frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 12 April 2023
DOI 10.3389/fnins.2023.1136934

OPEN ACCESS

EDITED BY
Yinghui Kong,
North China Electric Power University, China

REVIEWED BY
Gengshen Wu,
City University of Macau, Macao SAR, China
Aibin Chen,
Central South University of Forestry and Technology, China

*CORRESPONDENCE
Ke Zhang
✉ zhangkeit@ncepu.edu.cn

SPECIALTY SECTION
This article was submitted to Visual Neuroscience, a section of the journal Frontiers in Neuroscience

RECEIVED 03 January 2023
ACCEPTED 21 February 2023
PUBLISHED 12 April 2023

CITATION
Shi C, Zhao S, Zhang K, Wang Y and Liang L (2023) Face-based age estimation using improved Swin Transformer with attention-based convolution. *Front. Neurosci.* 17:1136934. doi: 10.3389/fnins.2023.1136934

Face-based age estimation using improved Swin Transformer with attention-based convolution

Chaojun Shi[1,2], Shiwei Zhao[1], Ke Zhang[1,2]*, Yibo Wang[1] and Longping Liang[1]

[1]Department of Electronic and Communication Engineering, North China Electric Power University, Baoding, Hebei, China, [2]Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding, Hebei, China

Recently Transformer models is new direction in the computer vision field, which is based on self multihead attention mechanism. Compared with the convolutional neural network, this Transformer uses the self-attention mechanism to capture global contextual information and extract more strong features by learning the association relationship between different features, which has achieved good results in many vision tasks. In face-based age estimation, some facial patches that contain rich age-specific information are critical in the age estimation task. The present study proposed an attention-based convolution (ABC) age estimation framework, called improved Swin Transformer with ABC, in which two separate regions were implemented, namely ABC and Swin Transformer. ABC extracted facial patches containing rich age-specific information using a shallow convolutional network and a multiheaded attention mechanism. Subsequently, the features obtained by ABC were spliced with the flattened image in the Swin Transformer, which were then input to the Swin Transformer to predict the age of the image. The ABC framework spliced the important regions that contained rich age-specific information into the original image, which could fully mobilize the long-dependency of the Swin Transformer, that is,extracting stronger features by learning the dependency relationship between different features. ABC also introduced loss of diversity to guide the training of self-attention mechanism, reducing overlap between patches so that the diverse and important patches were discovered. Through extensive experiments, this study showed that the proposed framework outperformed several state-of-the-art methods on age estimation benchmark datasets.

KEYWORDS
age estimation, Swin Transformer, attention mechanism, deep learning, neural networks

1. Introduction

A large amount of useful information in facial images, such as age, gender, identity, race, emotion, and so forth (Angulu et al., 2018), and research on techniques related to facial image analysis has become the focus of computer vision. What is the significance of the facial age as an important feature? As an important physical and social characteristic of human beings, age plays a fundamental role in human social interaction. Recently, age estimation based on facial images is already an important research topic (Zhao et al., 2020; Agbo-Ajala and Viriri, 2021), which predicts the age corresponding to the image containing the face in the image. The task has very good application prospects in various intelligent fields, such as cross-age face recognition, intelligent security surveillance, harmonious human-computer interaction, image and video retrieval, face-based age prediction, and marketing analysis (Bruyer and Scailquin, 1994; Geronimo et al., 2009; Song et al., 2011; Angulu et al., 2018; Pei et al., 2019).

Modern face-based age estimation methods typically consist of two directions. One is to improve the learning ability of the neural network, and the other is to use other features related to the age of the face to assist the learning of the network. Convolutional networks can learn age features in facial images by multilayer convolution and have achieved great success in the field of computer vision with a wide range of applications. With the growing popularity of convolutional neural networks (CNNs), recent work on face-based age estimation has used these networks as a backbone (Shen et al., 2018; Dagher and Barbara, 2021; Sharma et al., 2022; Zhang and Bao, 2022). Most of these works improve the learning ability of the network by increasing the number of layers of convolutional layers (Dornaika et al., 2020; Yi, 2022) and improving the structure of the network. However, as convolutional networks continue to improve, the potential of CNN-based facial age estimation models has been exploited, and the increasing number of network model parameters has raised the cost of training. Therefore, we proposed to use a new network model designed specifically for face-based age estimation. Some other recent studies on face-based age estimation (Deng et al., 2021; Lu et al., 2022; Wang et al., 2022) have improved the accuracy of age estimation by extracting age features of faces and the relationship between different ages (Akbari et al., 2020; Xia et al., 2020). These studies have enhanced the learning ability of the network using attention-related mechanisms, using features other than faces such as sex, gender (Liu et al., 2020), and label distribution (Zeng et al., 2020; Zhao et al., 2020). However, these efforts may destroy the features and the structure of the original image while extracting the age features of the images, resulting in the loss of age information. Therefore, how to extract features without destroying the extracted features and the structure of the original image is a research direction.

Recently, the self-attention mechanism and Transformer model (He et al., 2021) have attracted great attention in computer vision and natural language processing tasks. Vision Transformer (Dosovitskiy et al., 2020) has shown that the Transformer-based model indeed contains the capacity as the backbone network instead of former pure CNN model in image synthesis and classification tasks (He et al., 2016; Szegedy et al., 2017). Related research (Bourdev et al., 2011) shows that compared with CNN, the self-attention mechanism of Transformer is not limited by local interactions and allows both long-distance dependencies and computes in parallel and learn the most appropriate inductive bias according to different task goals, which has achieved good effects in many vision tasks. The attention mechanism (Niu et al., 2021) is also a direction to improve the prediction ability of the network, and the self-attention mechanism (Lin et al., 2017) of the Transformer can quickly extract the important features of sparse data, which is an improvement of the attention mechanism. It reduces the reliance of the network on external information and is better at capturing the correlation within the data or features. Therefore, the self-attention mechanism can be designed to exploit age-specific patches during training to boost the performance of face-based age estimation methods.

In this study, we proposed an attention-based convolution (ABC) age estimation framework called an improved Swin Transformer with ABC. The architecture of an improved Swin Transformer with ABC is illustrated in **Figure 1**. The core of the improved Swin Transformer with ABC was the ABC. In ABC, two separate regions were implemented, namely shallow convolution and a multiheaded self-attention mechanism. The feature extractor performed initial feature extraction of the image using a shallow convolutional network and then further extracted the corresponding patches that might contain rich age-specific information through the multiheaded attention mechanism, according to the number of probes. ABC extracted facial patches containing rich age-specific information by a shallow convolutional network and multiheaded attention mechanism. Subsequently, the features obtained by ABC were spliced with the flattened image in the Swin Transformer, which were then input to the Swin Transformer to predict the age of the image. ABC also introduced loss of diversity to guide the training of self-attention mechanism, reducing overlap between patches so that the diverse and important patches were discovered. The contributions of this study were as follows:

(1) We proposed a new module called ABC that used shallow convolution and shifted window self-attention mechanism. The image was initially extracted and refined by shallow image convolution, and some facial regions containing rich age-specific information were extracted by the multiheaded self-attention mechanism.
(2) We introduced the Swin Transformer as the backbone model, and the features obtained by ABC were spliced with the output of the embedding layer of the Swin Transformer in the channel dimension and subsequently input to the rest of the Swin Transformer so that the model could better capture the global and local information of the image and achieve better results by learning the relationship between different features of the image.
(3) We introduced a new loss function that made each probe of the multiheaded self-attention mechanism reveal a different age region and prevent overlapping feature extraction from the multiheaded feature extractor.

2. Related research

We reviewed and discussed related studies on face-based age estimation to lay the foundation for this study. We also reviewed the attention mechanism and Transformer, both of which were relevant to the proposed ABC.

2.1. Face-based age estimation

In the last few decades, many researches have been conducted on face-based age estimation. One of the earliest studies can be traced back to (Kwon and da Vitoria Lobo, 1999), in which

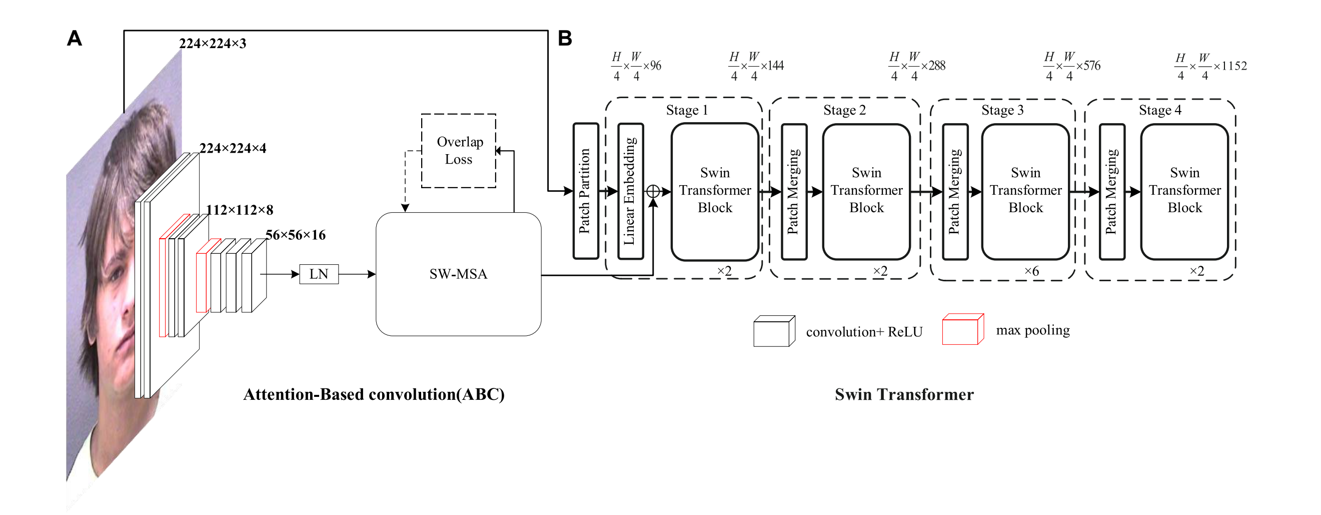

FIGURE 1
Structure of an improved Swin Transformer with attention-based convolution (ABC). **(A)** Structure of an attention based convolution. **(B)** Swin Transformer.

the researcher's classified faces into three age groups based on the craniofacial development theory and wrinkle analysis. In the traditional face-based age estimation methods, first the face-based age features are extracted, and then classification and regression models are established for face-based age estimation. With the rapid development of deep learning in recent years, deep learning–based facial age estimation methods have significantly improved the accuracy and robustness of face-based age estimation, especially the accuracy of face-based age estimation under unconstrained conditions.

Yi et al. (2014) proposed a multistream CNN to better leverage high-dimensional structured information in facial images. The authors cropped multiple patches from facial images so that each stream learned from one patch. Then, the features extracted from different patches were fused before the output layer. Ranjan et al. (2015) used a deep convolutional neural network (Chen et al., 2016) with 10 convolutional layers, 5 pooling layers, and 1 fully connected layer to extract the age characteristics of facial images. Then, age regression was performed using a three-layer artificial neural network. One of the first studies to use CNNs for the face-based age estimation problem was (Wang et al., 2015), in which a CNN with two convolutional layers was deployed. Han et al. (2017) used a modified AlexNet (Krizhevsky et al., 2017) to develop a multitask learning method for heterogeneous face attribute estimation including the age. Rothe et al. (2018) transformed the regression problem into a classification-regression problem and proposed a deep expectation network (DEX) for the age estimation of representations. The DEX network changed the number of neurons in the last layer of the VGG-16 network.

In general, CNN-based methods for face-based age estimation can be divided into two categories. The first approach is to use deeper and better deep learning models as backbone networks with better networks to extract features (Dornaika et al., 2020; Lu et al., 2022; Sunitha et al., 2022; Wang et al., 2022). The second approach is to improve the performance of the network in terms of other attributes of the face, such as race features and gender features (Zhang et al., 2017a; Liu et al., 2020; Deng et al., 2021), and relational features of different ages (Song et al., 2011; Gao et al., 2017; Zeng et al., 2020). In our previous study, we proposed a multilevel residual network model to further improve the performance of the network so as to better improve the accuracy of age estimation (Levi and Hassner, 2015). As the CNN-based network model for face-based age estimation may soon reach a bottleneck in the direction of face-based age estimation, we tried to incorporate a newer network model that combined the advantages of CNN with the advantages of the new network to achieve better results.

2.2. Attention-based facial age estimation

Attention mechanism (Vaswani et al., 2017) mimics the internal processes of biological observation behavior, increasing the fineness of observation in certain regions. The attention mechanism can quickly extract important features of sparse data and is therefore widely used in machine translation, speech recognition, and image processing (Wang et al., 2016), and other fields. A multiheaded attention mechanism can attend to multiple informative segments of the input with an attention head attending to one specific segment. Therefore, the number of segments that the multiheaded attention mechanism can attend to is determined by the number of attention heads. In computer vision, the self-attention layer takes the feature map as input and calculates the attention weights between each pair of features to generate an updated feature map where each position has information about any other feature in the same image. These layers can replace convolution directly or be combined with convolution layers. They are able to handle larger perceptual fields than conventional convolution, and therefore can acquire dependencies between some long-distance interval features on the space.

Multiple attention mechanisms have been proposed for visual tasks to address the weaknesses of convolutions due to the aforementioned characteristics. Bello et al. (2019) proposed to augment convolutional networks with self-attention by concatenating convolutional feature maps with a set of feature maps produced *via* a novel relative self-attention mechanism. Hu et al. (2018) proposed a simple, lightweight approach for better context exploitation in CNNs by introducing a pair of operators: gather (which efficiently aggregated feature responses from a large spatial extent) and excite (which redistributed the pooled information to local features). Chen et al. (2018) proposed the "double attention block," which was designed with a double attention mechanism in two steps, a novel component that aggregated and propagated informative global features from the entire spatiotemporal space of input images/videos, enabling subsequent convolution layers to access features from the entire space efficiently. The nonlocal operation proposed by Wang et al. (2018) computed the response at a position as a weighted sum of the features at all positions and achieved excellent results in the experiment. Wang et al. (2022) proposed a face-based age estimation framework called attention-based dynamic patch fusion (ADPF). The ADPF dynamically located and ranked age-specific patches by employing a novel ranking-guided multihead hybrid attention mechanism and used the discovered patches along with the facial image to predict the age of the subject.

A previously proposed work (Bello et al., 2019) adds an attention mechanism to the convolutional network to enhance the convolutional network. A previously proposed work (Hu et al., 2018) enhances the input image by using the attention mechanism to get the feature responses of adjacent regions of the image. A previously proposed work (Chen et al., 2018) get the global features and local features by two attention mechanisms, respectively. The previously proposed work (Wang et al., 2018, 2022) learn the important features by calculating the region weights through the attention mechanism. Different from the above work which uses the attention mechanism to enhance the deep convolutional network, our work uses shallow convolution and attention mechanism with the aim of finding the important regions without destroying the original image and stitching these regions with the original image to enhance the Transformer network.

2.3. Vision transformer

Transformer (Vaswani et al., 2017) is a kind of deep divine meridian based on the self-attention mechanism. In recent years, Transformer-based models have become a popular research direction in the field of computer vision. Another visual Transformer model, ViT, recently proposed by Dosovitskiy et al. (2020), achieved state-of-the-art performance on several image recognition benchmark tasks in a structure that fully adopted the standard structure of a Transformer, ViT. Liu Z. et al. (2021) proposed the Swin Transformer, which enabled the flexibility of the Transformer model to handle images of different scales by applying a hierarchical structure similar to that of CNN. The Swin Transformer used a windowed attention mechanism to greatly reduce the computational complexity. The architecture of a Swin Transformer is illustrated in **Figure 2**. Yuan et al. (2021) proposed CeiT, combined with the ability of CNN to extract low-level features, to design an Image-to-Tokens module, which extracted the patch from the generated low-level features. Wang et al. (2021) proposed the CrossFormer, which used a cross-scale embedding layer (CEL), generated patch embeddings using a CEL at each stage, and extracted features using four different-sized convolutional kernels in the first CEL layer. The features were extracted using four convolutional kernels of different sizes, and the results of the convolutional kernels were stitched into patch embeddings. Peng et al. (2021) proposed Conformer (Yuan et al., 2021)], which combined the features of a Transformer and CNN through a parallel structure to achieve feature fusion by bridging each other, so that the local features of CNN and the global features of the Transformer could be retained to the maximum extent. Xiao et al. (2021) proposed ViTc, which replaced the patch embedding module in the Transformer with convolution. It made the replaced Transformer more stable and converge faster. Liu Z. et al. (2021) proposed the TransCNN by introducing a CNN layer after the self-attention block so that the network could inherit the advantages of a Transformer and CNN. UniFormer (Li et al., 2022) seamlessly integrated the advantages of convolution and self-attention through the Transformer to aggregate the local and global features at shallow and deep layers, respectively, solving the problem of redundancy and dependency for efficient representation learning.

A previous study proposed (Yuan et al., 2021) replacing the original three structures of the Transformer with convolutional layers in the Transformer, thus integrating CNN into the Transformer. Another previous study proposed (Wang et al., 2021) designing a CEL and long short distance attention to replace the Transformer block and MSA (Multihead Self-Attention) in the Transformer. A related study (Li et al., 2022) seamlessly integrated the merits of convolution and self-attention in a concise Transformer format and proposed a new framework UniFormer. Some previous studies proposed (Wang et al., 2021; Yuan et al., 2021; Li et al., 2022) combining CNN with the structure in the Transformer block to improve the capabilities of the network by adding or replacing it so as to improve the framework of the Transformer. However, the present study designed an independent shallow convolution, which extracted features through a self-attentive mechanism to be stitched with the original image and input to the Transformer, without modifying the Transformer. The CNN performed the role of extracting shallow features in the network while preserving the original structural information of the image. Another study (Xiao et al., 2021) introduced convolution to process the input original image by replacing the patch layer with CNN. A related study (Liu Z. et al., 2021) viewed image patches as tokens by CNN, which continuously extracted features through a multilayer self-attention mechanism and finally input to the Transformer block. The present study extracted shallow features by shallow convolution without destroying the location structure information of the image. The feature regions rich in face-based age information were obtained through a self-attention mechanism and stitched with the original image after the patch layer processing in the Transformer, instead of directly processing the original image. A related study (Peng et al., 2021) combined the features of the Transformer and CNN through a parallel structure using a bridge to achieve feature fusion. This study introduced CNN and self-attention mechanism to extract shallow features while preserving the original structural information of the image. The purpose of the CNN and self-attention mechanism was to improve the ability of the Transformer to obtain feature dependencies at long distances.

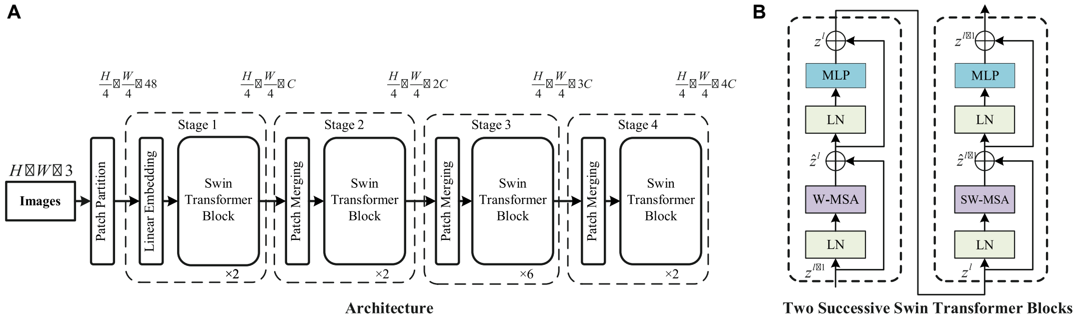

FIGURE 2
Architecture of a Swin Transformer: **(A)** Structure of a Swin Transformer (Swin-T). **(B)** Two successive Swin Transformer blocks. Window based self-attention (W-MSA) and shifted window based self-attention (SW-MSA) are multihead self-attention modules with regular and shifted windowing configurations, respectively (Vaswani et al., 2017).

3. Methodology

In this section, we first discussed the core of the Swin Transformer with the attention-based convolution mechanism, which was the proposed ABC. Then, we combined the ABC and Swin Transformer parts. Finally, we introduced the age expectation and loss function. The architecture of the Swin Transformer with the ABC mechanism is shown in **Figure 1**.

3.1. ABC

As the Swin Transformer with attention-based convolution is based on ABC and the critical component of ABC is the self-attention mechanism, we first introduced the self-attention mechanism followed by the proposed attention-based convolution. Finally, we detailed the complete mechanism. The architecture of attention-based convolution is shown in **Figure 3**.

We considered an input tensor X with a dimension of $h \times w \times c$, where h denotes the height, w denotes the width, and c denotes the number of channels. During training, the input to ABC was a fixed-size 224 × 224 RGB image.

This model started with the institutional area of the convolution. The image was passed through a stack of convolutional layers; this convolutional region had eight convolutional layers, where we used filters with a very small receptive field: 3 × 3. The convolution stride was fixed to 1 pixel; the spatial padding of the convolution layer input was such that the spatial resolution was preserved after convolution, that is, the padding was one pixel for 3 × 3 convolution layers. Spatial pooling was carried out by two max-pooling layers, which followed the second and fourth convolutional layers. Max-pooling was performed over a 2 × 2 pixel window, with stride two. All hidden layers were equipped with the rectification nonlinearity.

The convolution was followed by the region of the self-attention structure. The output of the convolutional layer X was used as the input of the attention mechanism. Let us consider an input tensor X with a dimension of $h \times w \times c$, where h denotes the height, w denotes the width, and c denotes the number of channels. X was convolved into three separate tensors: Q with a shape of $h \times w \times c_Q$, K with a shape of $h \times w \times c_K$, and V with a shape of $h \times w \times c_V$, where c_Q, c_k, and c_V indicate the number of channels in the corresponding tensor. The intention behind self-attention was to compute a weighted summation of the values, V, where the weights were computed as the similarities between the query, Q, and the corresponding key, K. Therefore, to compute the similarity, Q and K normally had the same shape, that is, $c_Q = c_V$. The output of a single self-attention mechanism was computed as:

$$S_{n_1} = soft\max\left(\frac{Q' \cdot K'^T}{\sqrt{c_K}}\right) \cdot V \tag{1}$$

where Q' and K' are flattened tensors to perform the dot product.

After the scaling operation, that is, dividing the similarity matrix $Q' \cdot K'^T$ by a factor of $\sqrt{c_K}$ and applying the softmax function, we performed a dot product between the normalized similarity matrix and V to generate the self-attention maps S_n with a dimension of $h \times w \times c_K$. n_i is the number of heads of attention probes in the multiheaded attention mechanism.

As we flattened the two-dimensional feature maps into a one-dimensional vector in Equation 1, the original structure of the feature maps was distorted. We adopted the relative positional encoding to make it efficient when dealing with structured data such as images and multidimensional features. Specifically, the relative positional encoding was represented by the attention logit, which encoded how much an entry in Q' attended to an entry in K'. The attention logit was computed as:

$$l_{i,j} = \frac{q_i^T}{\sqrt{c_K}}(k_j + r^w_{j_x - i_x} + r^h_{j_y - i_y}) \tag{2}$$

where q_i is the ith row in Q' indicating the feature vector for pixel $i := (i_x, i_y)$ and k_j is the jth row in K' indicating the feature vector for pixel $j := (j_x, j_y)$. $r^w_{j_x - i_x}$ and $r^h_{j_y - i_y}$ are learnable parameters encoding the positional information within the relative width $j_x - i_x$ and relative height $j_y - i_y$, respectively. With the relative positional encoding, the output of a single self-attention mechanism could be reformulated as:

$$S_{n_1} = soft\max\left(\frac{Q' \cdot K'^T + m_h + m_w}{\sqrt{c_K}}\right) \cdot V \tag{3}$$

where $m_h\left[i, j\right] = q_i^T r^h_{j_y - i_y}$ and $m_h\left[i, j\right] = q_i^T r^w_{j_x - i_x}$ are matrices of relative positional logits. In this study, the number of heads

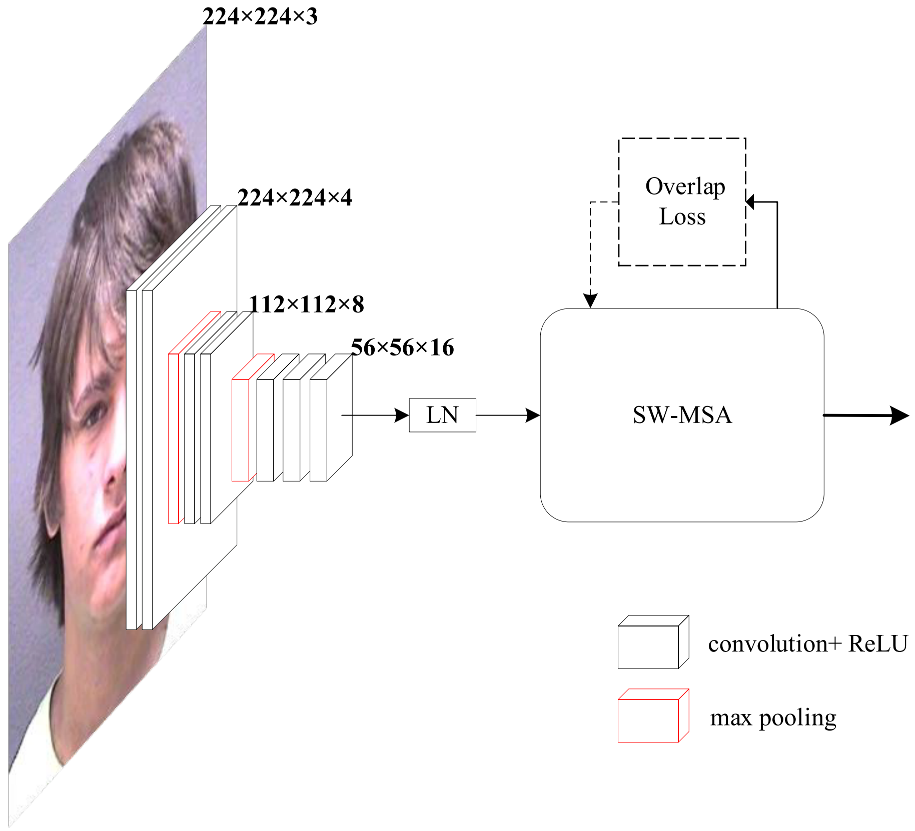

FIGURE 3
Structure of the proposed attention-based convolution (ABC).

of the attention probe in the multiheaded attention mechanism was set to four.

A key design element of ABC was its shift of the window partition between consecutive self-attention layers, as illustrated in Figure 2. The shifted windows bridged the windows of the preceding layer, providing connections among them that significantly enhanced the modeling power. As illustrated in Figure 2, the module used a regular window partitioning strategy that started from the top-left pixel, and the 56 × 56 feature map was evenly partitioned into 8 × 8 windows of size 7 × 7 ($M = 7$). Then, the module adopted a windowing configuration that was shifted from that of the preceding layer by displacing the windows by $([\frac{M}{2}], [\frac{M}{2}])$ pixels from the regularly partitioned windows.

With the shifted window partitioning approach, the attention blocks were computed as follows:

$$\hat{X}_1 = W - MSA(\hat{X}) \tag{4}$$

where $\hat{X}_1$ denotes the output features of the shallow convolution module.

3.2. Swin Transformer with ABC mechanism

We again used the sample image input at the beginning with the tensor Y as input to the Transformer. First, the tensor Y went through the patch partition layer and the dimension became 56 × 56 × 48. Then, Y was again mapped to the specified dimension by the linear embedding layer, and the dimension of Y was 56 × 56 × 128. The role of the patch partition module was to crop a patch_size * patch_size block of the input original image by conv2d. The patch_size was set to four. The specific structure of the Swin Transformer is shown in the Figure 2. The remaining structure of the Swin Transformer could be referred from the original study, and as no modifications were made to the Swin Transformer in this study, it was not repeated in this study.

In the output of the self-attention mechanism in Equation 4, which was the output of ABC, the output tensor X had a dimension of 56 × 56 × 16 and the output of the linear embedding layer of the Swin Transformer also had a dimension of 56 × 56 × 128, so we considered fusing the two tensors. We spliced the output X of ABC with the output Y of the embedding layer in the channel dimension to get Y_1, that had a dimension of 56 × 56 × 144. Then, we replaced Y with the spliced tensor Y_1, continued with the network layer behind the Swin Transformer, and finally got the final output Z at the end of the Swin Transformer.

Using a Transformer as the backbone network can make the model better to capture the global and local information of images, extract more powerful features, and achieve better effects by learning the relationship between different features of images. This specific framework can more effectively learn the long-distance dependencies in face semantic parts, which naturally helps model the strong attribute correlations (Liu et al., 2020).

Focusing on the ability of the Swin Transformer to mine both long-range dependencies and parallel computation, we first obtained several important age feature regions rich with age-specific information using ABC and stitched them with the original input image. This way, the Swin Transformer afterward could learn more dependencies between important age features and improve the learning ability of the network. We used the stitching in the connection part of the ABC and Swin Transformer. By stitching in the channel dimension, the stitched Y_1 at this point not only retained the structure and features of the original input image but also had feature regions that were processed by the shallow convolutional layer and the multiheaded attention mechanism. The ABC-derived feature regions not only highlighted the regions containing rich age information but also retained the structural information of the regions to the greatest extent by the shallow convolutional layer. The Swin Transformer could learn the relationship between the features of the original input image and the features obtained by ABC when processing the stitched Y_1, thus fully exploiting the advantage of the Swin Transformer. The use of the ABC+Swin Transformer enabled the model to better capture the global and local information of the image, which not only preserved the local sensitivity and translation invariance of CNN but also improved the ability of Swin Transformer to learn the dependencies between features for better results.

3.3. Loss

We used the label distribution learning to learn the exact age and Kullback–Leibler divergence to learn the age distribution to estimate the age.

Formally, let $x_i \in X$ denote the ith input instance with $i = 1, 2, ..., N$, and $\hat{y}_{i,j}$ denote the softmax distribution of the predicted values of the network, where $N$is the number of instances and c is the number of each age.

Deep label distributions transform y_i from a single class label to a label distribution and then predict $\hat{y}_i$ by label distribution learning. Instances with the same class label y_i share the identical Gaussian distribution:

$$l_{x_i}^{c_j} = \frac{1}{\sqrt{2\pi}\sigma M} \exp\left(-\frac{(c_j - y_i)^2}{2\sigma^2}\right) \tag{5}$$

where $l_i^{c_j}$ is the degree to which the value of each age c_j describes images x_i and $c_j = 0, 1, ..., C$; σ is the standard deviation of $l_b^{c_i}$; and the factor M lead $\sum_{j=0}^{a} l_{x_i}^{c_j} = 1$. In this study, σ was set to 2.

KL (Kullback-Leibler) tried to generate the predicted softmax probability distribution as similar to the ground truth distribution as possible. For the same random variable, with two different distributions $\hat{y}_i$ and $l_i^{c_j}$, the KL scatter of $\hat{y}_i$ to $l_i^{c_j}$ was expressed as:

$$KL = \sum_{x} \hat{y}_i \times \log_2 \frac{\hat{y}_i}{l_i^{c_j}} \tag{6}$$

The less the result of KL_{loss}, the less the difference between the two divisions, indicating that the less the difference between the true age and the estimated age, the more accurate the age estimated by the network.

The number of patches that could be discovered was determined by the number of attention heads implemented in ABC. However, during implementation, we found that patches tended to overlap, especially in informative regions. This overlap of attended patches might lead to redundant learning sources and leave other age-specific patches undiscovered. To alleviate this overlap issue, we used a diversity loss to learn diverse and nonoverlapping patches by minimizing the summation of products of corresponding entries in two attention heads, $S_{n_1}(h', w')$ and $S_{n_2}(h', w')$. The diversity loss was formulated as:

$$L_{Overlap} = \sum_{\substack{n_1, n_2 \\ n_1 \neq n_2}}^{n} \sum_{h'}^{h} \sum_{w'}^{w} S_{n_1}(h', w') \cdot S_{n_2}(h', w') \tag{7}$$

Each probe in the multiheaded attention mechanism generated a $S_n(h', w')$, and $S_n(h', w')$ represented the region that the probe focused on. $S_n(h', w')$ could be regarded as a weight matrix with dimension 56 × 56 × 16. In $S_n(h', w')$, the richer the region with age-specific information, the larger the weight matrix corresponding to that region. When the result obtained by multiplying $S_n(h', w')$ and the other $S_n(h', w')$ was 0, the regions attended by the two attention probes did not overlap. When the overlap loss obtained after multiplying all $S_n(h', w')$ with each other was zero, no overlap occurred between the regions attended by different probes, which prevented the redundancy of learning caused by multiple attention probes attending to the same region at the same time.

The overall loss to train this network was the summation of the two loss functions:

$$L = L_{KL} + \lambda_1 L_{Overlap} \tag{8}$$

where λ_1 is the hyperparameter that attempts to balance the influences of mean and residual sublosses in the combined loss function. In this study, λ_1 was set to 10.

The estimated age of the i-th test image could be calculated based on Equation 9. Each descriptive degree in the label distribution was multiplied by its corresponding label and then summed to obtain the final predicted age. Assuming that the network softmax layer output 62 probability values from age 16 years to age 77 years, $p_{i,j}$ corresponded to the probability of the prediction for 16–77 years, respectively, as shown in Equation 9:

$$\hat{y}_i = \sum_{j=16}^{77} \hat{y}_{i,j} \times c_j \tag{9}$$

4. Experiments

In this section, we first detailed the experimental settings and then compared our method with state-of-the-art studies on face-based age database MORPH Album II (Ricanek and Tesafaye, 2006), FG-NET dataset (Cootes et al., 2001), and Adience dataset (Eidinger et al., 2014). We removed the important design elements of ABC. In this study, Swin Transformer used the model parameter settings of the base.

4.1. Implementation details

The effectiveness of the methods in this study was demonstrated using the model as the backbone network for image classification tasks. For training and testing on the MORPH Album II, FG-NET, and Adience datasets, the size was set to 64, the number of iterations to 800, the weight decay to 0.005, the momentum to 0.9, and the default learning rate to 0.0001. The learning rate was multiplied by 0.1 when training to the 500th iteration, and by 0.01 when training to the 650th iteration.

The facial images in MORPH Album II and FG-NET datasets had corresponding specific age values. For the evaluation metrics of these two datasets, we used MAE. The age labels corresponding to the images in the Adience dataset were age groups, such as 0–2 and 4–6. For the evaluation metrics of the Adience dataset, we used the accuracy of a one-class classification.

The MAE was calculated as shown in Equation 10, where y_j is the age label value, $y_j^{'}$ is the age prediction value, and N is the number of test images. The accuracy of the one-class classification was calculated as shown in Equation 11, where TP is the correctly predicted sample and FP is the incorrectly predicted sample.

$$\mathrm{MAE} = \frac{1}{N}\sum_{j=1}^{N}|y_j - y_j^{'}| \tag{10}$$

$$precision = \frac{TP}{TP + FP} \tag{11}$$

4.2. Dataset

We conducted experiments on three commonly used face-based age estimation benchmark datasets: MORPH Album II, FG-NET, and Adience.

The MORPH Album II is one of the most common and largest longitudinal face databases in the public domain for age estimation, containing 55,134 facial images of 13,617 subjects. The age ranged from 16 to 77 years with an average age of 33 years, and the male-to-female ratio was 5.5:1. Each facial image in the MORPH II dataset was associated with identity, age, race, and sex labels. 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, no identity overlap occurred between the two subsets. Some images are shown in **Figure 4**.

The FG-NET dataset contained 1,002 facial images from 82 noncelebrity subjects with large variations in lighting, pose, and expression. The age ranged from 0 to 69 years (on average, 12 images per subject). Each subject in this dataset had more than 10 facial images taken over a long time span. In addition, the facial images in this dataset contained pose, illumination, and expression variations. For the FG-NET dataset, we use the leave-one-person-out (LOPO) strategy. 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-fold and the reported results are the average values. Some images are shown in **Figure 5**.

The Adience dataset was collected from the photos taken by users themselves from the Yahoo image-sharing website. It contained 26,580 images of 2,284 people with age ranging from 0 to 100 years. The labels used in this dataset were age group labels: 0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, and 60–100, categorized into eight groups. Some images with age labels were labeled as age values, some as other age ranges, and some others as empty. In this study, the images with age labels labeled as age values, other age ranges, and null were removed. The age classification experiments using the Adience dataset were treated as eight classes. Some images are shown in **Figure 6**.

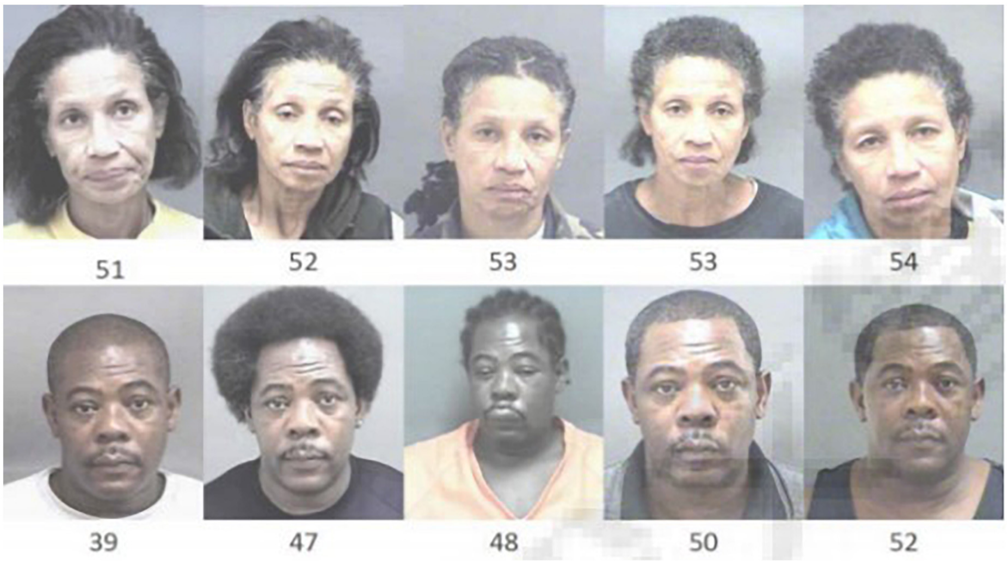

FIGURE 4
Images in the MORPH Album II dataset.

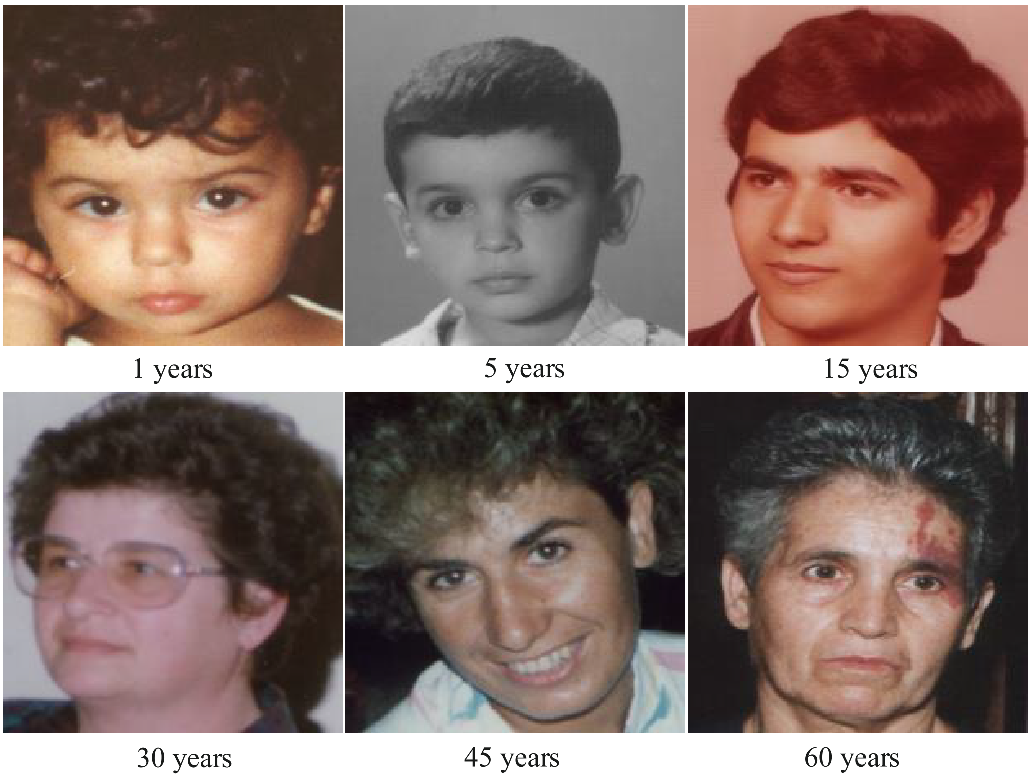

FIGURE 5
Images in the FG-NET dataset.

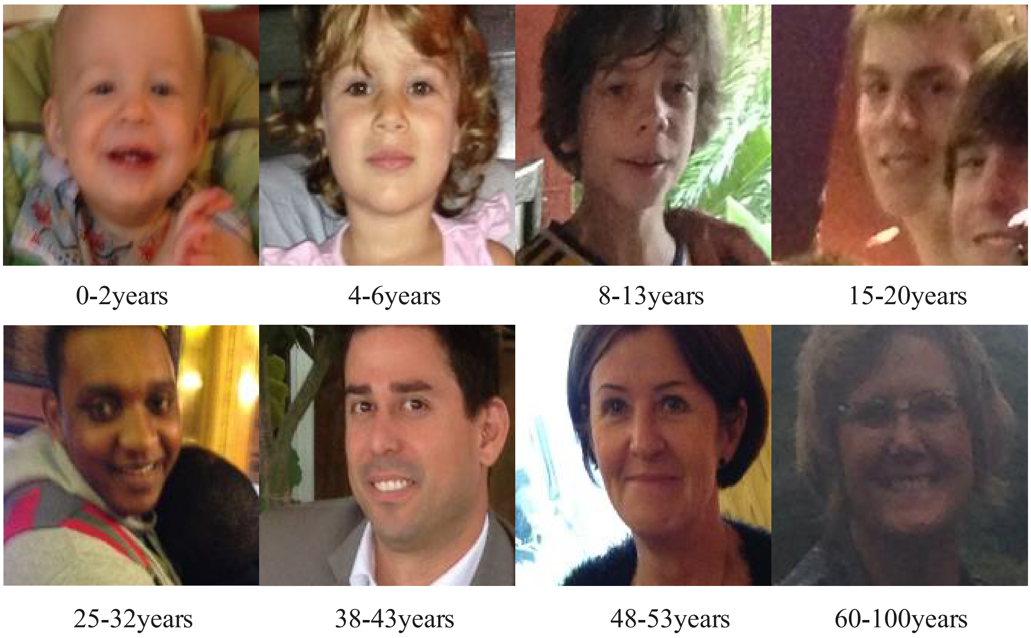

FIGURE 6
Images in the Adience dataset.

TABLE 1 Comparison of age classification on the MORPH Album II dataset.

Method	MAE
OHRank (Chang et al., 2011)	8.83
MTWGP (Zhang and Yeung, 2010)	6.28
OHRank (Chang et al., 2011)	6.07
CA-SVR (Chen et al., 2013)	5.88
SVR (Guo et al., 2008)	5.77
LDL (Geng et al., 2013)	4.87
DLA (Wang et al., 2015)	4.77
ALDL (Geng et al., 2013)	4.34
KPLS (Guo and Mu, 2011)	4.18
BIF+KCCA (Guo and Mu, 2013)	3.98
VGG (Rothe et al., 2016)	3.45
ARN (Agustsson et al., 2017)	3.25
VGGNetHybrid (Xing et al., 2017)	2.96
DAG-VGG16 (Li et al., 2019)	2.81
Mean-Variance Loss (Pan et al., 2018)	2.80
MSFCL-KL (Xia et al., 2020)	2.73
DEX (IMDB-WIKI) (Rothe et al., 2018)	2.68
VDAL (Liu H. et al., 2020)	2.57
ADPF (Wang et al., 2022)	2.54
Anet+Gnet+Rnet (Deng et al., 2021)	2.47
Swin Transformer (Liu Z. et al., 2021)	2.37
ABC+Swin Transformer (Ours)	**2.17**

Bold values represent the best data.

4.3. Evaluations on the MORPH Album II dataset

The MAE values for the aforementioned settings of the MORPH Album II dataset are tabulated in Table 1. As shown in the table, the ABC+Swin Transformer outperformed all state-of-the-art methods. Compared with the other methods on the MORPH Album II, our proposed method improved by 55.45% over the LDL method, by 37.2% over the VGG method, by 15.6% over the VADL method, and by 14.6% over the ADPF method. Compared with the original Swin Transformer, our proposed ABC+Swin Transformer improved by 8.4%. The experimental results showed that our method had high accuracy on the MORPH Album II dataset and improved the learning ability of the facial image age of the Swin Transformer.

4.4. Evaluations on the FG-NET dataset

As the number of images in the FG-NET dataset was too small, training the dataset directly might lead to a decrease in the accuracy of the results and slow convergence. Therefore, we used the pretraining weights obtained in the aforementioned FG-NET dataset as the initial weights for training. The MAE values for the aforementioned settings of the FG-NET dataset are tabulated in Table 2. As shown in the table, the ABC+Swin Transformer outperformed all state-of-the-art methods. Compared with the other methods on the FG-NET dataset, the method proposed in this study improved by 21.8% over the DAG-VGG16 method, by 15.8% over the ADPF method, and by 8.6% over the BridgeNet method. Compared with the original Swin Transformer, the performance of our proposed ABC+Swin Transformer improved by 4.8%. The experimental results showed that our method had high accuracy on the FG-NET dataset and improved the learning ability of facial image age of the Swin Transformer.

TABLE 2 Comparison of age estimates on the FG-NET dataset.

Method	MAE
Mean-Variance Loss (Pan et al., 2018)	4.10
DRFs (Dagher and Barbara, 2021)	3.85
M-LSDML (Liu et al., 2017)	3.74
DLDFL (Shen et al., 2019)	3.71
DFR (Shen et al., 2019)	3.47
DAG-VGG16 (Taheri and Toygar, 2019)	3.08
DAG-GoogleNet (Panis et al., 2016)	3.05
AGEn (Tan et al., 2017)	2.96
C3AE (Zhang et al., 2019)	2.95
ADPF (Wang et al., 2022)	2.86
Swin Transformer (Liu Z. et al., 2021)	2.69
Anet+Gnet+Rnet (Deng et al., 2021)	2.59
BridgeNet (Hou et al., 2016)	2.56
ABC+Swin Transformer (Ours)	**2.52**

Bold values represent the best data.

4.5. Evaluations on the Adience dataset

In this study, the images of the Adience dataset without age labels or confusing age labels in the dataset were removed. Experimentally, when using the Adience dataset for training and testing, the dataset was split into five groups using the 5-fold cross-validation method: 0-, 1-, 2-, 3-, and 4-fold. Each time when

TABLE 3 Comparison of age estimates on the Adience dataset.

Models	Accuracy of a one-class classification (%)
SVM-dropout (Shen et al., 2019)	45.1 ± 2.6
CNN+over-sampling (Hou et al., 2016)	50.7 ± 5.1
R-SAAFc1 (Zhang et al., 2017b)	52.9
R-SAAFc2 (Zhang et al., 2017b)	53.5
Classification (Liu et al., 2018)	54.0 ± 6.3
Chained Net (Zhang et al., 2017a)	54.5
DEX w/o IMDB-WIKI Pretrain (Rothe et al., 2018)	55.6 ± 6.1
ABC+Swin Transformer (Ours)	**56.1**

Bold values represent the best data.

TABLE 4 Accuracy of each age class on the Adience dataset.

Fold	0–2	4–6	8–12	15–20	25–32	38–43	48–53	60–100
0	82.59	37.70	39.38	32.92	72.36	39.51	30.31	52.73
1	85.31	59.38	41.29	35.68	58.65	37.11	31.32	52.19
2	64.45	65.19	58.04	17.03	76.95	35.16	19.38	61.46
3	50.00	71.63	38.02	36.13	61.98	40.63	21.88	44.38
4	84.58	62.95	61.97	29.84	71.38	33.99	27.40	50.00
Average	73.39	59.37	47.74	30.20	68.26	37.28	26.06	52.15

TABLE 5 MAE values for the attention-based convolution (ABC) framework on the MORPH Album II dataset for different numbers of attention probes.

Number of attention probes	0	1	2	4	8	16
ABC+Swin Transformer (Ours)	2.243	2.239	2.228	2.170	2.212	2.236

TABLE 6 MAE values for the attention-based convolution (ABC) framework on the MORPH II for different values of λ_1.

Value of λ_1	1	5	7.5	10	12.5	15	20
ABC+Swin Transformer (Ours)	2.37	2.25	2.19	2.17	2.20	2.27	2.38

training and testing, one group of images was used for testing, the remaining four groups were combined into a training set for training, and the training-testing was done a total of five times. The final test set of the age group classification results was taken as the average of the five times results. The Adience dataset used an evaluation method for single-age classification accuracy. The accuracy for the aforementioned settings of the Adience dataset are tabulated in Table 3.

The accuracy of each age class of folds on the Adience dataset is shown as follows. Due to the random grouping in the images, the accuracy of each fold is different. Overall, the accuracy of each age class is essentially equal between different folds. This demonstrates that the model is stable in training. The accuracy of each age class for the aforementioned settings of the Adience dataset are tabulated in Table 4.

4.6. Ablation study

We conducted ablation experiments to demonstrate the effectiveness of each component of the ABC+Swin Transformer. Specifically, we used the number of heads of attention probes in the multiheaded attention mechanism as a variable to demonstrate the necessity of each component. The MAE values for the aforementioned settings of the MORPH Album II dataset are tabulated in Table 5.

Table 5 shows that when the number of attention probes of ABC was zero, the MAE obtained by the ABC+Swin Transformer on the dataset was better than that obtained by the Swin Transformer, which demonstrated that the convolution performed an effective function in the network and enhanced the network's ability to learn the age information of facial images. The MAE obtained using different numbers of attention probes showed that the optimum results were obtained when the number of probes was four. The reason for this was the presence of four important regions with rich age-specific information in the face. The more the number of probes, the more completely the age-specific regions could be extracted. As the number of probes increased, the intervals noticed by different probes might overlap, leading to redundancy in information extraction, which was not conducive to the information extraction of the network.

As shown in Table 6, the network was difficult to converge when λ_1 was too small (KL_{loss} took a dominant role) and a big performance degradation occurred when λ_1 was too large (the overlap loss took a dominant role). Within a long reasonable range, our proposed method performed stably. The number of heads of the multihead attention mechanism was set to four.

5. Conclusion

In this study, we proposed the ABC framework to improve the performance of face-based age estimation tasks and combined ABC with the Swin Transformer to obtain better prediction performance. Our framework combined the shallow convolution and the multiheaded attention mechanism using the shifted window approach. The shallow convolution used several layers of a convolutional network with few convolutional kernels to condense the information, enhance the features of the image, and process the input to the same size for stitching with the information of the later attentional network computations and the Swin Transformer. The multiheaded attention mechanism allowed the network to learn and find regions that contained rich age-specific information, and display these regions. Finally, the age-rich regions obtained by the ABC framework were spliced with the image that was initially processed by the Swin Transformer along the channel dimension, and the stitched image tensor was carried out to the subsequent network of the Swin Transformer together to compute the final age prediction.

The significant age feature regions obtained by ABC were stitched with the original input images and then passed through the Swin Transformer for face-based age estimation, which made excellent use of the ability of the Swin Transformer to mine long-distance dependencies and parallel computation to learn more dependencies between important age features. The addition of the ABC framework well compensated the image local sensitivity and translation invariance of the Swin Transformer. The ABC framework spliced the important regions that contained rich age-specific information into the original image, which could fully mobilize the long-dependency of the Swin Transformer, that is, extracting stronger features by learning the dependency relationship between different features. As a result, the entire network could not only extract the important face-based age information regions but also further improve the prediction accuracy using the ability of the Swin Transformer to learn the interrelationship between features.

Based on the evaluation of several benchmark datasets, ABC significantly improved the prediction accuracy compared with several state-of-the-art methods. Future studies should investigate the design of custom estimators to further improve the performance, for example, by further augmenting the convolutional network.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

CS had full access to all the data in the study and was responsible for the integrity of the data and the accuracy of the data analysis. CS contributed to concept and design with input from KZ, SZ, YW, and LL and drafted the manuscript with input from all authors. CS, SZ, and KZ contributed to statistical analysis. CS and KZ contributed to funding acquisition, administrative, technical, or material support. All authors contributed to acquisition, analysis, or interpretation of data, critical revision of the manuscript for important intellectual content, and approved the submitted version.

Funding

This study was funded by the National Natural Science Foundation of China (62076093 and 62206095) and the Fundamental Research Funds for the Central Universities (2020MS099, 2020YJ006, and 2022MS078).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agbo-Ajala, O., and Viriri, S. (2021). Deep learning approach for facial age classification: A survey of the state-of-the-art. *Artif. Intell. Rev.* 54, 179–213. doi: 10.1007/s10462-020-09855-0

Agustsson, E., Timofte, R., and Van Gool, L. (2017). "Anchored regression networks applied to age estimation and super resolution," in *Proceedings of the IEEE international conference on computer vision*, Venice, 1643–1652. doi: 10.1109/ICCV.2017.182

Akbari, A., Awais, M., Feng, Z. H., Farooq, A., and Kittler, J. (2020). Distribution cognisant loss for cross-database facial age estimation with sensitivity analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1869–1887. doi: 10.1109/TPAMI.2020.3029486

Angulu, R., Tapamo, J. R., and Adewumi, A. O. (2018). Age estimation via face images: A survey. *EURASIP J. Image Video Process.* 2018, 1–35. doi: 10.1186/s13640-018-0278-6

Bello, I., Zoph, B., Vaswani, A., Shlens, J., and Le, Q. V. (2019). "Attention augmented convolutional networks," in *Proceedings of the IEEE/CVF international conference on computer vision Seoul, Korea (South).* Piscataway, NJ: IEEE, 3286–3295. doi: 10.1109/ICCV.2019.00338

Bourdev, L., Maji, S., and Malik, J. (2011). "Describing people: A poselet-based approach to attribute classification," in *Proceedings of the 2011 international conference on computer vision*, Barcelona, 1543–1550. doi: 10.1109/ICCV.2011.6126413

Bruyer, R., and Scailquin, J. C. (1994). Person recognition and ageing: The cognitive status of addresses-an empirical question. *Int. J. Psychol.* 29, 351–366. doi: 10.1080/00207599408246548

Chang, K. Y., Chen, C. S., and Hung, Y. P. (2011). "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proceedings of the CVPR 2011*, Colorado Springs, CO, 585–592. doi: 10.1109/CVPR.2011.5995437

Chen, J. C., Patel, V. M., and Chellappa, R. (2016). "Unconstrained face verification using deep CNN features," in *Proceedings of the 2016 IEEE winter conference on applications of computer vision (WACV)*, Lake Placid, NY, 1–9. doi: 10.1109/WACV.2016.7477557

Chen, K., Gong, S., Xiang, T., and Change Loy, C. (2013). "Cumulative attribute space for age and crowd density estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Portland, OR, 2467–2474. doi: 10.1109/CVPR.2013.319

Chen, Y., Kalantidis, Y., Li, J., Yan, S., and Feng, J. (2018). "A 2-Nets: double attention networks," in *Proceedings of the 32nd international conference on neural information processing systems*. Montréal: NIPS, 350–359.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 681–685. doi: 10.1109/34.927467

Dagher, I., and Barbara, D. (2021). Facial age estimation using pre-trained CNN and transfer learning. *Multimed. Tools Applic.* 80, 20369–20380. doi: 10.1007/s11042-021-10739-w

Deng, Y., Teng, S., Fei, L., Zhang, W., and Rida, I. (2021). A multifeature learning and fusion network for facial age estimation. *Sensors* 21:4597. doi: 10.3390/s21134597

Dornaika, F., Bekhouche, S. E., and Arganda-Carreras, I. (2020). Robust regression with deep CNNs for facial age estimation: An empirical study. *Exp. Syst. Appl.* 141:112942.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* [Preprint]. arXiv: 2010.11929.

Eidinger, E., Enbar, R., and Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Trans. Inform. Forensics Secur.* 9, 2170–2179. doi: 10.1109/TIFS.2014.2359646

Gao, B. B., Xing, C., Xie, C. W., Wu, J., and Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Trans. Image Process.* 26, 2825–2838. doi: 10.1109/TIP.2017.2689998

Geng, X., Yin, C., and Zhou, Z. H. (2013). Facial age estimation by learning from label distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2401–2412. doi: 10.1109/TPAMI.2013.51

Geronimo, D., Lopez, A. M., Sappa, A. D., and Graf, T. (2009). Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1239–1258. doi: 10.1109/TPAMI.2009.122

Guo, G., and Mu, G. (2011). "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proceedings of the CVPR 2011*, Colorado Springs, CO, 657–664. doi: 10.1109/CVPR.2011.5995404

Guo, G., and Mu, G. (2013). "Joint estimation of age, gender and ethnicity: CCA vs. PLS," in *Proceedings of the 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Shanghai, 1–6. doi: 10.1109/FG.2013.6553737

Guo, G., Fu, Y., Dyer, C. R., and Huang, T. S. (2008). Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Trans. Image Process.* 17, 1178–1188. doi: 10.1109/TIP.2008.924280

Han, H., Jain, A. K., Wang, F., Shan, S., and Chen, X. (2017). Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 2597–2609. doi: 10.1109/TPAMI.2017.2738004

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 770–778. doi: 10.1109/CVPR.2016.90

He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. (2021). "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, Montreal, QC, 15013–15022. doi: 10.1109/ICCV48922.2021.01474

Hou, L., Samaras, D., Kurc, T. M., Gao, Y., and Saltz, J. H. (2016). Neural networks with smooth adaptive activation functions for regression. *arXiv* [Preprint]. arXiv: 1608.06557.

Hu, J., Shen, L., Albanie, S., Sun, G., and Vedaldi, A. (2018). "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proceedings of the 32nd international conference on neural information processing systems*. Montréal: NIPS, 9423–9433.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kwon, Y. H., and da Vitoria Lobo, N. (1999). Age classification from facial images. *Comput. Vision Image Understand.* 74, 1–21. doi: 10.1006/cviu.1997.0549

Levi, G., and Hassner, T. (2015). "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, Boston, MA, 34–42. doi: 10.1109/CVPRW.2015.7301352

Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., et al. (2022). Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv* [Preprint]. arXiv: 2201.09450.

Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., and Tian, Q. (2019). "Bridgenet: A continuity-aware probabilistic network for age estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition Seoul, Korea (South)*. Piscataway, NJ: IEEE, 1145–1154. doi: 10.1109/CVPR.2019.00124

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., et al. (2017). A structured self-attentive sentence embedding. *arXiv* [Preprint]. arXiv: 1703.03130.

Liu, H., Lu, J., Feng, J., and Zhou, J. (2017). Label-sensitive deep metric learning for facial age estimation. *IEEE Trans. Inform. Forensics Secur.* 13, 292–305. doi: 10.1109/TIFS.2017.2746062

Liu, H., Sun, P., Zhang, J., Wu, S., Yu, Z., and Sun, X. (2020). Similarity-aware and variational deep adversarial learning for robust facial age estimation. *IEEE Trans. Multimed.* 22, 1808–1822. doi: 10.1109/TMM.2020.2969793

Liu, N., Zhang, F., and Duan, F. (2020). Facial age estimation using a multi-task network combining classification and regression. *IEEE Access* 8, 92441–92451. doi: 10.1109/ACCESS.2020.2994322

Liu, Y., Kong, A. W. K., and Goh, C. K. (2018). "A constrained deep neural network for ordinal regression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, 831–839.

Liu, Y., Sun, G., Qiu, Y., Zhang, L., Chhatkuli, A., and Van Gool, L. (2021). Transformer in convolutional neural networks. *arXiv* [Preprint]. arXiv: 2106.03180.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision Montreal*. Piscataway, NJ: IEEE , 10012–10022.

Lu, D., Wang, D., Zhang, K., and Zeng, X. (2022). Age estimation from facial images based on Gabor feature fusion and the CIASO-SA algorithm. *CAAI Trans. Intell. Technol.* doi: 10.1049/cit2.12084

Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing* 452, 48–62. doi: 10.1016/j.neucom.2021.03.091

Pan, H., Han, H., Shan, S., and Chen, X. (2018). "Mean-variance loss for deep age estimation from a face," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, 5285–5294. doi: 10.1109/CVPR.2018.00554

Panis, G., Lanitis, A., Tsapatsoulis, N., and Cootes, T. F. (2016). Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics* 5, 37–46.

Pei, W., Dibeklioǒlu, H., Baltrušaitis, T., and Tax, D. M. (2019). Attended end-to-end architecture for age estimation from facial expression videos. *IEEE Trans. Image Process.* 29, 1972–1984. doi: 10.1109/TIP.2019.2948288

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., et al. (2021). "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF international conference on computer vision Montreal*. Piscataway, NJ: IEEE, 367–376. doi: 10.1109/ICCV48922.2021.00042

Ranjan, R., Zhou, S., Cheng Chen, J., Kumar, A., Alavi, A., Patel, V. M., et al. (2015). "Unconstrained age estimation with deep convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision workshops*, Santiago, 109–117.

Ricanek, K., and Tesafaye, T. (2006). "Morph: A longitudinal image database of normal adult age-progression," in *Proceedings of the 7th international conference on automatic face and gesture recognition (FGR06)*, Southampton, 341–345.

Rothe, R., Timofte, R., and Van Gool, L. (2016). "Some like it hot-visual guidance for preference prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition Las Vegas*. Piscataway, NJ: IEEE, 5553–5561. doi: 10.1109/CVPR.2016.599

Rothe, R., Timofte, R., and Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vision* 126, 144–157. doi: 10.1007/s11263-016-0940-3

Sharma, N., Sharma, R., and Jindal, N. (2022). Face-based age and gender estimation using improved convolutional neural network approach. *Wireless Pers. Commun.* 124, 3035–3054. doi: 10.1007/s11277-022-09501-8

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., and Yuille, A. (2019). Deep differentiable random forests for age estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 404–419. doi: 10.1109/TPAMI.2019.2937294

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., and Yuille, A. L. (2018). "Deep regression forests for age estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, 2304–2313. doi: 10.1109/CVPR.2018.00245

Song, Z., Ni, B., Guo, D., Sim, T., and Yan, S. (2011). "Learning universal multi-view age estimator using video context," in *Proceedings of the 2011 international conference on computer vision*, Barcelona, 241–248. doi: 10.1109/ICCV.2011.6126248

Sunitha, G., Geetha, K., Neelakandan, S., Pundir, A. K. S., Hemalatha, S., and Kumar, V. (2022). Intelligent deep learning based ethnicity recognition and classification using facial images. *Image Vision Comput.* 121:104404. doi: 10.1016/j.imavis.2022.104404

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the 31st AAAI conference on artificial intelligence San Francisco, CA*. Menlo Park: AAAI, 4278–4284.

Taheri, S., and Toygar, Ö (2019). On the use of DAG-CNN architecture for age estimation with multi-stage features fusion. *Neurocomputing* 329, 300–310. doi: 10.1016/j.neucom.2018.10.071

Tan, Z., Wan, J., Lei, Z., Zhi, R., Guo, G., and Li, S. Z. (2017). Efficient group-n encoding and decoding for facial age estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 2610–2623. doi: 10.1109/TPAMI.2017.2779808

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st international conference on neural information processing systems*, Long Beach, CA: NIPS, 6000–6010.

Wang, H., Sanchez, V., and Li, C. T. (2022). Improving face-based age estimation with attention-based dynamic patch fusion. *IEEE Trans. Image Process.* 31, 1084–1096. doi: 10.1109/TIP.2021.3139226

Wang, W., Shen, J., Yu, Y., and Ma, K. L. (2016). Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Trans. Visual. Comput. Graph.* 23, 2014–2027. doi: 10.1109/TVCG.2016.2600594

Wang, W., Yao, L., Chen, L., Cai, D., He, X., and Liu, W. (2021). Crossformer: A versatile vision transformer based on cross-scale attention. *arXiv* [Preprint]. arXiv-2108.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition Salt Lake City.* Piscataway, NJ: IEEE, 7794–7803. doi: 10.1109/CVPR.2018.00813

Wang, X., Guo, R., and Kambhamettu, C. (2015). "Deeply-learned feature for age estimation," in *Proceedings of the 2015 IEEE winter conference on applications of computer vision*, Waikoloa, HI, 534–541. doi: 10.1111/1556-4029.13798

Xia, M., Zhang, X., Weng, L., and Xu, Y. (2020). Multi-stage feature constraints learning for age estimation. *IEEE Trans. Inform. Forensics Secur.* 15, 2417–2428.

Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., and Girshick, R. (2021). Early convolutions help transformers see better. *Adv. Neural Inform. Process. Syst.* 34, 30392–30400.

Xing, J., Li, K., Hu, W., Yuan, C., and Ling, H. (2017). Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recogn.* 66, 106–116. doi: 10.1001/jamanetworkopen.2021.11176

Yi, D., Lei, Z., and Li, S. Z. (2014). "Age estimation by multi-scale convolutional network," in *Proceedings of the Asian conference on computer vision*, Cham: Springer, 144–158.

Yi, T. (2022). Estimation of human age by features of face and eyes based on multilevel feature convolutional neural network. *J. Electron. Imaging* 31:041208. doi: 10.1117/1.JEI.31.4.041208

Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., and Wu, W. (2021). "Incorporating convolution designs into visual transformers," in *Proceedings of the IEEE/CVF international conference on computer vision Seattle.* Piscataway, NJ: IEEE, 579–588. doi: 10.1109/ICCV48922.2021.00062

Zeng, X., Huang, J., and Ding, C. (2020). Soft-ranking label encoding for robust facial age estimation. *IEEE Access* 8, 134209–134218.

Zhang, B., and Bao, Y. (2022). Age estimation of faces in videos using head pose estimation and convolutional neural networks. *Sensors* 22:4171. doi: 10.3390/s22114171

Zhang, C., Liu, S., Xu, X., and Zhu, C. (2019). "C3AE: Exploring the limits of compact model for age estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, 12587–12596. doi: 10.1109/CVPR.2019.01287

Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T. X., et al. (2017a). Age group and gender estimation in the wild with deep RoR architecture. *IEEE Access* 5, 22492–22503. doi: 10.1109/ACCESS.2017.2761849

Zhang, K., Sun, M., Han, T. X., Yuan, X., Guo, L., and Liu, T. (2017b). Residual networks of residual networks: Multilevel residual networks. *IEEE Trans. Circuits Syst. Video Technol.* 28, 1303–1314. doi: 10.1109/TCSVT.2017.2654543

Zhang, Y., and Yeung, D. Y. (2010). "Multi-task warped Gaussian process for personalized age estimation," in *Proceedings of the 2010 IEEE computer society conference on computer vision and pattern recognition*, San Francisco, CA, 2622–2629. doi: 10.1109/CVPR.2010.5539975

Zhao, Q., Dong, J., Yu, H., and Chen, S. (2020). Distilling ordinal relation and dark knowledge for facial age estimation. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 3108–3121. doi: 10.1109/TNNLS.2020.3009523