasthuriarachchi and S. Rasnayaka, "Multi-Dataset Benchmarks for Masked Identification using Contrastive Representation Learning," 2021 Digital Image Computing: Techniques and Applications (DICTA), 2021, pp. 01-08, doi: 10.1109/DICTA52665.2021.9647194.

[8] J. Lu, V. E. Liong and J. Zhou, "Cost-Sensitive Local Binary Feature Learning for Facial Age Estimation," in IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5356-5368, Dec. 2015, doi: 10.1109/TIP.2015.2481327.

[9] N. Yu, L. Qian, Y. Huang and Y. Wu, "Ensemble Learning for Facial Age Estimation Within Non-Ideal Facial Imagery," in IEEE Access, vol. 7, pp. 97938-97948, 2019, doi: 10.1109/ACCESS.2019.2928843.

[10] H. Dibeklioğlu, F. Alnajar, A. Ali Salah and T. Gevers, "Combining Facial Dynamics With Appearance for Age Estimation," in IEEE Transactions on Image Processing, vol. 24, no. 6, pp. 1928-1943, June 2015, doi: 10.1109/TIP.2015.2412377.

[11] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu and T. Gevers, "Expression-invariant age estimation using structured learning", IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 2, pp. 365-375, Feb. 2018.

[12] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo and S. Z. Li, "Efficient group-n encoding and decoding for facial age estimation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 11, pp. 2610-2623, Nov. 2018.

[13] Sun Y, Wang X, Tang X. Deep convolutional network cascade for facial point detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3476-3483.

[14] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).

[15] Seneviratne, Sachith, et al. "Does a Face Mask Protect my Privacy?: Deep Learning to Predict Protected Attributes from Masked Face Images." arXiv preprint arXiv:2112.07879 (2021).

[16] Savchenko, Andrey V. "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet." PeerJ Computer Science 5 (2019): e197.

[17] Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1867-1874.

[18] Rothe, Rasmus, Radu Timofte, and Luc Van Gool. "Dex: Deep expectation of apparent age from a single image." Proceedings of the IEEE international conference on computer vision workshops. 2015.