Improving Face-Based Age Estimation With Attention-Based Dynamic Patch Fusion

Haoyi Wang, Victor Sanchez, *Member, IEEE*, and Chang-Tsun Li, *Senior Member, IEEE*

***Abstract*— With the increasing popularity of convolutional neural networks (CNNs), recent works on face-based age estimation employ these networks as the backbone. However, state-of-the-art CNN-based methods treat each facial region equally, thus entirely ignoring the importance of some facial patches that may contain rich age-specific information. In this paper, we propose a face-based age estimation framework, called Attention-based Dynamic Patch Fusion (ADPF). In ADPF, two separate CNNs are implemented, namely the AttentionNet and the FusionNet. The AttentionNet dynamically locates and ranks age-specific patches by employing a novel Ranking-guided Multi-Head Hybrid Attention (RMHHA) mechanism. The FusionNet uses the discovered patches along with the facial image to predict the age of the subject. Since the proposed RMHHA mechanism ranks the discovered patches based on their importance, the length of the learning path of each patch in the FusionNet is proportional to the amount of information it carries (the longer, the more important). ADPF also introduces a novel diversity loss to guide the training of the AttentionNet and reduce the overlap among patches so that the diverse and important patches are discovered. Through extensive experiments, we show that our proposed framework outperforms state-of-the-art methods on several age estimation benchmark datasets.**

***Index Terms*— Age estimation, convolutional neural networks, attention mechanism, feature fusion.**

I. Introduction

FACE-BASED age estimation is an active and challenging research topic that keeps attracting attention from the research community [1]–[10]. The aim of the face-based age estimation task is to predict the real age (accumulated years after birth) of a subject from their facial images. This task has several applications in diverse scenarios like security control, video surveillance, and merchandise recommendation [11], [12].

Modern face-based age estimation methods typically consist of two components, a feature extractor and an estimator. The feature extractor is used to extract age-specific features from raw facial images and the estimator is used to predict the age based on the extracted features. Many recent works [6], [7], [13]–[20] focus on designing customized estimators while treating the facial image as an ordinary input, hence paying no attention to the relative importance of the extracted features. However, related studies [2], [5], [9] show that age-specific patches are useful when predicting the age of the subject from an image. In other words, customized feature extractors can be designed to exploit age-specific patches during training to boost the performance of face-based age estimation methods. As a consequence, many works now tackle the face-based age estimation problem by leveraging cropped age-specific patches as complementary inputs to their estimator [5], [9], [10], [21], [22]. The patches used in most of these works are those depicting dominant facial attributes like the eyes, nose, and mouth. However, early studies on face-based age estimation [23]–[29] show that the most informative patches for this problem are where wrinkles typically appear, like eye bags and laugh lines. To locate these age-specific patches, Han *et al.* [5] leverage the Bio-Inspired Features (BIF) proposed in [2]. Later, Wang *et al.* [9] design a customized CNN to fuse the features learned from the facial image and the BIF-based patches. Unfortunately, the computed BIF-based patches in these methods are fixed in every image, which prevents extracting features that are robust to the location and shape variations of age-specific regions.

In this paper, we propose a novel framework named Attention-based Dynamic Patch Fusion (ADPF) based on our preliminary work [9] to tackle the face-based age estimation problem. ADPF comprises a customized feature extractor that consists of an AttentionNet and a FusionNet. The AttentionNet dynamically discovers age-specific patches by employing a novel attention mechanism, while the FusionNet predicts the age of the subject by fusing features learned from the facial image and the discovered age-specific patches. To improve performance, the discovered patches are fed into the FusionNet sequentially in a descending order based on the amount of age-specific information they carry. To this end, we introduce the Ranking-guided Multi-Head Hybrid Attention (RMHHA) mechanism into the AttentionNet. RMHHA is inspired by the Multi-Head Self-Attention (MHSA) mechanism [30]. However, instead of using the multi-channel feature maps produced by MHSA, each attention head in RMHHA yields a compact single-channel attention map, which is used to crop the corresponding age-specific patch from the facial image. RMHHA assigns a learnable weights to the produced attention maps to rank their importance. Hence, RMHHA not only helps to dynamically learn age-specific patches, but it also

Manuscript received July 29, 2020; revised January 12, 2021, June 26, 2021, and November 26, 2021; accepted December 19, 2021. Date of publication January 6, 2022; date of current version January 12, 2022. This work was supported by the European Union (EU) Horizon 2020—Marie Skł odowska-Curie Actions through the project Computer Vision Enabled Multimedia Forensics and People Identification (IDENTITY) under Project 690907. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Heng Tao Shen. *(Corresponding author: Haoyi Wang.)*

Haoyi Wang and Victor Sanchez are with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: h.wang.16@warwick.ac.uk; v.f.sanchez-silva@warwick.ac.uk).

Chang-Tsun Li is with the School of Information Technology, Deakin University, Geelong, VIC 3216, Australia (e-mail: changtsun.li@deakin.edu.au).

Digital Object Identifier 10.1109/TIP.2021.3139226

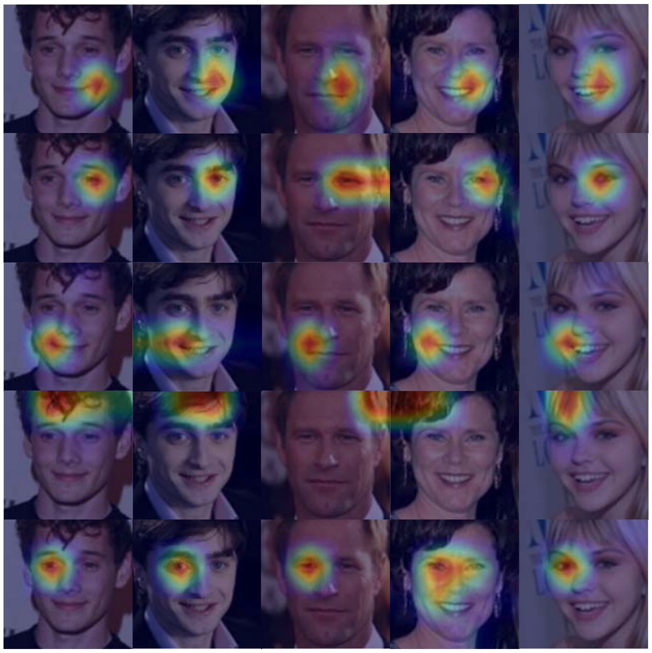


Fig. 1. Five most informative age-specific patches, indicated as heat maps, used in ADPF. Each row depicts an age-specific patch from one head in RMHHA for five different samples. For each sample, patches are presented in a descending order in terms of importance from top to bottom.

ensures the discovered patches are fed into the FusionNet in the desired order. The age-specific patches revealed by ADPF are exemplified in Fig. 1.

The contributions of this paper are as follows:

- We introduce ADPF, a framework that contains two networks, an AttentionNet and a FusionNet, to improve the face-based age estimation performance.
- Instead of using the BIF and AdaBoost algorithms in [9] to locate age-specific patches, ADPF uses the AttentionNet, which includes the novel RMHHA mechanism. RMHHA dynamically produces ranked single-channel attention maps, where each attention map highlights an age-specific patch.
- To reduce the overlap among patches, we propose a diversity loss to force RMHHA to reveal diverse age-specific regions.
- Through extensive experiments, we show that ADPF achieves state-of-the-art performance on several face-based age estimation benchmark datasets. We also show that, compared to our prior work [9], ADPF dramatically decreases training times.

The rest of this paper is organized as follows. In Section II, we review the related works on the face-based age estimation task and attention mechanisms. In Section III, we present the details of ADPF, including the RMHHA mechanism, the formulation of the diversity loss and the FusionNet. In Section IV, we explain the experimental settings and show the performance of ADPF and other state-of-the-art methods as evaluated on several age estimation benchmark datasets. Finally, we conclude our work in Section V.

II. Related Work

To lay the foundation of our work, in this section, related research on face-based age estimation is reviewed and discussed. We also review MHSA and other channel-wise attention mechanism, which are both related to the proposed RMHHA.

A. Face-Based Age Estimation

In the past few decades, many works have been conducted on face-based age estimation. One of the earliest works can be traced back to [31], in which the researchers classify faces into three age groups based on the cranio-facial development theory and wrinkle analysis. Later, [23] reveals that wrinkles play an important role in modeling aging faces and determining ages.

Before deep learning-based methods dominated the computer vision field, researchers used to develop face-based age estimation methods with hand-crafted features. For example, the Statistical Face Model [32] used in [33] is adopted to extract features and reveal the relationship between features and the corresponding age labels. Geng *et al.* [1], [34] propose the AGing pattErn Subspace (AGES) to learn aging pattern vectors in a representative subspace from training images. Unseen faces are then projected to this newly constructed subspace to predict their ages. Later, [35] reveals the ambiguity of mapping ages to age groups and proposes the Fuzzy Linear Discriminant Analysis (LDA) to build the classifier as an estimator. The authors define an Age Membership Function to encode the relevance between ages and age groups and integrate this function as a weighting factor into the conventional LDA. Guo and Mu [36] propose a kernel-based regression method to tackle the face-based age estimation problem. A worth-noting algorithm designed to extract hand-crafted features for face-based age estimation is BIF [2]. The BIF algorithm is based on the HMAX feature extraction method [37], which models the visual processing in the cortex. Specifically, it adopts the first two layers of HMAX, where the first layer convolves facial images with a set of Gabor filters [38] and the second layer performs maximum (max) pooling over the features extracted by the first layer. The authors improve this bio-inspired method by adding a normalization operation after max pooling. They find that using only the first two layers of HMAX achieves better results in the age estimation scenario than using the entire HMAX method. Recently, Han *et al.* [5] attach binary decision trees after the feature extraction process performed by the BIF algorithm to predict the age, gender and race simultaneously.

With the growing size of age-oriented datasets [39], [40], CNNs are now the foundation of feature extraction methods. One of the first works to use CNNs for the face-based age estimation problem is [41], in which a CNN with two convolutional layers is deployed. Han *et al.* [8] use a modified AlexNet [42] to construct a multi-task learning method for heterogeneous face attribute estimation including the age.

In general, CNN-based face-based age estimation methods can be classified into two categories. Works in the first category aim to design customized estimators after the feature extraction stage to better model the mapping between the features and the corresponding age label. For example, Niu *et al.* [7] treat face-based age estimation as an ordinal regression problem. In their work, a classifier with parallel fully-connected (FC) layers is constructed, where each FC layer produces a binary output that solves a binary classification sub-problem with respect to the corresponding age label. Chen *et al.* [16] also consider the ordinal relationship between different ages and propose the Ranking-CNN for face-based age estimation. Later, Pan *et al.* [18] propose the mean-variance loss that consists of a mean loss and a variance loss aiming to learn a concentrated age distribution with a mean value close to the ground-truth. Recently, Shen *et al.* [43] argue that the mapping between the facial features and the age label is inhomogeneous and introduce deep forests attached to CNNs to deal with such inhomogeneous mappings.

While the aforementioned CNN-based methods focus on learning a sophisticated mapping between the features and the corresponding age label, the works in the second category try to boost performance with customized feature extractors. Yi *et al.* [21] propose a multi-stream CNN to better leverage high-dimensional structured information in facial images. The authors crop multiple patches from facial images so that each stream learns from one patch. Then, the features extracted from different patches are fused before the output layer. Angeloni *et al.* [10] and Chen *et al.* [22] also follow the same multi-stream CNN strategy. The patches used in these works are mainly dominant facial attributes such as the eyes, the nose, and the mouth, and not age-specific patches, which are those where wrinkles typically appear like eye corners and laugh lines [23]–[29].

To locate informative age-specific patches, our prior work [9] uses the BIF and Adaboost algorithms. Specifically, the facial image is the primary input to the network as it carries more age-specific information than patches. The cropped patches are then subsequently fed into the CNN based on their importance, which is determined by the Adaboost algorithm. Due to the high-dimensional inputs to the BIF and Adaboost algorithms, the process of computing and ranking patches is extremely time consuming. In addition, the method proposed in our prior work consists of two separate stages, patch acquisition and CNN-training, which further increases the training complexity.

Since the facial image and cropped patches are processed by a different number of convolutional layers, i.e., the length of the learning path varies for different learning sources, the FusionNet in both our current and prior works involves fusing different levels of features. One work that also fuses different levels of features is [44]. However, the fused features in our work are from various inputs while the fused features in [44] are all from the input facial image.

B. Attention Mechanisms

1) Multi-Head Self-Attention: MHSA is first proposed in [30] and has been widely deployed as the backbone model for various Natural Language Processing (NLP) tasks [45]. MHSA can attend to multiple informative segments of the input with an attention head attending to one specific segment. Therefore, the number of segments MHSA can attend to is determined by the number of attention heads. MHSA has been recently used for imaging data. For example, Zhang *et al.* [46] uses MHSA for the image synthesis task. Specifically, the authors propose the self-attention generative adversarial network (SAGAN) by adding MHSA layers to both the generator and the discriminator of a generative adversarial network (GAN) [47]. With the help of MHSA layers, SAGAN can synthesize images with finer details than other state-of-the-art GAN models like [48]. Several recent works [49], [50] also use MHSA for image classification and object detection tasks.

2) Channel-Wise Attention: Ever since Zeiler *et al.* [51] visualized the feature maps learned by each channel in each layer of the AlexNet [42] trained on the ImageNet dataset [52], researchers have been exploiting channel-wise attention mechanism to guide the network to pay attention to those channels that learn representative feature maps. Hu *et al.* [53] integrate channel-wise attention into various CNN architectures [54]–[57] to boost their performance on image classification and object detection tasks. Similarly, Zhang *et al.* [58] and Chen *et al.* [59] employ channel-wise attention to generate high-resolution images and image captions, respectively. Different from the aforementioned works where channel-wise attention is used to highlight informative channels in the input, in the proposed RMHHA mechanism, we use the computed channel-wise attention weights to merge the multi-channel self-attention maps into a single-channel attention map that reveals a particular age-specific patch.

Some previously proposed works also use attention mechanisms to discover multiple informative regions in an image. Specifically, Ba *et al.* [60] first use a recurrent neural network (RNN) to locate multiple regions, where each iteration of the recurrence outputs one region. Later, Chen *et al.* [61], Rao *et al.* [62], and Shi *et al.* [63] leverage reinforcement learning into the RNN to help the localization. However, such RNN-based methods produce regions with significant overlap, which may result in redundant post-processing. Different from these methods, we implement a diversity loss to force the attention mechanism in our model to reveal diverse regions.

Another related work is the one proposed in [64], which uses a residual attention network to detect one or multiple objects in an image. However, their method cannot produce a consistent number of patches for each image, which is an important aspect of the attention mechanism in our model.

It is worth noting that the hybrid attention mechanism proposed in this paper is similar to the non-local block in [65]. The non-local block can be treated as the self-attention in transformers [30] with an additional short-cut connection. The difference between the non-local block and the hybrid attention mechanism is that instead of using a short-cut connection, we use channel-wise attention to produce a single-channel output.

III. Attention-Based Dynamic Patch Fusion

In this section, we explain in detail ADPF by first discussing the core of the AttentionNet, i.e., the proposed RMHHA

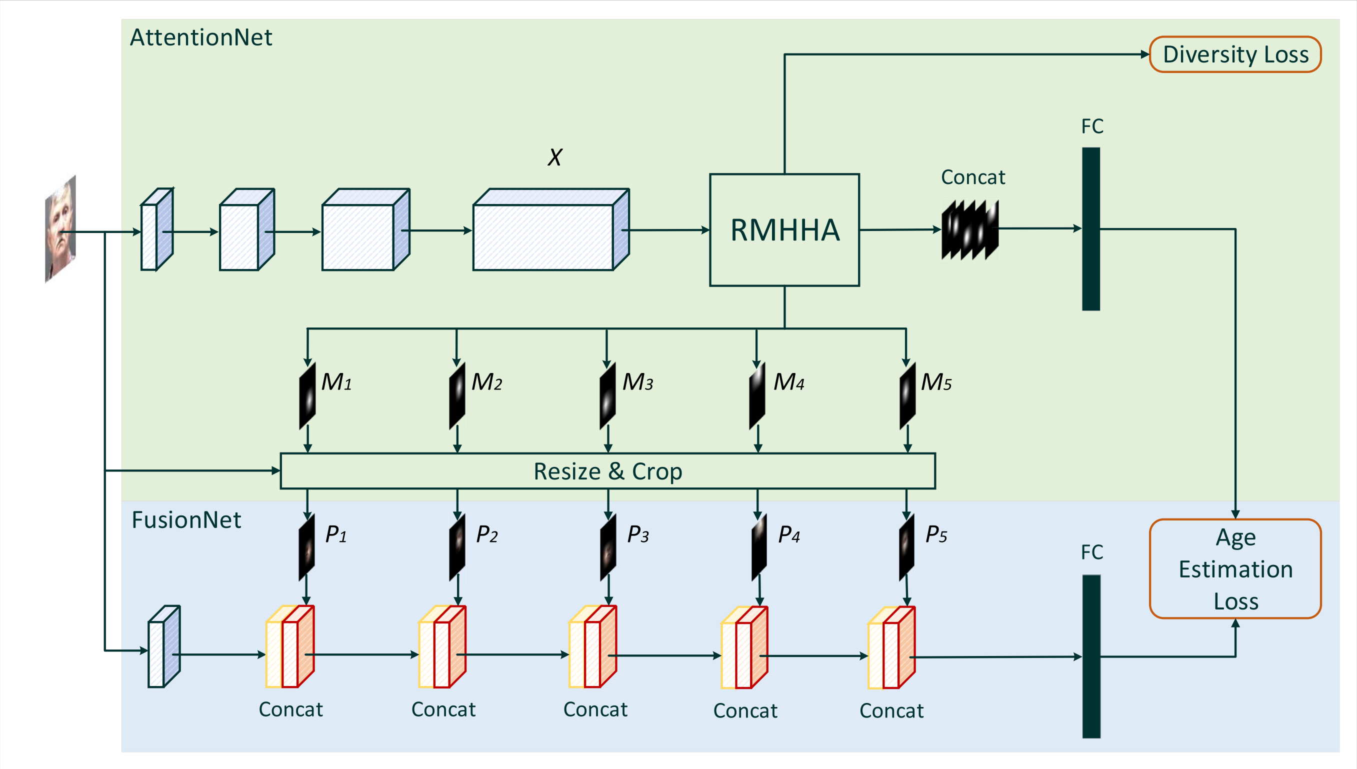


Fig. 2. Architecture of ADPF. It consists of two networks, the AttentionNet and the FusionNet. The AttentionNet is used to train the proposed RMHHA to learn and rank age-specific features. Once the features are learned and ranked, denoted as $\boldsymbol{M1}$ to $\boldsymbol{M5}$ in the figure, we resize them to crop the corresponding patches from the input facial image. The cropped patches are listed as **P1** to **P5** in a descending order based on the amount of age-specific information they carry. Blocks represents CNN layers, *Concat* indicates a concatenation operation, and *FC* indicates a fully-connected layer. In particular, yellows blocks are from the previous layer in the main stream and red ones are from one particular age-specific patch. In addition, **X** is the input tensor to the RMHHA mechanism with dimension of $32 \times 32 \times 500$.

mechanism. Then, we formulate the diversity loss followed by explaining the FusionNet used to fuse features from various learning sources. The architecture of ADPF is illustrated in Fig. 2.

A. Ranking-Guided Multi-Head Hybrid Attention

Since RMHHA is based on MHSA and the key component in MHSA is the self-attention mechanism, we first discuss the self-attention mechanism followed by the proposed hybrid attention mechanism. Then, we detail the complete RMHHA mechanism.

Let us consider an input tensor **X** that has a dimension of $h \times w \times c$, where h denotes the height, w denotes the width and the c denotes the number of channels. **X** is convolved into three separate tensors: **Q** with a shape of $h \times w \times c_{\mathbf{Q}}$, **K** with a shape of $h \times w \times c_{\mathbf{K}}$, and **V** with a shape of $h \times w \times c_{\mathbf{V}}$, where $c_{\mathbf{Q}}$, $c_{\mathbf{K}}$, and $c_{\mathbf{V}}$ indicate the number of channels in the corresponding tensor. The intuition behind self-attention is to compute a weighted summation of the values, **V**, where the weights are computed as the similarities between the query, **Q**, and the corresponding key, **K**. Therefore, in order to compute the similarity, **Q** and **K** normally have the same shape, i.e., $c_{\mathbf{Q}} = c_{\mathbf{K}}$. The output of a single self-attention mechanism is computed as:

$$\mathbf{SA} = Softmax(\frac{\mathbf{Q}' \cdot \mathbf{K}'^{T}}{\sqrt{c_{\mathbf{K}}}}) \cdot \mathbf{V}, \tag{1}$$

where $\mathbf{Q}'$ and $\mathbf{K}'$ are flattened tensors in order to perform the dot product.

After the scaling operation, i.e., dividing the similarity matrix $\mathbf{Q}' \cdot \mathbf{K}'^{T}$ by a factor of $\sqrt{c_{\mathbf{K}}}$ and applying the softmax function, we perform a dot product between the normalized similarity matrix and **V** to generate the self-attention maps **SA** with a dimension of $h \times w \times c_{\mathbf{K}}$.

Since we flatten two-dimensional feature maps into an one-dimensional vector in Eq. 1, the original structure of the feature maps is therefore distorted. To make it efficient when dealing with structured data like images and multi-dimensional features, we adopt the relative positional encoding in [66] and [49]. Specifically, the relative positional encoding is represented by the attention logit, which encodes how much an entry in $\mathbf{Q}'$ attends to an entry in $\mathbf{K}'$. The attention logit is computed as:

$$\boldsymbol{l}_{i,j} = \frac{\boldsymbol{q}_i^{T}}{\sqrt{c_{\mathbf{K}}}}(\boldsymbol{k}_j + \boldsymbol{r}^{w}_{j_x - i_x} + \boldsymbol{r}^{h}_{j_y - i_y}), \tag{2}$$

where $\boldsymbol{q}_i$ is the i-th row in $\mathbf{Q}'$ indicating the feature vector for pixel $i := (i_x, i_y)$ and $\boldsymbol{k}_j$ is the j-th row in $\mathbf{K}'$ indicating the feature vector for pixel $j := (j_x, j_y)$. $\boldsymbol{r}^{w}_{j_x - i_x}$ and $\boldsymbol{r}^{h}_{j_y - i_y}$ are learnable parameters encoding the positional information within the relative width $j_x - i_x$ and relative height $j_y - i_y$. With the relative positional encoding, the output of a single

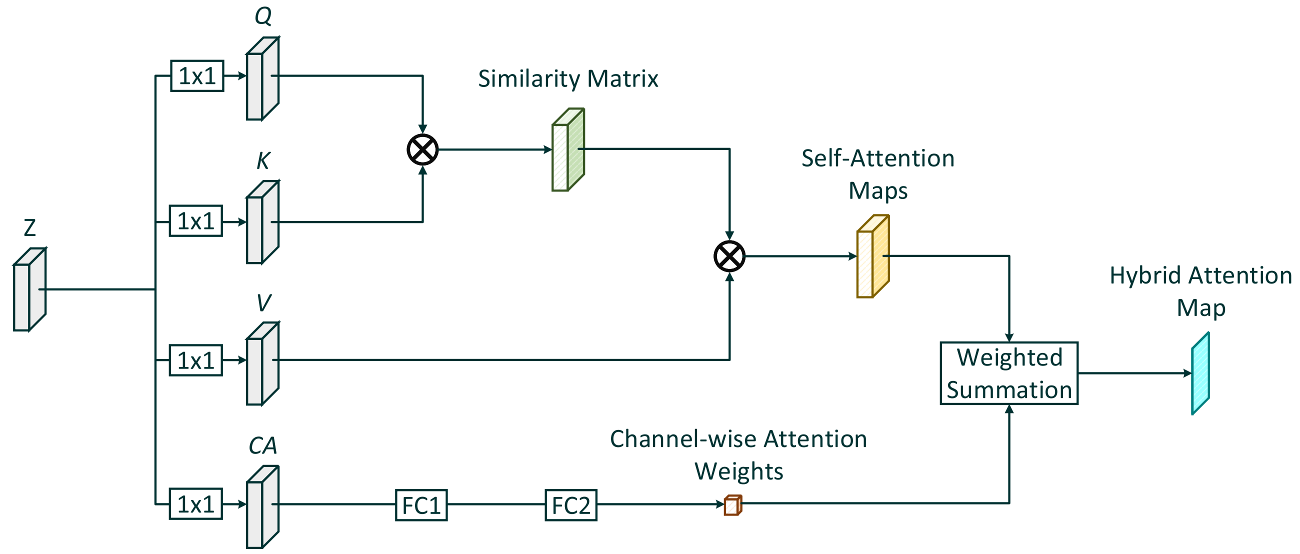


Fig. 3. Structure of the proposed hybrid attention mechanism. **Q**, **K**, and **V** are the *query*, *keys*, and *value*, respectively, for the self-attention mechanism, and CA is the input tensor to the channel-wise attention mechanism. The final hybrid attention map is computed as weighted summation, where the input tensor comprises the attention maps from the self-attention mechanism and the weights are computed from the channel-wise attention mechanism. 1 × 1 represents convolutional layers with kernel size of 1 and FC1 and FC2 indicate two fully-connected layers.

self-attention mechanism can be reformulated as:

$$\mathbf{SA} = Softmax(\frac{\mathbf{Q}' \cdot \mathbf{K}'^{T} + \boldsymbol{m}_h + \boldsymbol{m}_w}{\sqrt{c_{\mathbf{K}}}}) \cdot \mathbf{V}, \tag{3}$$

where $\boldsymbol{m}_h[i,j] = \boldsymbol{q}_i^T \boldsymbol{r}^h_{j_y - i_y}$ and $\boldsymbol{m}_w[i,j] = \boldsymbol{q}_i^T \boldsymbol{r}^w_{j_x - i_x}$ are matrices of relative positional logits.

The output of the self-attention mechanism in Eq. 3 has a dimension of $h \times w \times c_{\mathbf{V}}$. However, we want each attention head to produce a single-channel attention map to depict one particular age-specific patch. To this end, we use channel-wise attention alongside self-attention to form a hybrid attention mechanism. Channel-wise attention is used to compute weights for each channel and a weighted summation is performed along the channel axis of the self-attention maps to generate the final single-channel attention map, indicated as the hybrid attention map in Fig. 3.

As depicted in Fig. 3, in the proposed hybrid attention mechanism, we first use a 1×1 convolutional layer on the input tensor, **Z**, to ensure the number of channels before computing the channel-wise attention weights matches the number of channels in the self-attention maps, i.e., $c_{\mathbf{V}}$. The tensor after this 1×1 convolution is denoted as **CA**. We then aggregate each feature map in **CA** with a pooling operation to produce a feature vector, in which each entry represents the features for the corresponding channel. Different from [16], [53], in which average pooling is used, we use max pooling to emphasize the most important features with high activation values. Following the procedure in [53], we use a gating mechanism with two sequential FC layers to form a bottleneck. The first FC layer reduces the dimentionality, i.e., the number of channels, and the second FC layer increases the dimentionality of the previous layer to match the original shape. The output from the second FC layer is the set of channel-wise attention weights that we need, which are computed as:

$$\boldsymbol{w}_{CA} = \sigma(\boldsymbol{W}_{FC2}\delta(\boldsymbol{W}_{FC1}\delta(\mathbf{CA}))), \tag{4}$$

where δ indicates the non-linear ReLU function, σ refers to the Sigmoid function used to normalize the attention weights, and $\boldsymbol{W}_{FC1}$ and $\boldsymbol{W}_{FC2}$ are learnable parameters in the two FC layers.

After the self-attention maps and channel-wise attention weights are computed, we perform a weighted summation over these two tensors along the channel dimension to get the single-channel hybrid attention map. The hybrid attention map is then computed as:

$$\boldsymbol{HA} = \sum_{c}^{c_{\mathbf{V}}} \mathbf{SA}_c \cdot \boldsymbol{w}_{CA_c}, \tag{5}$$

where c is the channel index and **SA** is computed using Eq. 3.

To perform hybrid attention in a multi-head manner, each hybrid attention head takes a certain number of feature maps from the previous convolutional layer as the input. Specifically, assume there are c_p feature maps in the tensor produced by the previous layer. Then, we have $c_p = c_{head} \times n$, where n denotes the number of heads.

Different from MHSA [30], in which the attention maps from each head are concatenated right after the attention operation, we assign a learnable scale to each hybrid attention map to rank their importance when predicting ages, as shown in Fig. 4. RMHHA can then be formulated as:

$$\mathbf{RMHHA} = \{\boldsymbol{HA}_1 \cdot a_1, \boldsymbol{HA}_2 \cdot a_2, \ldots, \boldsymbol{HA}_n \cdot a_n\}, \tag{6}$$

where a_n indicates the learnable scale, which is updated by using the age estimation loss function presented in subsection III.D. $\boldsymbol{HA}_n \cdot a_n$ is equivalent to $\boldsymbol{HA}'_n$ in Fig. 4. All weighted hybrid attention maps used in ADPF are then concatenated before the final FC layer in the AttentionNet.

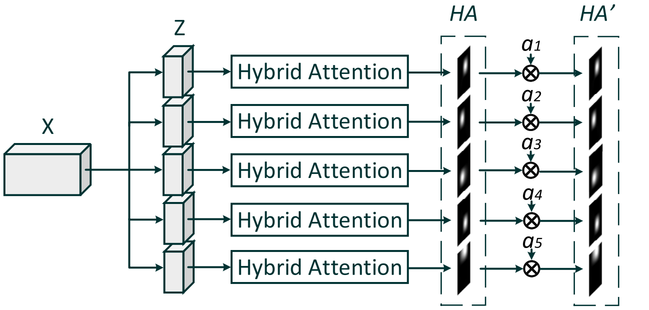


Fig. 4. Architecture of the proposed RMHHA, where five attention heads are implemented.

It is worth noting that multi-head attention methods always involve heavy matrix multiplications, which may be computationally expensive especially when the input matrices have a high dimentionality, which is common in CNNs. Therefore, differently from [49], [50], which stack dozens of MHSA models to compute the output, our work only uses one multi-head attention model to discover age-specific patches.

B. Diversity Loss

The number of patches that can be discovered is determined by the number of attention heads implemented in RMHHA. However, during implementation, we find that when using more than four heads, patches tend to overlap especially in informative regions. As demonstrated in Section IV, without further supervision, two attention maps may overlap in the nose region. This overlap of attended patches may led to redundant learning sources and leave other age-specific patches undiscovered. To alleviate this overlap issue, we propose a diversity loss to learn diverse and non-overlapping patches by minimizing the summation of products of corresponding entries in two hybrid attention maps, $\boldsymbol{HA}_{n_1}$ and $\boldsymbol{HA}_{n_2}$. The diversity loss is formulated as:

$$\mathcal{L}_{diversity} = \sum_{\substack{n_1,n_2 \\ n_1 \neq n_2}}^{n} \sum_{h'}^{h} \sum_{w'}^{w} \boldsymbol{HA}_{n_1}(h', w') \cdot \boldsymbol{HA}_{n_2}(h', w'), \quad (7)$$

where (h', w') denotes the location of the corresponding entry in a hybrid attention map.

C. FusionNet

The architecture of the FusionNet is illustrated in Fig. 2. Since the attention maps (e.g., $\mathbf{M_1}$ to $\mathbf{M_5}$) produced by the AttentionNet only contain the location of the patches, they do not contain sufficient information to aid in the learning process of the FusionNet. It is then necessary to retrieve the corresponding patches from the input image. To get these patches, i.e., $\mathbf{P_1}$ to $\mathbf{P_5}$, we first rank the learned hybrid attention maps based on their associated weights, i.e., a_1 to a_5. $\boldsymbol{M_1}$ has the highest weight indicating that the corresponding age-specific patch represents the most age-specific information. After the hybrid attention maps are ranked, they are resized into the same spatial size as the original facial image and used to crop the corresponding highlighted area by performing the contour detection based on the boundary in attention maps.

Instead of training separate shallow CNNs for each input and concatenating the information before the final FC layer, we merge the features in the convolution stage. In the FusionNet, the length of the path to learn from an input is directly proportional to the amount of information it carries. This approach also allows extracting and emphasizing common age-specific features among all inputs. For example, the skin feature, which has an ordinal relationship with the age, can be emphasized since all inputs are expected to share the same skin texture.

In the FusionNet, we preform concatenation operations on pairs of feature maps, one from the previous layer in the main stream (yellow blocks in Fig. 2), $\mathbf{I}$, and the other representing the features learned from one particular age-specific patch (red blocks in Fig. 2), $\mathbf{P}$. Therefore, the concatenation in the FusionNet is formulated as:

$$\mathbf{R} = Concate[\mathbf{I}, \mathbf{P}]. \quad (8)$$

This formulation is also commonly used in modern CNN architectures like the ResNet [56] and the DenseNet [67]. Therefore, a sub-stream in the FusionNet can be treated as a shortcut connection, which emphasizes the learning of the age-specific information shared by all inputs.

D. Age Estimation Loss

To estimate the age, we use a regression loss to learn the exact age and a divergence loss to learn the age distribution (i.e., the label distribution learning [68]). Specifically, after the features are processed by a Softmax function, we eliminate all the negative values in the output vector and normalize the remaining values so that they can form a probability distribution that sums up to 1:

$$o_p := \begin{cases} 0 & o_t \leq 0 \\ \dfrac{\sum_{p=1}^{q} \max(0, o_p)}{o_p} & o_t > 0, \end{cases} \quad (9)$$

where o_p is the p-th element in the output vector $\boldsymbol{o} \in \mathbb{R}^q$ and q is the total number of classes.

The final prediction is the summation of products of the probabilities by the corresponding age labels:

$$E = \sum_{p=1}^{q} o_p g_p, \quad (10)$$

where o_p denotes the normalized probability from Eq. 9 and g_p is the associated age label for class p.

We use the mean absolute error (MAE) to compute the error between the prediction and the corresponding ground truth label:

$$\mathcal{L}_{MAE} = \frac{1}{b} \sum_{b'}^{b} |E_{b'} - GT_{b'}|, \quad (11)$$

where b is the batch size and GT refers to the ground truth label.

Recent works [6], [18] also include a soft label technique to model the age distributions. Specifically, since there is no noticeable visual change of a face over a few years, adopting such technique can explicitly create more training samples for each label (e.g. age). Following these works, we use the KL-divergence to measure the difference between a Gaussian distribution derived from the label [18] and the learned distribution. The KL-divergence is formulated as:

$$\mathcal{L}_{KL} = \sum_{p=1}^{q} P(p) log \left(\frac{P(p)}{P'(p)} \right), \tag{12}$$

where P is the ground truth distribution and P' is the learned distribution. The complete age estimation loss is then defined as a summation of these two losses:

$$\mathcal{L}_{AE} = \mathcal{L}_{MAE} + \mathcal{L}_{KL}. \tag{13}$$

E. Training Strategy

Since the training of the FusionNet requires well-learned and stabilized patches, we first train the AttentionNet with RMHHA until convergence. The overall loss to train this network is the summation of two loss functions:

$$\mathcal{L}_{AttentionNet} = \mathcal{L}_{AE} + \lambda \mathcal{L}_{diversity}, \tag{14}$$

where λ controls the relative importance between two learning objectives.

When the AttentionNet converges, we freeze its parameters and start training the FusionNet. The loss function used to train the FusionNet is the loss formulated in Eq. 13.

IV. Experiments

A. Dataset

We conduct experiments on three commonly used face-based age estimation benchmark datasets, the MORPH II dataset [39], the FG-NET dataset [69], and the Cross-Age Celebrities Dataset (CACD) [40].

The MORPH II dataset contains more than 55,000 facial images from about 13,000 subjects with ages ranging from 16 to 77 and an average age of 33. The distribution of race labels in the MORPH II dataset is extremely unbalanced as more than 96% of subjects are annotated as *African* or *European* and individuals from *Asia* and other regions only occupy less than 4%. Each facial image in the MORPH II dataset is associated with identity, age, race and gender labels.

The FG-NET dataset has 1002 facial images belonging to 82 subjects. Each subject in this dataset has more than 10 facial images taken over a long time span. In addition, the facial images in this dataset contain pose, illumination and expression (PIE) variations.

The CACD contains more than 160,000 facial images from 2000 celebrities with ages ranging from 16 to 62. Similar to the images in the FG-NET dataset, facial images in the CACD contain PIE variations. The characteristics of these three datasets are presented in Table I.

TABLE I
Statistics of Three Benchmark Datasets

Dataset	#images	#subjects	age range
MORPH II	55,134	13,618	16-77
FG-NET	1,002	82	0-69
CACD	163,446	2000	16-62

B. Experimental Settings

1) Data Pre-Processing: We use the open-source computer vision library dlib [70] for image pre-processing. Firstly, 68 facial points are detected in each facial image to crop them based on the location of the eyes to a size of 128 × 128 pixels.

Further, data augmentation is used to increase the dataset size. Specifically, images are zero-padded first and then cropped to the original size. Finally, the cropped images are randomly flipped horizontally.

2) Dataset Partition: For the MORPH II dataset, three commonly used settings are adopted. In the first setting, i.e., *Setting I*, following prior works [7], [9], [14], [16], [18], [44], [78], we randomly split the whole dataset into two subsets, one with 80% of the data for training and the other with 20% for testing. In this setting, there is no identity overlap between the two subsets. To perform statistical analysis, we use 20 different partitions (with the same ratio but different distribution) and report mean values. In the second setting, i.e., the *Setting II*, to compensate for the imbalance of race distribution, we randomly split the dataset into three subsets, denoted as *S1*, *S2*, and *S3*, and ensure the ratio between Black and White labels is 1:1 and that between Male and Female labels is 1:3. In order to follow the same protocol as other works [19], [21], [22], [36], [80], the results under this setting are reported in three different ways: 1) training on *S1* and testing on *S2*+*S3*; 2) training on *S2* and testing on *S1*+*S3* and 3) the average value from the previous two scenarios. Finally, in the third setting, i.e., the *Setting III*, we select 5,492 facial images of White people to reduce the variance caused by imbalanced race distribution [41], [75], [77], [81]. Then, these 5,492 facial images are randomly split into two subsets, 80% of the them are used for training and the remaining 20% for testing. To further reduce the data distribution variance, in this setting, we use 5-fold cross validation to produce the final results.

For the FG-NET dataset, we use the leave-one-person-out (LOPO) strategy [14], [34], [43], [44], [72], [82]. In each fold, we use facial images of one subject for testing and the remaining images for training. Since there are 82 subjects, this process consists of 82 folds and the reported results are the average values.

For the CACD, following the setup in [75], [80], and [43], the whole dataset is divided into three subsets, denoted as the training set, validation set, and testing set. The training set has facial images from 1,800 subjects, the validation set has facial images from 120 subjects, and the testing set has facial images from 80 subjects. The reported results are computed by training either on the training set or the validation set and evaluating on the testing set.

3) Evaluation Metrics: Results are reported based on two metrics, Mean Absolute Error (MAE) and Cumulative Score (CS). The MAE measures the average absolute difference between the predicted age and the ground truth:

$$MAE = \frac{\sum_{z'}^{z} e_{z'}}{z}, \tag{15}$$

where $e_{z'}$ is the absolute error between the predicted age $\hat{u_{z'}}$ and the input label $u_{z'}$ for the z'-th sample, and z is the total number of testing samples. The CS measures the percentage of images that are correctly classified in a certain age range as:

$$CS(\upsilon) = -\frac{z_\upsilon}{z} \times 100\%, \tag{16}$$

where Z_υ is the number of images whose predicted age $\hat{u_z}$ is in the range of $[u_z - \upsilon, u_z + \upsilon]$ and υ is the age margin.

4) Implementation Details: ADPF is implemented based on the open-source deep learning framework Pytorch [87] and trained with the SGD algorithm with a batch size of 32. We first train the AttentionNet for 200 epochs and then the FusionNet for another 200 epochs with the parameters of the AttentionNet fixed. The initial learning rate for both networks is set to 0.1 and drops by a factor of 0.1 after every 50 epochs. When training the AttentionNet, we empirically set λ in Eq. 14 to 0.01. Following our prior work, we use 5 patches when comparing with other state-of-the-art methods. All experiments are run on a single NVIDIA GTX 2080Ti GPU. To have a fair comparison against our prior work, we replace the age regression model used by our prior work with the age estimation loss in Eq. 13.

C. Evaluations on the MORPH II Dataset

The MAE values for the three aforementioned settings of the MORPH II dataset are tabulated in Table II-IV, respectively. In Table III, the headings indicate the subsets used to compute the results. For example, *S1/S2+S3* indicates the model is trained on the *S1* subset and evaluated on the *S2* and *S3* subsets, and the *Average* column tabulates the mean value of the two columns on the left. The CS curves for the three settings are presented in Fig. 5-7, respectively. Note that not all methods report the results under this metric. As can be seen from these tables and figures, ADPF outperforms all state-of-the-art methods that focus on improving the feature extractor like the DAG family (DAG-GoolgeNet and DAG-VGG16) [78], MSFCL family (MSFCL, MSFCL-LR, and MSFCL-KL) [44], and our prior work [9]. Also note that ADPF achieves comparable results to other methods that use customized estimators. For all three settings, the superior performance demonstrate that ADPF can predict ages accurately regardless of the imbalanced data distribution caused by other information like race. We also include comparisons of the number of parameters in Table II, IV, and V to provide more information into the performance of the evaluated methods. As tabulated, our method is among the ones with the smallest number of parameters, using only 12% of the total number of parameters used by the BridgeNet.

TABLE II

MAE Values for Several State-of-the-Art Face-Based Age Estimation Methods on the MORPH II Dataset Under Setting I. Params Indicates the Number of Parameters

Method	Params	MAE
OHRank [71]	-	6.07
IIS-LLD [72]	-	5.67
CPNN [72]	-	4.87
OR-SVM [73]	-	4.21
BFGS-LDL [74]	-	3.94
OR-CNN [7]	7M	3.27
DEX [75]	138M	3.25
SMMR [76]	-	3.24
ARN [77]	139M	3.00
Ranking-CNN [16]	26M	2.96
MSFCL [44]	15M	2.90
DAG-GoogleNet [78]	24M	2.87
DAG-VGG16 [78]	131M	2.81
Mean-Variance Loss [18]	20M	2.80
MSFCL-LR [44]	15M	2.79
Hu *et al.* [6]	24M	2.78
BIF + FusionNet [9]	5M	2.76
MSFCL-KL [44]	15M	2.73
VDAL [79]	19M	2.57
ADPF (ours)	14M	**2.54**

TABLE III

MAE Values for Several State-of-the-Art Face-Based Age Estimation Methods on the MORPH II Dataset Under Setting II

Method	MAE		
	S1/S2+S3	S2/S1+S3	Average
KPLS [36]	4.21	4.15	4.18
MS-CNN [21]	3.63	3.63	3.63
MRNPE (AlexNet) [80]	2.98	2.73	2.86
MRNPE (VGG16) [80]	2.85	2.60	2.73
ARAN [22]	2.77	**2.48**	2.63
BridgeNet [19]	2.74	2.51	2.63
ADPF (ours)	**2.63**	2.50	**2.56**

TABLE IV

MAE Values for Several State-of-the-Art Face-Based Age Estimation Methods on the MORPH II Dataset Under Setting III. Params Indicates the Number of Parameters

Method	Params	MAE
AGES [1]	-	8.83
MTWGP [83]	-	6.28
CA-SVR [84]	-	5.88
DLA [41]	6M	4.77
Rothe *et al.* [85]	20M	3.45
DLDLF [43]	14M	2.94
DRF [43]	14M	2.80
deep-JREAE [86]	138M	2.77
BridgeNet [19]	120M	**2.38**
ADPF (ours)	14M	2.71

D. Evaluations on the FG-NET Dataset

The MAE values and the CS curve are tabulated in Table V and depicted in Fig. 8, respectively, for the FG-NET dataset.

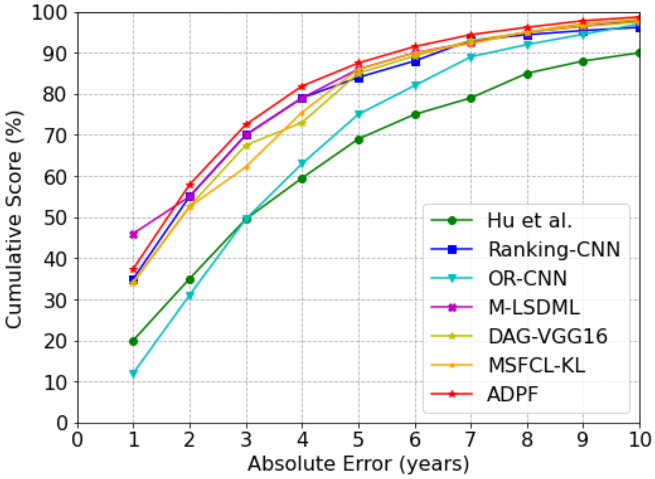


Fig. 5. CS curves for several state-of-the-art face-based age estimation methods on the MORPH II Dataset under Setting I.

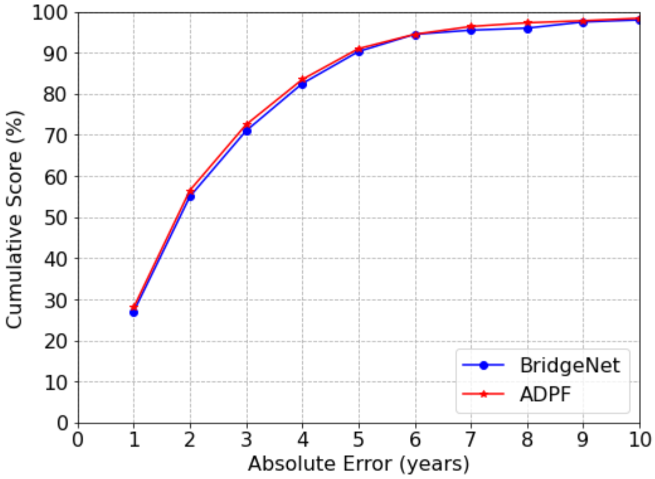


Fig. 7. CS curves for several state-of-the-art face-based age estimation methods on the MORPH II Dataset under Setting III.

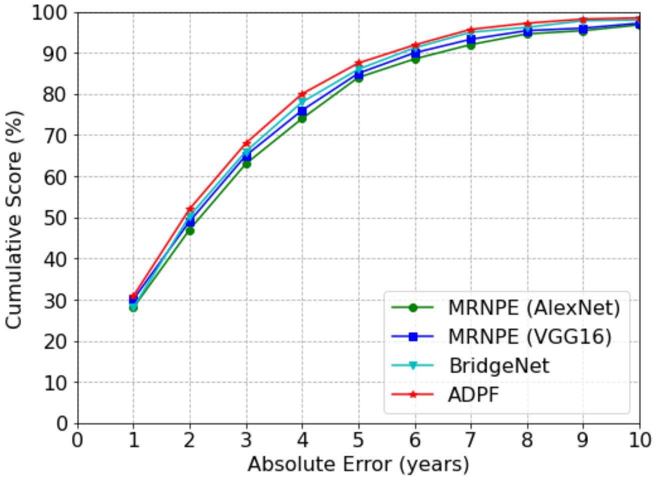


Fig. 6. CS curves for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting II.

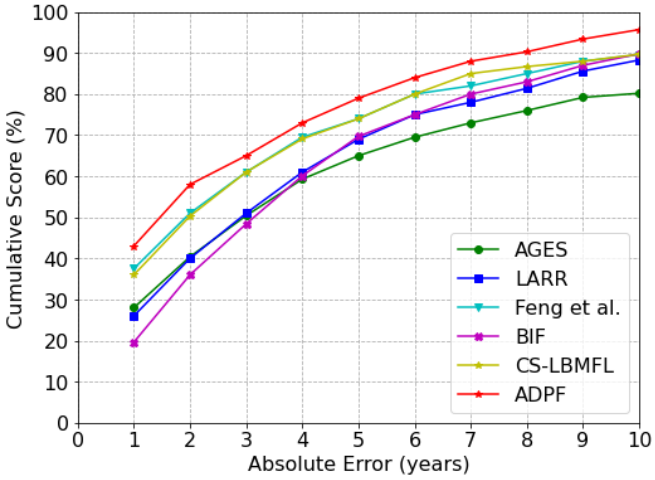


Fig. 8. CS curves for several state-of-the-art face-based age estimation methods on the FG-NET Dataset.

Again, not all methods report the results under the CS metric for the FG-NET dataset. It can be seen from Table V that ADPF achieves an MAE value under 3.00, which shows that it can perform well even with small datasets.

E. Evaluations on the CACD

Evaluation results for the CACD under the MAE metric are tabulated in Table VI. ADPF achieves the best performance when trained on the validation dataset but only achieves the third best performance when trained on the training set. This may due to the age labels in the training set not being accurate. Since the input to the FusionNet of ADPF is sixfold, i.e., it includes one facial image and five patches, compared to other single-input networks, inaccurate labels may confuse the model due to mis-information.

F. Ablation Study

We conduct ablation experiments to demonstrate the effectiveness of each component of ADPF. Specifically, we aim to demonstrate that: 1) the hybrid attention mechanism is more effective than the self-attention mechanism when discovering age-specific patches; 2) the ranking operation in RMHHA is beneficial for feature learning in the FusionNet; 3) the effectiveness of the diversity loss; and 4) the importance of combining the FusionNet and the AttentionNet in a single framework. To this end, we design several baseline models as follows:

- *ADPF w/SA*: ADPF with the self-attention mechanism instead of the hybrid attention mechanism in the AttentionNet. The single channel feature maps are then generated by performing summation along the channel axis of the self-attention maps.
- *ADPF w/o ranking*: ADPF without the ranking operation for age-specific patches.
- *ADPF w/o diversity*: ADPF without the diversity loss.
- *AttentionNet*: ADPF with no FusionNet.

The evaluation results on the MORPH II dataset, Setting I, for the aforementioned baseline models and ADPF are tabulated in Table VII. Example attention maps computed

TABLE V

MAE VALUES FOR SEVERAL STATE-OF-THE-ART FACE-BASED AGE ESTIMATION METHODS ON THE FG-NET DATASET. PARAMS INDICATES THE NUMBER OF PARAMETERS

Method	Params	MAE
AGES [1]	-	6.77
IIS-LLD [72]	-	5.77
LARR [81]	-	4.87
Feng *et al.* [13]	-	5.05
BIF [2]	-	4.77
CPNN [72]	-	4.76
DEX [75]	120M	4.63
CS-LBFL [82]	-	4.43
CS-LBMFL [82]	-	4.36
Mean-Variance Loss [18]	20M	4.10
GA-DFL [15]	138M	3.93
LSDML [14]	44M	3.92
ARAN [22]	414M	3.79
M-LSDML [14]	44M	3.74
DLDLF [43]	14M	3.71
DRF [43]	14M	3.47
DAG-VGG16 [78]	24M	3.08
DAG-GoogleNet [78]	131M	3.05
BridgeNet [19]	120M	**2.56**
ADPF (ours)	14M	2.86

TABLE VI

MAE VALUES FOR SEVERAL STATE-OF-THE-ART FACE-BASED AGE ESTIMATION METHODS ON THE CACD

Method	MAE	
	train	val
DEX [75]	4.79	6.52
DLDLF [43]	4.68	6.16
DRF [43]	**4.61**	5.63
ADPF (ours)	4.72	**5.39**

TABLE VII

MAE VALUES FOR SEVERAL BASELINE MODELS AND THE COMPLETE ADPF FRAMEWORK ON THE MORPH II DATASET UNDER SETTING I

Method	MAE
ADPF w/SA	2.90
ADPF w/o ranking	2.74
ADPF w/o diversity	2.65
AttentionNet	3.31
ADPF	**2.54**

by the *ADPF w/SA* baseline model are shown in Fig. 9. As shown in this figure, although *ADPF w/SA* can reveal key regions for age estimation, it may also reveal non-important regions, including sections of the background, which may be treated as noise during the feature learning process and eventually hinder the performance. In *ADPF w/o ranking*, we feed the patches into the FusionNet based on their original order in the input tensor along the channel axis as produced by RMHHA. This feeding strategy cannot guarantee that the learning path for the most informative patch is long enough to extract meaningful features.

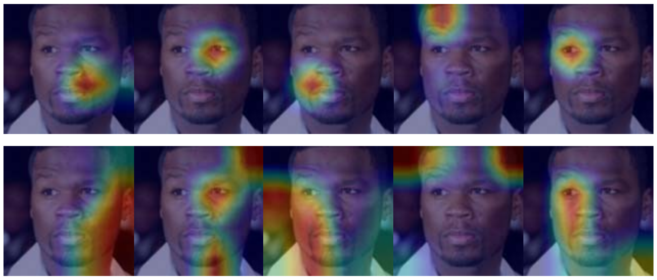


Fig. 9. Attention maps computed by (upper row) the ADPF framework and (bottom row) the *ADPF w/SA* baseline model.

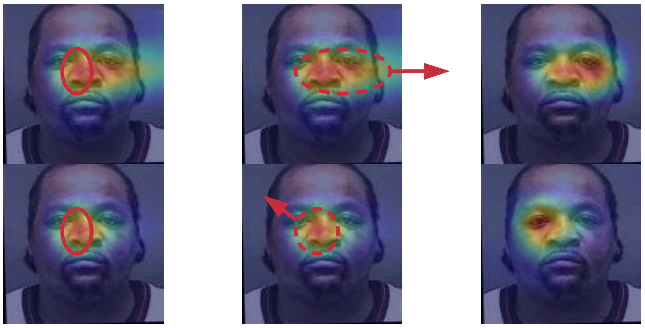


Fig. 10. **Left**: Two attention maps overlap in the annotated area with out the supervision from the diversity loss. **Middle**: By minimizing the diversity loss, the two attention maps are forced to move in opposite directions. **Right**: attention maps generated by using the diversity loss.

It is worth noting that we do not compare the proposed hybrid attention mechanisms with other attention mechanisms like non-local blocks due to different output formats, i.e., outputs from non-local blocks have multiple channels while outputs from hybrid attention mechanisms only have one channel. This difference in number of channels makes hybrid attention mechanisms and other attention mechanisms not interchangeable.

It would be interesting to compare the proposed hybrid attention mechanisms with other attention mechanisms like non-local blocks. However, since the output from such attention mechanisms has multiple channels.

To demonstrate the effectiveness of the proposed diversity loss, we visualize the attention maps learned on the MORPH II dataset, Setting I, by ADPF and the baseline model *ADPF w/o diversity*. As shown in Fig. 10, in the *ADPF w/o diversity* baseline model, the two attention maps overlap in the highlighted nose region, which leads to redundant input information to the network. With the aid of the diversity loss, these key regions detected by these two attention maps are forced to move in opposite directions resulting in two attention maps with negligible overlap.

MAE values tabulated in Table VII confirm the importance of combining the AttentionNet and the FusionNet in a single framework instead of using the AttentionNet exclusively. As we can see from this table, the performance of the *AttentionNet* baseline model significantly drops compared to that of ADPF. This is mainly due to the limited number of feature maps available to the FC layer in the AttentionNet. With such a limited number of feature maps, the estimator cannot

TABLE VIII
TRAINING TIME OF THE FUSIONNET IN OUR PRIOR WORK AND ADPF

Method	Hours	MAE
BIF + FusionNet [9]	70	2.76
ADPF	**25**	**2.54**

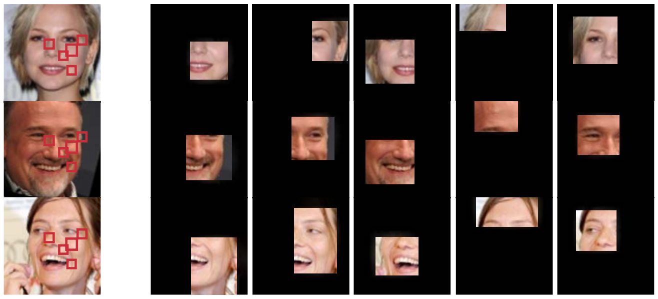


Fig. 11. Sample age-specific patches computed by our prior work [9] and the ADPF framework. The left column depicts the original facial images with patches computed by [9] highlighted in red. The five patches computed by the ADPF framework are depicted in the last five columns. Within these columns, the patches are depicted from left to right in descending order in terms of their importance.

get enough information from the feature extractor. However, implementing the AttentionNet in this way is essential to learn and rank multiple single-channel attention maps, which shows the importance of combining the AttentionNet and the FusionNet in a single framework.

G. Discussions

1) Training Efficiency: We compare the training time required by our prior work [9] and the ADPF on the MORPH II dataset with *Setting I*. The training times are tabulated in Table VIII. Note that it takes about 70 hours to train the whole method in [9] out of which 60 hours are required to compute and rank BIF-based patches and 10 hours to train the CNN. Thanks to the proposed RMHHA mechanism, ADPF only takes about one third of this time to converge with significantly boosted performance (see MAE values). In addition, the process of acquiring patches and training the CNN can only be done separately in [9]. On the contrary, in ADPF, the training of the FusionNet can be done directly after the AttentionNet converges, which further makes the training process more time-efficient.

2) Robustness of Age-Specific Patches: We visually compare the patches computed by the BIF and Adaboost algorithms used in [9] and those computed by RMHHA. This comparison is conducted on the CACD dataset as the facial images in this dataset contain PIE variations. Fig. 11 depicts sample patches, where the most informative patches computed by [9] are marked with red boxes. It is clear that the location and shape of each patch computed by [9] are identical for all the images. On the contrary, the location and shape of the patches computed by the RMHHA vary from image to image. For example, in the bottom row, the patch capturing the right laughline is larger than that of the other two images, which allows capturing the complete skin texture of this key region.

TABLE IX
PERFORMANCE OF ADPF WITH DIFFERENT NUMBER OF ATTENTION HEADS ON THE MORPH II DATASET UNDER SETTING I

# Heads	3	4	5	6	7	8
MAE	2.77	2.62	**2.54**	**2.54**	2.55	2.61

3) Number of Heads: The performance of ADPF with different number of attention heads is tabulated in Table IX. We can see that the best performance can be achieved when 5 or 6 attention heads are implemented. This may due to the fact that with less heads, some age-specific patches may remain undiscovered. Moreover, since most of the facial regions are already revealed when 5 attention heads are used, adding more heads only forces the framework to attend to irrelevant regions like the background, which as discussed previously, can be treated as noise and degrade the performance. Since 6 heads requires more time to train with no significant performance gains, 5 is an appropriate number to be used by ADPF.

4) Limitations: Although the training efficiency has dramatically increased compared to our prior work [9], the pipeline requires more time during inference due to the involvement of the hybrid attention mechanism. Such attention mechanisms can have a quadratic complexity with respect to the input dimension due to the similarity comparison operations [88].

V. CONCLUSION

In this paper, we proposed the ADPF framework to improve the performance of the face-based age estimation task. Our framework merges an AttentionNet and a FusionNet. The AttentionNet includes a novel hybrid attention mechanism, namely RMHHA, which allows learning multiple single-channel attention maps to reveal age-specific patches. After ranking them, these patches are used by the FusionNet, along with the facial image to compute the final age prediction. Based on evaluations on several benchmark datasets, ADPF significantly improves prediction accuracy compared to several state-of-the-art methods. ADPF also outperforms our previous work, both in terms of accuracy and training times. Since this work focuses on building customized feature extractors, in the future, we will investigate the design of customized estimators to further boost performance by, for example, considering the ordinal information among ages and further minimizing the distance between label distributions and feature distributions.

REFERENCES

[1] X. Geng, Z.-H. Zhou, and K. Smith-Miles, “Automatic age estimation based on facial aging patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[2] G. Guo, G. Mu, Y. Fu, and T. S. Huang, “Human age estimation using bio-inspired features,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 112–119.

[3] Y.-L. Chen and C.-T. Hsu, “Subspace learning for facial age estimation via pairwise age ranking,” *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 12, pp. 2164–2176, Dec. 2013.

[4] C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu, and H. Lu, “Human age estimation based on locality and ordinal information,” *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2522–2534, Nov. 2014.

[5] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[6] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3087–3097, Jul. 2017.

[7] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

[8] H. Han, A. K. Jain, X. Chen, F. Wang, and S. Shan, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2597–2609, Nov. 2018.

[9] H. Wang, X. Wei, V. Sanchez, and C.-T. Li, "Fusion network for face-based age estimation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2675–2679.

[10] M. Angeloni, R. de Freitas Pereira, and H. Pedrini, "Age estimation from facial parts using compact multi-stream convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–7.

[11] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Sep. 2010.

[12] H. Wang, V. Sanchez, W. Ouyang, and C.-T. Li, "Using age information as a soft biometric trait for face image analysis," in *Deep Biometrics*. Cham, Switzerland: Springer, 2020, pp. 1–20.

[13] S. Feng, C. Lang, J. Feng, T. Wang, and J. Luo, "Human facial age estimation by cost-sensitive label ranking and trace norm regularization," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 136–148, Jan. 2017.

[14] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 292–305, Feb. 2018.

[15] H. Liu, J. Lu, J. Feng, and J. Zhou, "Group-aware deep feature learning for facial age estimation," *Pattern Recognit.*, vol. 66, pp. 82–94, Jun. 2017.

[16] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5183–5192.

[17] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep regression forests for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2304–2313.

[18] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5285–5294.

[19] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "BridgeNet: A continuity-aware probabilistic network for age estimation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1145–1154.

[20] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," 2019, *arXiv:1901.07884*.

[21] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 144–158.

[22] Y. Chen, S. He, Z. Tan, C. Han, G. Han, and J. Qin, "Age estimation via attribute-region association," *Neurocomputing*, vol. 367, pp. 346–356, Nov. 2019.

[23] Y. Wu, N. M. Thalmann, and D. Thalmann, "A dynamic wrinkle model in facial animation and skin ageing," *J. Vis. Comput. Animation*, vol. 6, no. 4, pp. 195–205, Oct. 1995.

[24] L. Boissieux, G. Kiss, N. M. Thalmann, and P. Kalra, "Simulation of skin aging and wrinkles with cosmetics insight," in *Proc. Comput. Animation Simulation*. Vienna, Austria: Springer, 2000, pp. 15–27.

[25] S. Akazaki *et al.*, "Age-related changes in skin wrinkles assessed by a novel three-dimensional morphometric analysis," *Brit. J. Dermatol.*, vol. 147, no. 4, pp. 689–695, Oct. 2002.

[26] S. Mukaida and H. Ando, "Extraction and manipulation of wrinkles and spots for facial image synthesis," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Jan. 2004, pp. 749–754.

[27] J. M. Lagarde, C. Rouvrais, and D. Black, "Topography and anisotropy of the skin surface with ageing," *Skin Res. Technol.*, vol. 11, no. 2, pp. 110–119, May 2005.

[28] A. Golovinskiy, W. Matusik, H. Pfister, S. Rusinkiewicz, and T. Funkhouser, "A statistical model for synthesis of detailed facial geometry," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1025–1034, 2006.

[29] M. Lai, I. Oruç, and J. J. S. Barton, "The role of skin texture and facial shape in representations of age and identity," *Cortex*, vol. 49, no. 1, pp. 252–265, Jan. 2013.

[30] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[31] Y. H. Kwon and D. V. Lobo, "Age classification from facial images," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 1994, pp. 762–767.

[32] G. J. Edwards, A. Lanitis, C. J. Taylor, and T. F. Cootes, "Statistical models of face images—Improving specificity," *Image Vis. Comput.*, vol. 16, no. 3, pp. 203–211, Mar. 1998.

[33] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[34] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai, "Learning from facial aging patterns for automatic age estimation," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 307–316.

[35] F. Gao and H. Ai, "Face age classification on consumer images with Gabor feature and fuzzy LDA method," in *Proc. Int. Conf. Biometrics*. Berlin, Germany: Springer, 2009, pp. 132–141.

[36] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. CVPR*, 2011, pp. 657–664.

[37] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[38] D. Gabor, "Theory of communication. Part 1: The analysis of information," *J. Inst. Elect. Eng.*, vol. 93, no. 26, pp. 429–441, Jul. 1946.

[39] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 341–345.

[40] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 768–783.

[41] X. Wang, R. Guo, and C Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 534–541.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[43] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille, "Deep differentiable random forests for age estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 404–419, Feb. 2019.

[44] M. Xia *et al.*, "Multi-stage feature constraints learning for age estimation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2417–2428, 2020.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[46] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2014, pp. 2672–2680.

[48] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–35.

[49] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jan. 2019, pp. 3286–3295.

[50] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 68–80.

[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2018, pp. 7132–7141.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[55] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2016, pp. 770–778.

[57] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[58] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[59] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2017, pp. 5659–5667.

[60] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *Proc. ICLR*, 2015, pp. 1–10.

[61] T. Chen, Z. Wang, G. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 6730–6737.

[62] Y. Rao, J. Lu, and J. Zhou, "Attention-aware deep reinforcement learning for video face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jan. 2017, pp. 3931–3940.

[63] Y. Shi, G. Li, Q. Cao, K. Wang, and L. Lin, "Face hallucination by attentive sequence optimization with reinforcement learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2809–2824, Nov. 2020.

[64] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[65] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[66] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, 2018, pp. 464–468.

[67] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[68] W. Shen, K. Zhao, Y. Guo, and A. L. Yuille, "Label distribution learning forests," in *Proc. NIPS*, 2017, pp. 834–843.

[69] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001.

[70] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[71] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR*, 2011, pp. 585–592.

[72] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[73] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "A ranking approach for human ages estimation based on face images," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3396–3399.

[74] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[75] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, pp. 144–157, Apr. 2018.

[76] D. Huang, L. Han, and F. De la Torre, "Soft-margin mixture of regressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6532–6540.

[77] E. Agustsson, R. Timofte, and L. Van Gool, "Anchored regression networks applied to age estimation and super resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1643–1652.

[78] S. Taheri and O. Toygar, "On the use of DAG-CNN architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300–310, Jan. 2019.

[79] H. Liu, P. Sun, J. Zhang, S. Wu, Z. Yu, and X. Sun, "Similarity-aware and variational deep adversarial learning for robust facial age estimation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1808–1822, Jul. 2020.

[80] Y. Chen, Z. Tan, A. P. Leung, J. Wang, and J. Zhang, "Multi-region ensemble convolutional neural networks for high accuracy age estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12.

[81] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[82] J. Lu, V. E. Liong, and J. Zhou, "Cost-sensitive local binary feature learning for facial age estimation," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.

[83] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2622–2629.

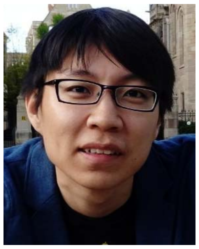
[84] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2467–2474.

[85] R. Rothe, R. Timofte, and L. Van Gool, "Some like it hot—Visual guidance for preference prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5553–5561.

[86] Q. Tian *et al.*, "Facial age estimation with bilateral relationships exploitation," *Neurocomputing*, vol. 444, pp. 158–169, Jul. 2021.

[87] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.

[88] F. Babiloni, I. Marras, F. Kokkinos, J. Deng, G. Chrysos, and S. Zafeiriou, "Poly-NL: Linear complexity non-local layers with polynomials," 2021, *arXiv:2107.02859*.



Haoyi Wang received the B.Sc. degree in engineering from the North China University of Technology, China, and the University of Central Lancashire, U.K., in 2013, and the M.Sc. degree from The University of Manchester, U.K., in 2014. He is currently pursuing the Ph.D. degree in computer science with the University of Warwick, U.K. His research interests include deep supervised and unsupervised learning, generative models, and computer vision.

Victor Sanchez (Member, IEEE) received the M.Sc. degree from the University of Alberta, Canada, in 2003, and the Ph.D. degree from The University of British Columbia, Canada, in 2010. From 2011 to 2012, he was with the Video and Image Processing Laboratory, University of California at Berkeley, as a Postdoctoral Researcher. In 2012, he was a Visiting Lecturer with the Group on Interactive Coding of Images, Universitat Autònoma de Barcelona. From 2018 to 2019, he was a Visiting Scholar with the School of Electrical and Information Engineering, The University of Sydney, Australia. He is currently an Associate Professor with the Department of Computer Science, University of Warwick, U.K. His main research interests are in the area of signal and information processing with applications to multimedia analysis, image and video coding, security, and communications. He has authored several technical papers in these areas and coauthored a book (Springer, 2012). His research has been funded by the Consejo Nacional de Ciencia y Tecnologia, Mexico; the Natural Sciences and Engineering Research Council of Canada; the Canadian Institutes of Health Research; the FP7 and H2020 programs of the European Union; the Engineering and Physical Sciences Research Council, U.K.; and the Defence and Security Accelerator, U.K.

Chang-Tsun Li (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the National Defence University (NDU), Taiwan, in 1987, the M.Sc. degree in computer science from the U.S. Naval Postgraduate School, USA, in 1992, and the Ph.D. degree in computer science from the University of Warwick, U.K., in 1998. He was an Associate Professor with the Department of Electrical Engineering, NDU, from 1998 to 2002, and a Visiting Professor with the Department of Computer Science, U.S. Naval Postgraduate School, in 2001. He was a Professor with the Department of Computer Science, University of Warwick, U.K., until January 2017, and a Professor with Charles Sturt University, Australia, from January 2017 to February 2019. He is currently a Professor with the School of Information Technology, Deakin University, Australia. His research interests include multimedia forensics and security, biometrics, data mining, machine learning, data analytics, computer vision, image processing, pattern recognition, bioinformatics, and content-based image retrieval. The outcomes of his multimedia forensics research have been translated into award-winning commercial products protected by a series of international patents and have been used by a number of police forces and courts of law around the world. He involved in the organization of many international conferences and workshops and also served as a member of the international program committees for several international conferences. He is also actively contributing keynote speeches and talks at various international events. He is currently an Editor of the *EURASIP Journal on Image and Video Processing* (JIVP) and an Associate Editor of *IET Biometrics*.