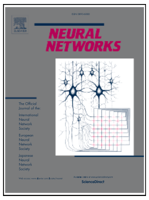
Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

CILF-CIAE: CLIP-driven image-language fusion for correcting inverse age estimation

Yuntao Shou [a], Tao Meng [a,*], Wei Ai [a], Nan Yin [b], Keqin Li [c]

[a] College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, Hunan, 410004, China
[b] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, 44737, UAE
[c] Department of Computer Science, State University of New York, New Paltz, New York, 12561, USA

ARTICLE INFO

Keywords:
Age estimation
CLIP
Transformer
Image-language fusion
Fourier transform
Error correction

ABSTRACT

The age estimation task aims to predict the age of an individual by analyzing facial features in an image. The development of age estimation can improve the efficiency and accuracy of various applications (e.g., age verification and secure access control, etc.). In recent years, contrastive language-image pre-training (CLIP) has been widely used in various multimodal tasks and has made some progress in the field of computer vision. However, the promotion of CLIP and error feedback mechanisms for age estimation has not been investigated, and existing Transformer-based methods require high memory usage (quadratic complexity) when globally modeling images. To tackle the above issues, we propose a novel CLIP-driven Image-Language Fusion for Correcting Inverse Age Estimation (CILF-CIAE). Specifically, we first introduce the CLIP model to extract image features and text semantic information respectively, and map them into a highly semantically aligned high-dimensional feature space. Next, we designed a new Transformer architecture (i.e., FourierFormer) to achieve channel evolution and spatial interaction of images, and to fuse image and text semantic information. Compared with the quadratic complexity of the attention mechanism, the proposed FourierFormer is of linear log complexity. To further narrow the semantic gap between image and text features, we utilize an efficient contrastive multimodal learning module that supervises the multimodal fusion process of FourierFormer through contrastive loss for image-text matching, thereby improving the interaction effect between different modalities. Finally, we introduce reversible age estimation, which uses end-to-end error feedback to reduce the error rate of age predictions. Extensive experiments on six benchmark datasets demonstrate that CILF-CIAE consistently outperforms advanced methods such as LRA-GNN and MCGRL. For example, our method achieves an MAE of 1.68 on MORPH-S2, significantly lower than 2.21 (LRA-GNN) and 1.77 (MCGRL), highlighting its superior accuracy and robustness in real-world age estimation scenarios.

1. Introduction

1.1. Motivation

The task of age estimation aims to determine the age based on the facial features in the image. In recent years, due to the massive increase in image data sets and the widespread application of deep learning (DL), age estimation methods have also achieved important achievements and attracted widespread research attention (Shen et al., 2019), (Liu et al., 2020), (Yin et al., 2023b). Furthermore, age estimation is also widely used in many scenarios. For example, age estimation in finance and insurance can help detect fraud where age is falsely stated to obtain improper benefits (Rothe et al., 2018), (Bao et al., 2023), (Yin et al., n.d.).

The current mainstream age estimation methods are divided into three categories: CNN (Niu et al., 2016), (Duan et al., 2017), attention network (Wang et al., 2022), (Zhang et al., 2019), and GCN (Shou et al., 2023). To extract global information and multi-scale information in images, a CNN-based age estimation algorithm is applied. For example, Rothe et al. (2018) estimated an individual's true age and apparent age from a single face image based on a CNN method. Unlike many traditional machine learning methods (Cao et al., 2012), this method does not require the use of facial feature point markers and only requires the input of face images for age estimation. However, CNN-based methods cannot capture the semantic features in images that are most relevant to age features. To give higher weight to the semantic features in the image that are most relevant to the age feature, attention networks began to be

* Corresponding author.
E-mail addresses: shouyuntao@stu.xjtu.edu.cn (Y. Shou), mengtao@hnu.edu.cn (T. Meng), aiwei@hnu.edu.cn (W. Ai), nan.yin@mbzuai.ac.ae (N. Yin), lik@newpaltz.edu (K. Li).

https://doi.org/10.1016/j.neunet.2025.108518
Received 9 October 2024; Received in revised form 16 October 2025; Accepted 21 December 2025
Available online 26 December 2025

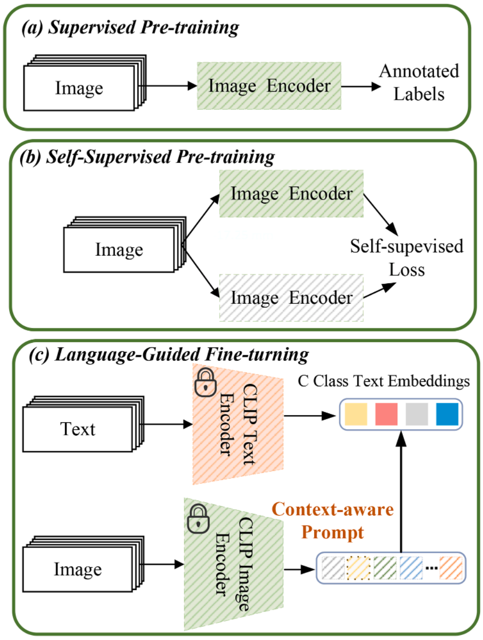


Fig. 1. We compare the differences between existing image learning paradigms and the paradigm proposed in this paper. As shown in Fig. 1(a), most image learning methods perform supervised learning by inputting images and then using manually annotated labels as supervision signals. As shown in Fig. 1(b), since manual annotation requires a large amount of resources, existing methods begin to build self-supervised learning models by contrasting input images. As shown in Fig. 1(c), we perform text-image contrastive learning by using the CLIP pre-trained model and transfer the learned knowledge to the age estimation prediction task.

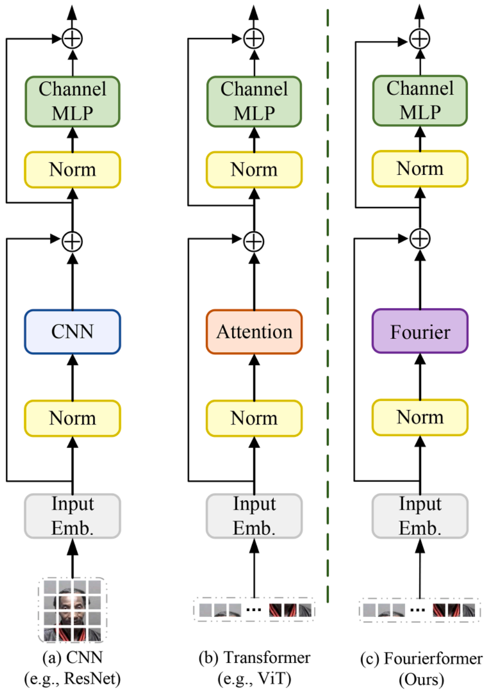


Fig. 2. We compare the differences between existing image architectures and the architectures proposed in this paper. As shown in Fig. 2(a) and (b), existing methods are mainly based on CNN architecture and Transformer architecture based on attention mechanism to extract feature information of images. As shown in Fig. 2(c), we replace the attention module in the Transformer architecture with a Fourier prior module.

applied. For instance, Shen et al. (2022) introduced an attention mechanism so that the model can automatically focus on regions in the image that are relevant for age estimation, which helps improve the model's perception of important features related to age. In addition to the attention structure for age estimation tasks, many task-specific Transformer variants have been proposed to address challenges in specific domains. For instance, the Top-k Token Selective Transformer (Xiao et al., 2024b) introduces a token selection strategy that retains only the most informative patches for remote sensing image super-resolution, effectively reducing computational overhead while preserving global context. Similarly, the Medical Transformer (Valanarasu et al., 2021) employs gated axial-attention to enhance spatial dependency modeling in medical image segmentation, showcasing the adaptability of Transformer-based designs in complex structural domains. However, attention network-based methods cannot flexibly model irregular objects. To overcome the above problems, Shou et al. (2023) proposed a contrastive multi-view GCN for age estimation (CMGCN). CMGCN improves the feature representation capabilities of images by extending image representation into topological semantic space. However, the methods mentioned above are all supervised learning methods and ignore the CLIP-based multimodal learning paradigm. Taking Fig. 1(a) and (b) as an example, existing age estimation algorithms mainly focus on supervised, or self-supervised algorithm design (Bao et al., 2022), (Deng et al., 2021), ignoring the contrastive image-language pre-training (CLIP) paradigm. CLIP can learn the prior information of faces from a large number of text-image pairs and provide better generalization for downstream tasks. Specifically, CLIP learns the correlation between images and text from a large number of image-text pairs through contrastive learning. Furthermore, existing algorithms directly predict age and lack an error information feedback mechanism, which may lead to a large error between the model's predicted age and the true label. Therefore, it is necessary to take CLIP multimodal learning paradigm and error-controllable generation as the starting point for model design.

To tackle the above problem, we propose a novel CLIP-driven Image-Language Fusion for Correcting Inverse Age Estimation (CILF-CIAE) to perform age estimation. CILF-CIAE mainly includes four modules: CLIP-based visual and language feature encoder, FourierFormer-based feature fusion, age prediction and error-controllable generation module. Firstly, we use Image Encoder and Text Encoder in CLIP to encode image and text features respectively and obtain corresponding feature representations. After obtaining the image and text feature representations, we jointly input them into the N-dimensional feature space for contrastive learning to obtain aligned text and image semantic vectors, and utilize obtained image semantic vectors to perform age estimation. Secondly, as shown in Fig. 2(a) and (b), unlike previous CNN-based and attention-based Transformer architectures, CNN-based methods can only extract local information of the image and it is difficult to use contextual prompts modules to enhance age estimation, while attention-based methods require large memory usage (quadratic complexity). We introduce the Transformer architecture based on Fourier transform to realize the spatial interaction and channel evolution of image features, so as to fuse text and image feature information to improve the age estimation performance. Specifically, we replace the attention module in Transformer with Fourier transform and input image features into FourierFormer to achieve spatial interaction and channel evolution. To further narrow the semantic gap between image and text features, we utilize an efficient contrastive multimodal learning module that supervises the multimodal fusion process of FourierFormer through contrastive loss for image-text matching, thereby improving the interaction effect between different modalities. Thirdly, we construct age estimation prediction loss and text and image matching loss to complete the parameter optimization of the model. Finally, we build an error-correcting reversible age estimation module to ensure that the predicted age is within a high-confidence interval in an end-to-end learning manner.

1.2. Our contributions

Therefore, CLIP multimodal learning, spatial interaction of images, and channel evolution should be the core of age estimation algorithm design. Inspired by the above analysis, we propose a novel CLIP-driven Image-Language Fusion for Correcting Inverse Age Estimation (CILF-CIAE) to perform age estimation. The main contributions of this paper are summarized as follows:

1. We propose a novel CLIP-driven Image-Language Fusion framework (CILF-CIAE) tailored for age estimation, which goes beyond simple CLIP fine-tuning by integrating a vision-guided semantic alignment pipeline and a dedicated correction mechanism.
2. We design a new lightweight Transformer variant called FourierFormer, which replaces the self-attention mechanism with a learnable frequency-based spatial and channel interaction module. Unlike FNet or frequency-assisted Mamba, our design is optimized for image-language fusion and incorporates nonlinear filtering and residual pathways to enhance representational expressiveness.
3. We introduce a contrastive multimodal learning module with context-aware prompt enhancement, which strengthens image-text alignment through Fourier-enhanced visual context. This differs from prior CLIP-based works (e.g., CoOp, CoCoOp) by utilizing dynamic, vision-driven text guidance.
4. We develop an end-to-end reversible error feedback mechanism, which combines explicit and implicit error modeling using an ensemble of lightweight regressors. Unlike standard post-processing methods, our mechanism is integrated into the training loop and iteratively refines predictions until the estimated error falls below a learned threshold.

2. Related work

2.1. Age estimation

Traditional age estimation methods usually rely on hand-designed feature extraction and machine learning algorithms, which are limited by feature selection and age estimation performance (Cao et al., 2012), (Yin et al., 2023a), (Yin et al., 2022). With the popularity of the Internet and social media (e.g., meta, twitter, and Youtube, etc.), large-scale face image datasets have also been widely grown. The rapid growth of data sets provides rich training data for deep learning (DL), making DL's learning capabilities more powerful. Age estimation has potential applications in social media analysis, ad targeting, security monitoring, medical image analysis, etc. For example, in security and legal applications, image age estimation can assist police in identifying possible underage criminal suspects.

Existing age estimation algorithms are mainly divided into two categories, i.e., age estimation algorithms based on machine learning and algorithms based on deep learning. Machine learning-based age estimation algorithms mainly rely on hand-designed rules to extract age-related features of images. Age estimation algorithms based on deep learning mainly use some deep learning models (e.g., CNN, Transformer, and GCN, etc) with powerful adaptive learning capabilities and massive data sets to estimate age in an end-to-end manner.

Machine learning methods: In the age estimation algorithms based on traditional machine learning algorithms, Shin et al. (2022) proposed an ordinal regression algorithm (MWR) based on moving window regression, which first ranks the input and reference labels and designs global and local regressors to achieve prediction of global ranking and local ranking. MWR achieves fine-grained age estimation by continuously iteratively optimizing the ranking order. However, the computational complexity of MWR is relatively high. Cao et al. (2020) proposed a consistent ranking logic algorithm to solve the inconsistency problem of multiple binary ordinal regression algorithms. CORAL ensures ranking consistency by introducing confidence scores. Cao et al. (2012) proposed the Ranking SVM algorithm to achieve age estimation of images. This algorithm estimates age by first grouping ages and then sorting ages. RSVM can reduce the hypothesis space of model learning. Zhang et al. (2017) achieved age estimation by learning the probability distribution of label information. This algorithm achieves age prediction by calculating the posterior probability of the image. There are some other typical traditional machine learning algorithms (Li et al., 2019), (Shen et al., 2018).

Deep learning methods: In the age estimation algorithms based on deep learning algorithms, CNN (Levi & Hassner, 2015), attention network (Wang et al., 2022), and hybrid neural network systems (Xie et al., 2015) are currently common age estimation algorithms. For example, Levi and Hassner (2015) proposed an age estimation algorithm based on deep CNN to solve the problem of insufficient performance of traditional machine learning algorithms. DeepCNN can achieve better prediction results even on a small amount of data sets. Duan et al. (2017) proposed the CNN2ELM algorithm to combine the advantages of CNN and regression algorithms. CNN2ELM constructed three feature extraction networks of age, gender, and race, and used a fusion mechanism to fuse the complementary information of the three networks, and used ELM for regression prediction of age. Wang et al. (2022) proposed the Attention-based Dynamic Patch Fusion algorithm to solve the problem that CNN cannot extract the most beneficial semantic information in the image for the age estimation task. ADPF introduces attention network and fusion network to dynamically extract image patches with rich semantic features and adaptively fuse the extracted feature information. Zhang et al. (2019) proposed a fine-grained attention LSTM algorithm to solve the problem that existing methods only focus on the global information of the image and ignore the fine-grained features of the image. This method first uses the residual network to extract the global information of the image, and then uses the attention LSTM to capture the sensitive area information of the image to obtain local important semantic features in the image. Xie et al. (2015) integrated CNN's feature extraction capabilities, domain generalization capabilities, and local information discrimination capabilities based on dictionary algorithms. This method first uses a pre-trained CNN to extract the feature representation of the image, and then builds a dictionary representation to extract the local feature information and Fisher vector representation of the image.

2.2. Contrastive image-language pre-training

The recent success of contrastive vision-language pre-training (e.g., CLIP (Lee et al., 2022)) has paved the way for a new generation of models that learn joint image-text representations by aligning large-scale image-text pairs. These models have achieved remarkable performance on downstream tasks such as zero-shot classification, image retrieval, and captioning. Building on CLIP, many extensions have been proposed to enhance its adaptability. CoOp (Zhou et al., 2022b) and CoCoOp (Zhou et al., 2022a) use learnable and conditional prompts to fine-tune CLIP on new categories with limited supervision. DenseCLIP (Rao et al., 2022) improves regional alignment by integrating patch-wise visual features. Flamingo (Alayrac et al., 2022) combines frozen visual backbones with pretrained LLMs to support multimodal reasoning. Beyond CLIP-style models, BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) introduce bootstrapped training pipelines that unify image-text understanding and generation, achieving strong performance in both captioning and VQA tasks. Similarly, MiniGPT-4 (Zhu et al., 2024), LLaVA (Liu et al., 2023), and mPLUG-Owl (Ye et al., 2023) bridge vision encoders with large language models for instruction-following and multimodal dialogue. However, these models primarily focus on open-ended generation, not continuous regression, and they typically require large-scale GPU resources for training and inference. On the other hand, SigLIP (Zhai et al., 2023) introduces a sigmoid-based contrastive loss to replace softmax, enabling more stable multi-label training. GRILL (Jin et al., 2023) focuses on

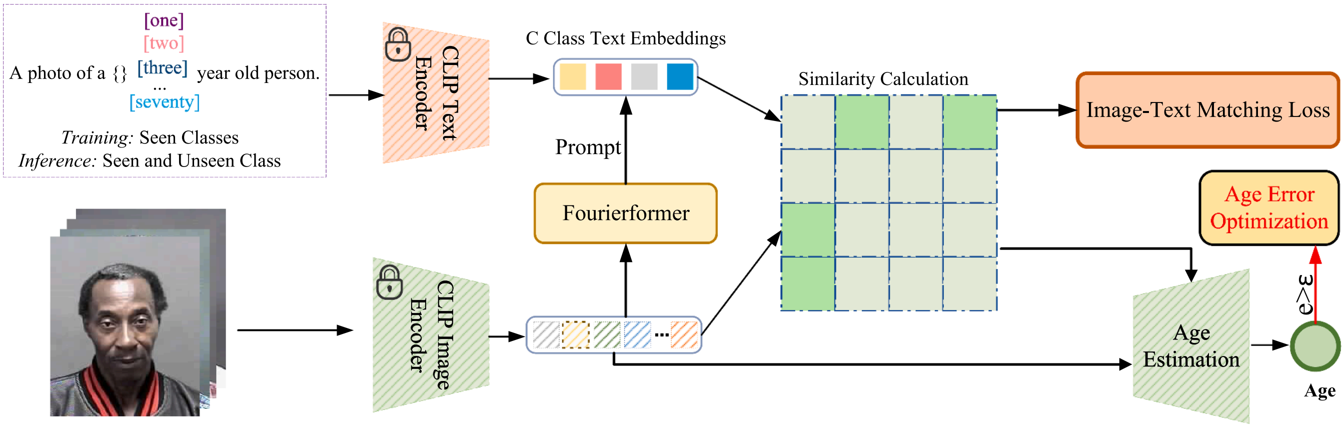


Fig. 3. The overall framework for age prediction using CILF-CIAE. Specifically, we first use CLIP to extract image features and C-type text features, and then calculate the pixel-text similarity score. The similarity scores of the pixel-text pairs are fed into the age estimation module, and the age label is used as a supervision signal. To better utilize the prior knowledge of images, we introduce FourierFormer to extract contextual information in images to prompt the language model. Finally, we perform error optimization on the predicted age.

region-phrase alignment, improving grounding accuracy in fine-grained image-language matching.

Despite these advancements, most existing vision-language methods lack components for regression refinement, feedback correction, or frequency-based context modeling. In contrast, our proposed method introduces a lightweight Fourier-enhanced Transformer (FourierFormer), a vision-guided prompt module, and a two-stage error feedback mechanism-making it the first to bring frequency-domain priors and end-to-end correction into multimodal age estimation.

2.3. Frequency-domain representation and fourier transformers

Recent advances in vision models have increasingly explored frequency-domain representations to enhance global context modeling and structural sensitivity, especially for high-resolution or multi-channel imagery. Compared to conventional spatial-domain convolutions or attention mechanisms, Fourier transforms enable compact and efficient global operations by projecting inputs into the frequency spectrum, where both spatial structure and semantic patterns can be more easily disentangled. For example, the Frequency-assisted Mamba network (Xiao et al., 2024a) applies Fourier encoding to enhance Mamba's temporal aggregation capability in remote sensing image super-resolution, revealing the potential of frequency priors for improving detail preservation. In hyperspectral video understanding, ProFiT (Chen et al., 2025a) leverages prompt-guided frequency-aware filtering to enhance template-matching through global signal manipulation. Similarly, SST-track and Spectral-Spatial Fusion with Memory Enhancement (Chen et al., 2025b) propose spatiotemporal fusion pipelines in the frequency domain to improve tracking robustness in dynamic hyperspectral sequences. These works demonstrate that Fourier-domain modeling is increasingly effective for domains requiring high-dimensional, multi-modal, or globally-aware representations. However, most of these models are not designed for image-language fusion or continuous regression tasks, and often rely on fixed or non-learnable transforms. In contrast, our proposed FourierFormer module adopts a learnable frequency modeling design, introducing both spatial-frequency interaction and channel evolution, combined with residual pathways and contrastive supervision for semantic alignment. This makes it particularly suitable for multi-modal alignment tasks such as age estimation, where preserving both visual granularity and global semantic alignment is crucial.

3. Methodology

3.1. The design of the CILF-CIAE structure

The CILF-CIAE architecture proposed in this paper is shown in Fig. 3, which contains age prediction stages and age error optimization. Specifically, we first use age estimation models based CLIP with a Fourier prior module to predict the age of images. To further narrow the semantic gap between image and text features, we utilize an efficient contrastive multimodal learning module that supervises the multimodal fusion process of FourierFormer through contrastive loss for image-text matching, thereby improving the interaction effect between different modalities. Furthermore, if the predicted and actual values exceed a given threshold, the optimization branch is activated. The age errors are then used in the training of an ensemble error correction model to update the predicted age x^*. This training process continues until $e(x^*) \le \epsilon$ terminates. The details of the CILF-CIAE architecture proposed in this paper will be described.

3.1.1. Language-guided visual age prediction

As shown in Fig. 3, we briefly introduce the CLIP-based visual language pre-training model for age estimation. CLIP consists of an image encoder and a text encoder (Zhang et al., 2022b). Image encoders aim to extract the underlying features of an image and map them into a low-dimensional embedding space. The architecture of image encoders usually uses ViT (Han et al., 2022) with superior performance. The text encoder often use Transformers (Khan et al., 2022) to generate text representations with rich semantic information. Given a text prompt, such as “A photo of a 12 year old person," the text encoder first converts each character into a lowercase byte-pair encoded representation, which uniquely identifies each character. The beginning and end of each text sequence are marked by [SOS] and [EOS]. Afterwards, the text representation is mapped into a 512-dimensional feature space, and then text Transformer is used for sequence modeling. Then, given an image feature obtained by the image encoder, the cosine similarity function is used to calculate the similarity between the image and the text prompt. The similarity formula is defined as follows:

$$\mathbf{S} = \frac{\exp(sim(T_i, I_i)/\tau)}{\sum_{j=1}^{N} \exp(sim(T_j, I_i)/\tau)} \quad (1)$$

where $\mathbf{S}$ is the similarity matrix, T_i is the feature vector of the ith text sequence obtained by the text encoder, I_i is the feature vector of the ith image obtained by the image encoder, N represents the total number of training samples, $sim(\cdot)$ represents cosine similarity, and τ represents temperature attenuation coefficient.

To further narrow the semantic gap between image and text features, we design an efficient contrastive multimodal learning module to supervise the fusion process of FourierFormer. Specifically, we introduce an image-text contrastive loss, denoted as $\mathcal{L}_{\text{text-image}}$, which encourages matched image-text pairs to be close in the embedding space and non-matching pairs to be distant. Formally, let T_i and I_i represent the ith text and image embeddings respectively, both normalized and extracted by the CLIP encoders. We define the cosine similarity between them as $\text{sim}(T_i, I_j)$, and use a temperature parameter τ to control sharpness. The loss $\mathcal{L}_{\text{text-image}}$ is composed of two symmetric components: (1) image-to-text matching and (2) text-to-image matching. The final contrastive loss is computed as:

$$\begin{aligned}\mathcal{L}_{\text{text-image}} = &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(sim(T_i, I_i)/\tau)}{\sum_{j=1}^{N}\exp(sim(T_i, I_j)/\tau)} \\ &+ \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\log\frac{\exp(S(T_i, I_j)/\tau)}{\sum_{k=1}^{N}\exp(S(T_i, I_k)/\tau)} \end{aligned} \tag{2}$$

where N is the number of the training samples.

3.1.2. Context-aware prompting

Previous work has demonstrated that feature alignment of visual and language modalities can significantly improve the performance of CLIP models on downstream tasks (Zhang et al., 2022a), (Zhou et al., 2022a). Therefore, we consider whether we can design a customized context-aware prompting method to improve text features.

Vision-to-language prompting. The textual features that fuse visual global context information can make age estimation predictions more accurate. For example, “a photo of a 68-year-old man with gray hair" is a more accurate prediction than “a photo of a 68-year-old man." Therefore, we design a customized Fourier prior module to utilize visual global context information to improve text features in fine granularity. Specifically, we use the FourierFormer decoder to realize image spatial information interaction and channel evolution, and model the interaction between vision and language.

3.1.3. Fourier prior embedded block

In contrast to traditional self-attention mechanisms used in Transformer architectures, which enable powerful global feature interaction but incur quadratic computational complexity with respect to the input sequence length, we adopt a more efficient frequency-domain modeling approach using the discrete Fourier transform (DFT) (Zhou et al., 2023). Self-attention computes pairwise similarity between all token pairs, which becomes computationally intensive for high-resolution visual inputs. By transforming image features into the frequency domain, the Fourier transform provides a global receptive field with significantly lower complexity, typically linear or log-linear, depending on implementation. This property allows the model to retain essential structural and spatial information while being more scalable. In multimodal settings, such as age estimation from image-text pairs, this efficiency is particularly beneficial for fusing visual and linguistic semantics without incurring the memory overhead of attention-based fusion. In our work, we apply the Fourier transform not directly to raw images but to the intermediate visual feature maps extracted by the image encoder (e.g., CLIP). This enables the model to capture global spatial patterns within high-level semantic features in a computationally efficient way. Formally, let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ denote the input feature map produced by the image encoder, where H and W are spatial dimensions and C is the number of channels. The Fourier transform is applied channel-wise to convert each 2D feature map into the frequency domain. The resulting complex-valued representation is denoted by $\mathcal{F}(\mathbf{X})$. The transformation for a single channel is defined as:

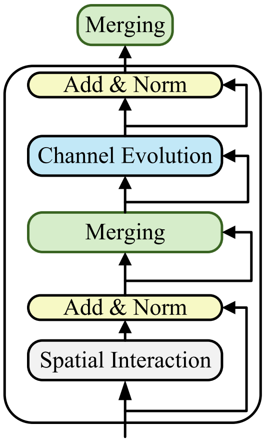


Fig. 4. The overall framework of the proposed FourierFormer. FourierFormer includes a spatial interaction module, a channel evolution module, a discrete Fourier transform (DFT) and an inverse discrete Fourier (IDFT) module, which can effectively extract information from the global context of an image.

$$\mathcal{F}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v}) = \frac{1}{\sqrt{\mathrm{H}\times\mathrm{W}}}\sum_{\boldsymbol{h}=0}^{\mathrm{H}-1}\sum_{\boldsymbol{w}=0}^{\mathrm{W}-1}\boldsymbol{x}(\boldsymbol{h},\boldsymbol{w})e^{-j2\pi(\frac{h}{\mathrm{H}}u+\frac{\boldsymbol{w}}{\mathrm{W}}\boldsymbol{v})} \tag{3}$$

where u and v represent the horizontal and vertical coordinates of the Fourier domain. The phase component $P(x)(u,v)$ and the amplitude component $A(x)(u,v)$ are obtained as follows:

$$\begin{aligned}&\mathcal{A}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v})) = \sqrt{\mathcal{R}^2(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v})) + \mathcal{I}^2(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v}))}, \\ &\mathcal{P}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v})) = \arctan\left[\frac{\mathcal{I}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v}))}{\mathcal{R}(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v}))}\right],\end{aligned} \tag{4}$$

where $I(x)(u,v)$ and $R(x)(u,v)$ represent imaginary numbers and real numbers, respectively. These components are further processed through spatial interaction and channel evolution modules inside the FourierFormer block to enhance cross-channel contextual modeling.

Structure Flow. The main goal of designing the Fourier prior module in this paper is to achieve an effective and efficient global context image information modeling paradigm and improve the representation ability of text features, as shown in Fig. 4. For a given image $x \in \mathbb{R}^{H\times W\times C_{in}}$, we first use a text encoder based CLIP to extract the shallow features of the image $X_0 \in \mathbb{R}^{H\times W\times C}$. Shallow features are encoded by using N stacked image encoders. The Fuoriformer module designed in this paper consists of a stack of spatial interaction module, channel evolution module, residual and layer normalization module and Fourier prior module. Similarly, for the image decoder, we use a stack of the proposed core modules for image feature decoding.

As shown in Fig. 5, the core module of FourierFormer consists of two parts: spatial interaction and channel evolution, which are implemented by depth convolution and 1×1 convolution with DFT and IDFT respectively.

Fourier Spatial Interaction. Fourier spatial interaction first takes the image feature maps obtained by the image encoder as the input of FourierFormer, and then applies DFT to convert them into a spatial feature representation. Assuming that the features are expressed as $X \in R^{H\times W\times C}$, the corresponding DFT formula is defined as:

$$\mathbf{X}_{\mathcal{I}}^{(c)}, \mathbf{X}_{\mathcal{R}}^{(c)} = \mathcal{F}(\mathbf{X}^{(c)}) \tag{5}$$

where $c = 1, \ldots, C$, $\mathbf{X}_{\mathcal{I}}$ and $\mathbf{X}_{\mathcal{R}}$ represent the real and imaginary parts in the Fourier space. We then perform Fourier spatial interaction to filter and compress the frequency domain signal of the image through a deep-wise convolution (DWconv) operation with LeakyReLU activation function. The spatial interaction process of images can be defined as:

$$\mathbf{S}_{\mathcal{I}}^{(b)} = LeakyReLU\left(DWconv^{(\boldsymbol{b})}(\mathbf{X}_{\mathcal{I}}^{(\boldsymbol{b})})\right)$$

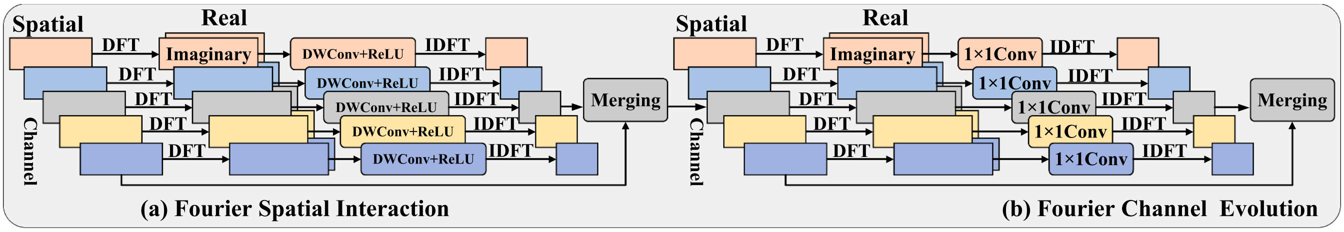


Fig. 5. Details of the Fourier Prior Embedding module (FPE). FPE follows the global context information modeling idea of spatial interaction and channel evolution.

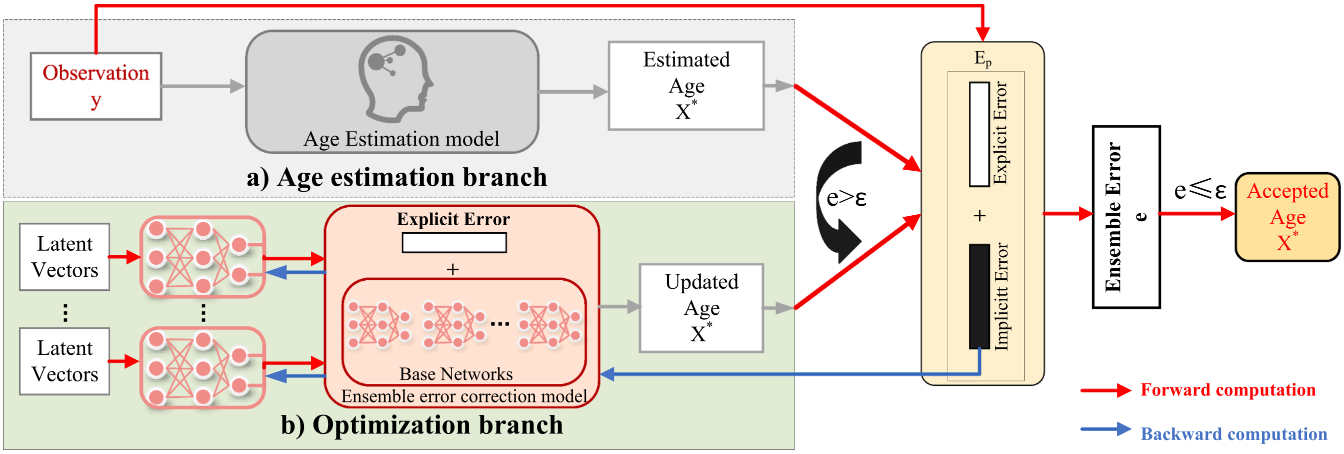


Fig. 6. The flowchart of the correcting inverse age estimation. Existing age estimation models give a first age estimate, which is assessed by evaluations E_P. If failed, the optimization branch will be activated. The age estimation error estimated by the ensemble error model is used for training to update the predicted age x^*. The process terminates until $e(x^*) \le \epsilon$.

$$\mathbf{S}_{\mathcal{R}}^{(\boldsymbol{b})} = LeakyReLU\left(DWconv^{(\boldsymbol{b})}(\mathbf{X}_{\mathcal{R}}^{(\boldsymbol{b})})\right) \tag{6}$$

Then we apply inverse DFT to the learned $\mathbf{S}_{\mathcal{I}}$ and $\mathbf{S}_{\mathcal{R}}$ with low-frequency signals to transform them back into the spatial domain. The formula for $\mathbf{S}_{\mathcal{I}}$ and $\mathbf{S}_{\mathcal{R}}$ to achieve time-frequency conversion is defined as follows:

$$\mathbf{X}_{\mathbf{S}}^{\mathbf{b}} = \mathcal{F}^{-1}(\mathbf{S}_{\mathcal{I}}^{(\boldsymbol{b})}, \mathbf{S}_{\mathcal{R}}^{(\boldsymbol{b})}) \tag{7}$$

The spectral convolution theorem in Fourier theory states that the convolution operation of signals in the frequency domain is equivalent to their product operation in the time domain, which reveals the overall frequency composition. The spectral convolution theorem provides an efficient way to process signals in the frequency domain because convolution operations in the frequency domain are generally easier to process than multiplication operations in the time domain. Therefore, we concatenate the $\mathbf{X}_{\mathbf{S}}^{\mathbf{b}}$ obtained by Fourier transform and normalize it to obtain the output S_X of the Fourier spatial interaction.

Fourier Channel Evolution. Fourier channel evolution performs channel-by-channel evolution by applying a 1×1 convolution operator to decompose the output S_X of the Fourier space interaction into real and imaginary parts $\mathbf{C}_{\mathcal{I}}$ and $\mathbf{C}_{\mathcal{R}}$. The Fourier channel evolution formula can be defined as:

$$\begin{aligned} \mathbf{CX}_{\mathcal{I}} &= LeakyReLU\left(\mathbf{conv}\left(cat[\mathbf{C}_{\mathcal{I}}^{1}, \dots, \mathbf{C}_{\mathcal{I}}^{c}]\right)\right) \\ \mathbf{CX}_{\mathcal{R}} &= LeakyReLU\left(\mathbf{conv}\left(cat[\mathbf{C}_{\mathcal{R}}^{1}, \dots, \mathbf{C}_{\mathcal{R}}^{c}]\right)\right) \end{aligned} \tag{8}$$

where $cat(\cdot)$ is the concatenation operation. Then we perform IDFT to convert $\mathbf{CX}_{\mathcal{R}}$ and $\mathbf{CX}_{\mathcal{I}}$ to time domain space as follows:

$$\mathbf{C}_{\mathbf{S}}^{\mathbf{b}} = \mathcal{F}^{-1}(\mathbf{CX}_{\mathcal{I}}^{(\boldsymbol{b})}, \mathbf{CX}_{\mathcal{R}}^{(\boldsymbol{b})}) \tag{9}$$

3.1.4. *Two-stage error selection*

To enhance prediction reliability, we design a two-stage error correction mechanism that refines the age estimation results based on the magnitude and nature of prediction errors. In the first stage, we compute the explicit error, which is the directly observable difference between the initial age prediction and the ground-truth age label. If this error exceeds a threshold ϵ, adaptively set based on the validation set MAE, we consider the prediction unreliable and activate the second-stage correction process. This threshold-based decision acts as a lightweight filter to determine whether further refinement is necessary. In the second stage, we introduce an ensemble error correction module that estimates the implicit error-that is, the model's internal estimate of residual error without reference to ground truth. This module operates on the fused multimodal features produced by the FourierFormer encoder and consists of an ensemble of five lightweight regressors. Each regressor is implemented as a two-layer MLP with independent parameters but identical architecture, encouraging diversity in correction strategies while maintaining computational efficiency. To integrate the ensemble outputs, we adopt a learnable voting mechanism. Each regressor contributes to the final correction through a weight that is learned during training. These weights are normalized to ensure stability and allow the model to automatically prioritize more reliable regressors. The refined age prediction is then obtained by applying the aggregated correction to the original prediction. The entire correction process is fully differentiable and end-to-end trainable. It operates during both training and inference, requiring no additional post-processing. This unified design ensures that the model not only predicts age but also self-corrects when confidence is low, improving robustness in the presence of noise or out-of-distribution samples.

As shown in Fig. 6, we first use a CLIP-based learning model to predict age. If the error exceeds the threshold, an optimization branch is used to optimize the error and give a predicted age with high confidence.

For a given observation y, we use multiple models and metrics to evaluate the predicted age, resulting in an h-dimensional error vectors, expressed as:

$$\mathbf{e}(\mathbf{x}, \mathbf{y}) = \left[E_1(\mathbf{x}, \mathbf{y}), E_2(\mathbf{x}, \mathbf{y}), \dots, E_h(\mathbf{x}, \mathbf{y})\right] \tag{10}$$

where $E_i(,)$ represents the error estimate calculated by the ith model, x is the input image.

Each associated age estimate obtained from an observation y follows the i.i.d. criterion, so y is treated as a constant. Therefore, we can simplify Eq. (10) and obtain optimal model parameters by minimizing the error $e(x)$:

$$\min_{\mathbf{x}\in\mathcal{X}} e(\mathbf{x}) = \sum_{i=1}^{h} w_i E_i(\mathbf{x}) \tag{11}$$

where w_i is determined using a voting mechanism, which is learnable.

Leveraging ensemble learning (Kang et al., 2023) enables a more robust representation of the hypothesis space, we integrate multiple neural networks to estimate implicit errors. Each neural network uses a mapping function $\phi(x, w)$, $\mathcal{R}^D \times \mathcal{R}^{|\mathbf{w}|} \to \mathcal{R}^k$ for error. We train L regressors with the same network architecture and use a voting algorithm to obtain the final prediction. Therefore, for a given input state x, the implicit error $\hat{\mathbf{e}}$ is estimated by the ensemble network as follows:

$$\hat{\mathbf{e}}\left(\mathbf{x}, \{\mathbf{w}_i\}_{i=1}^{L}\right) = \frac{1}{L} \sum_{i=1}^{L} \boldsymbol{\phi}(\mathbf{x}, \mathbf{w}_i) \tag{12}$$

where $\mathbf{w}_i$ is the learnable network parameters.

According to Eq. (12), we can obtain the cumulative age estimation error as follows:

$$\hat{e}\left(\mathbf{x}, \{\mathbf{w}_i\}_{i=1}^{L}\right) = \underbrace{\sum_{j=1}^{k} w_j \left(\frac{1}{L} \sum_{i=1}^{L} \boldsymbol{\phi}_j(\mathbf{x}, \mathbf{w}_i) \right)}_{\text{approximated implicit error}} \tag{13}$$

$$+ \underbrace{\sum_{j=k+1}^{h} w_j E_j(\mathbf{x})}_{\text{true explicit error}} \tag{14}$$

We divide the error of Eq. (14) into two parts, one is the estimated implicit error, and the other is the true explicit error. The estimated implicit error is obtained by learning the feature representation of the image encoder by the ensemble regressor we built, and the real explicit error is obtained by the age estimation model based on CLIP we built. At the same time, we optimize the network parameters of the ensemble regressor by minimizing the distance between the estimated implicit error and the true explicit error. The optimization goal is defined as follows:

$$\min_{\mathbf{w}_i} \mathbb{E}_{(\mathbf{x},\mathbf{e})\sim D}\left[\mathrm{dist}\left(\boldsymbol{\phi}(\mathbf{x}, \mathbf{w}_i), \mathbf{e}_{1:k}\right)\right] \tag{15}$$

where $\mathrm{dist}\left(\boldsymbol{\phi}(\mathbf{x}, \mathbf{w}_i), \mathbf{e}_{1:k}\right) = ||\boldsymbol{\phi}(\mathbf{x}, \mathbf{w}_i), \mathbf{e}_{1:k}||_2^2$

To achieve controllable generation of predicted states, we use the feature representation decoded by the image encoder as the input of the ensemble regressor to learn and sample candidate predicted ages. Therefore, the update target of network parameters is defined as follows:

$$\boldsymbol{\theta}^{(t)} = \arg \min_{\theta\in\mathcal{R}^d} \mathbb{E}_{\mathbf{z}}\left[\hat{e}\left((\mathbf{z}, \boldsymbol{\theta}), \left\{\mathbf{w}_i^{(t-1)}\right\}_{i=1}^{L}\right)\right] \tag{16}$$

where z is the latent vectors. Finally, among the candidate age estimation states generated by the ensemble regressor with the trained network parameters θ^t, we select the final prediction result as follows:

$$\mathbf{x}_{\Pi}^{(t)} = \arg \min_{\mathbf{x}\sim p\left(\mathbf{x}|\theta^{(t)}\right)} \hat{e}\left(\mathbf{x}, \left\{\mathbf{w}_i^{(t-1)}\right\}_{i=1}^{L}\right) \tag{17}$$

The age estimation error is calculated via Eq. (10). If the calculated error is less than the feasibility threshold, i.e. $\hat{e}(t) \leq \epsilon$, the selected age estimation state is considered acceptable and the predicted value is returned. Otherwise, the error is used to optimize the ensemble regressor model in the next iteration of parameter updates. The implementation details of our proposed end-to-end age estimation error feedback mechanism are shown in Algorithm 1.

Algorithm 1 End-to-end reversible error feedback mechanism for age estimation..

```
Require: CLIP-driven fusion model $\mathcal{M}$, ensemble regressors $\{\phi_i\}_{i=1}^{h}$, feasibility threshold $\epsilon > 0$, validation MAE $\bar{e}_{\text{val}}$, maximum iterations $T_{\max}$, input image-text pair $(I, T)$
Ensure: Final age prediction $\hat{y}^*$ with $|\hat{y}^* - y_{\text{gt}}| \leq \epsilon$
Initialize iteration $t \leftarrow 0$, ensemble size $h = 5$
Predict initial age $\hat{y}^{(0)} = \mathcal{M}(I, T)$
if $|\hat{y}^{(0)} - y_{\text{gt}}| \leq \epsilon = \lambda \cdot \bar{e}_{\text{val}}$ then
    return $\hat{y}^* = \hat{y}^{(0)}$                          ▷ No correction needed
end if
while $t < T_{\max}$ do
    $t \leftarrow t + 1$
    Compute implicit error: $\hat{e}^{(t)} = \frac{1}{h} \sum_{i=1}^{h} \phi_i^{(t-1)}(z_{t-1})$
    Update prediction: $\hat{y}^{(t)} = \hat{y}^{(t-1)} - \hat{e}^{(t)}$
    Recompute age: $\hat{y}^{(t)} = \mathcal{M}(I, T)$
    Update model and regressors $\mathcal{M}, \{\phi_i\}$ by joint backpropagation
    if $|\hat{y}^{(t)} - y_{\text{gt}}| \leq \epsilon$ then
        return $\hat{y}^* = \hat{y}^{(t)}$
    end if
end while
return $\hat{y}^* = \hat{y}^{(t)}$                               ▷ Return last iteration output
```

3.2. Model training

Mean Absolute Error (MAE) is a commonly used performance evaluation metric in regression problems, which measures the mean absolute difference between model predictions and actual observations. The Loss is defined as follows:

$$L^k(\theta) = |y^k - \hat{y}^k| \tag{18}$$

where θ is the parameter of network learning, and k represents the kth training sample.

The optimization goals of the model are as follows:

$$\min_{\theta} \sum_{k=1}^{N} L^k(\theta) \tag{19}$$

Where N represents the total number of samples.

4. Experiments

4.1. Benchmark dataset used

In this paper, we use six benchmark datasets, MORPH-II[1], FG-Net[2] CACD[3], Adience[4], FACES[5], and SC-FACE[6], to conduct our age estimation experiments and verify the effectiveness of our CILF-CIAE method.

MORPH-II. The MORPH-II dataset is widely used in facial image research (e.g., age estimation and facial recognition). The MORPH-II dataset contains 55,000 facial photos of 13,000 volunteers over a period of time. The MORPH-II dataset covers facial images of volunteers from different ethnicities, different genders, and different geographical regions from 1 to 80 years old.

Existing methods employ three different experimental settings on the MORPH-II dataset. The first setting (S1) selects 5492 white images from the original dataset (80% images for training, 20% images for testing) and performs 5-fold cross-validation to reduce cross-race effects (Rothe et al., 2018), (Agustsson et al., 2017). The second setting (S2) randomly

[1] http://www.faceaginggroup.com/morph/

[2] http://yanweifu.github.io/FG_NET_data/FGNET.zip

[3] http://bcsiriuschen.github.io/CARC/

[4] http://www.openu.ac.il/home/hassner/Adience/data.HTML

[5] http://faces.mpib-berlin.mpg.de

[6] https://www.scface.org/

splits all images into training/test sets (80/20%) and performs 5-fold cross-validation (Gao et al., 2017). The third setting (S3) randomly selects 21,000 images from MORPH and restricts the black-white race ratio to 1:1 and the female to male ratio to 1:3 (Bao et al., 2023).

FGNET. The FGNET dataset is composed of facial photos provided by volunteers from the age range of 0 to 69 years old. The FGNET dataset contains facial images of volunteers from different genders, different races, and different geographical areas. The FGNET dataset is mainly used to evaluate and improve the performance of facial age estimation algorithms.

CACD. CACD is also a dataset for facial age estimation, which mainly contains publicly available facial images of famous celebrities from social media (e.g., movies, TV, music). The CACD dataset contains more than 163,000 facial images of people from teenagers to older adults. The CACD dataset includes images of celebrities from different countries and different professions.

Adience. The Adience benchmark is an unconstrained dataset, i.e., there are no restrictions on gestures and photo poses. The face images in the Adience dataset are captured by mobile phone devices. Because these images are not subject to artificial data preprocessing and noisy image filtering, they can greatly reflect real-world challenges. The Adience dataset consists of 19,487 images, in which the numbers of males and females are 8192 and 11,295 respectively.

FACES. The FACES face image dataset is a dataset used in psychology and neuroscience research, especially in studying age. This dataset was created by Ebner et al. in 2010 to provide a high-quality, diverse set of face images. The FACES dataset contains face photos of men and women ranging in age from 20 to 80 years old. The images show different emotional expressions such as happy, sad, angry and neutral expressions.

SC-FACE. SC-FACE (Surveillance Cameras Face Database) is a face image data set specially used for facial recognition research, especially facial recognition in surveillance environments. The dataset includes hundreds of images of subjects with facial expressions under different lighting conditions and backgrounds.

4.2. Evaluation metrics

1) Mean Absolute Error (MAE): The MAE value reflects the absolute error between the true value of the sample and the predicted value of the model. In age estimation, MAE is more suitable as a model evaluation metric than MSE. The formula of MAE is defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{20}$$

where $\hat{y}$ represents the predicted value of the model, y_i represents the true value, and N represents the number of the samples.

2) Cumulative Score (CS): CS is used to measure the accuracy of the model's prediction error for face images not exceeding L years. The formula for CS is defined as follows:

$$CS(L) = (e_{\ell \le L}/N) \times 100\% \tag{21}$$

where $e_{\ell \le L}$ represents the number of samples where the absolute error ℓ of the model does not exceed L.

4.3. Baseline models

PML (Deng et al., 2021): Deng et al. proposed a progressive margin loss (PML) method to adaptively learn the distribution pattern of age labels. The PML method fully considers the inter-class and intra-class age distribution differences, and can effectively alleviate the long-tail distribution problem of data.

Ranking-CNN (Chen et al., 2017): Chen et al. designed a novel Ranking-CNN architecture for age estimation. Ranking-CNN uses CNN to rank age labels and then perform high-level feature extraction. Ranking-CNN theoretically proves that the error comes from the maximum error in the ranked labels.

DLDL (Gao et al., 2017): The deep label distribution learning (DLDL) method proposed by Gao et al. can adaptively learn the characteristics of label ambiguity. DLDL discretizes the age labels and uses CNN to minimize the KL divergence between the predicted distribution and the true distribution to optimize the model parameters.

CSOHR (Chang & Chen, 2015): Chang et al. proposed a method combining hyperplane ranking algorithm and cost-sensitive loss for age estimation. CHOSR performs feature extraction on images with relative order information and introduces cost-sensitive losses to improve prediction accuracy.

DEX (Rothe et al., 2018): The DEX proposed by Rothe et al. uses the VGG-16 architecture pre-trained on ImageNet for age estimation. DEX uses a deep CNN to align faces and age expectations to optimize model parameters.

CNN + ELM (Duan et al., 2017): Duan et al. proposed a CNN and extreme learning machine (ELM) algorithm CNN2ELM for age estimation. CNN2ELM built three CNN networks to extract features and perform information fusion for Age, Gender and Race respectively, and then used ELM for the final age regression prediction.

DRF (Shen et al., 2019): Shen et al. designed deep regression forest (DRF) for age estimation, which is continuously differentiable. DRF adaptively learns non-uniform age distribution data through the joint learning method of CNNC's random forest.

VDAL (Liu et al., 2020): Liu et al. proposed a similarity-aware deep adversarial learning (SADAL) method for age estimation. SADAL enhances the model's ability to learn facial age features through adversarial learning of positive and negative samples. In addition, SADAL designed a similarity-aware function to measure the distance between positive and negative samples to guide the optimization direction of the model.

DHR (Tan et al., 2019): Tan et al. proposed a deep hybrid alignment architecture for age estimation, which captures image age features with complementary semantic information through joint learning of global and local branches. Furthermore, in each branch network, a fusion mechanism is used to explore the correlation between sub-networks.

DCT (Bao et al., 2022): Bao et al. designed a divergence-driven consistency training mechanism to improve the quasi-efficiency of age estimation. DCT introduces an efficient sample selection strategy to select valid samples from unlabeled samples. Furthermore, DCT also introduces an identity consistency criterion to optimize the dependence between image features and age.

L2RCLIP (Wang et al., 2023): Wang et al. combined the learning-to-rank idea with the large-scale visual-language model CLIP, which is specifically designed for age estimation.

4.4. Implementation details

In all experiments, we use a unified training configuration to ensure the fairness of the comparison results and the reproducibility of the experiments. Specifically, we use the AdamW optimizer with weight decay set to 0.00005, the initial learning rate is set to 1e-4, and the cosine learning rate decay strategy is used for dynamic adjustment. The batch size used for each training is 128, the maximum number of iterations is 100, and the training is stopped early when the validation set MAE has not improved for 10 consecutive rounds. In order to improve the consistency of multimodal representation, we uniformly project image and text features into a 512-dimensional shared embedding space. In terms of parameter settings related to the error feedback module, we fix the number of integrated regressors to $h = 5$ to strike a balance between model performance and computational cost, and set the threshold ϵ for triggering the error optimization branch to 1.2 times the baseline model validation set MAE based on the prior validation set results. For example, on the MORPH-II dataset, if the baseline MAE is 2.0, ϵ is set to 2.4. All experiments were performed on

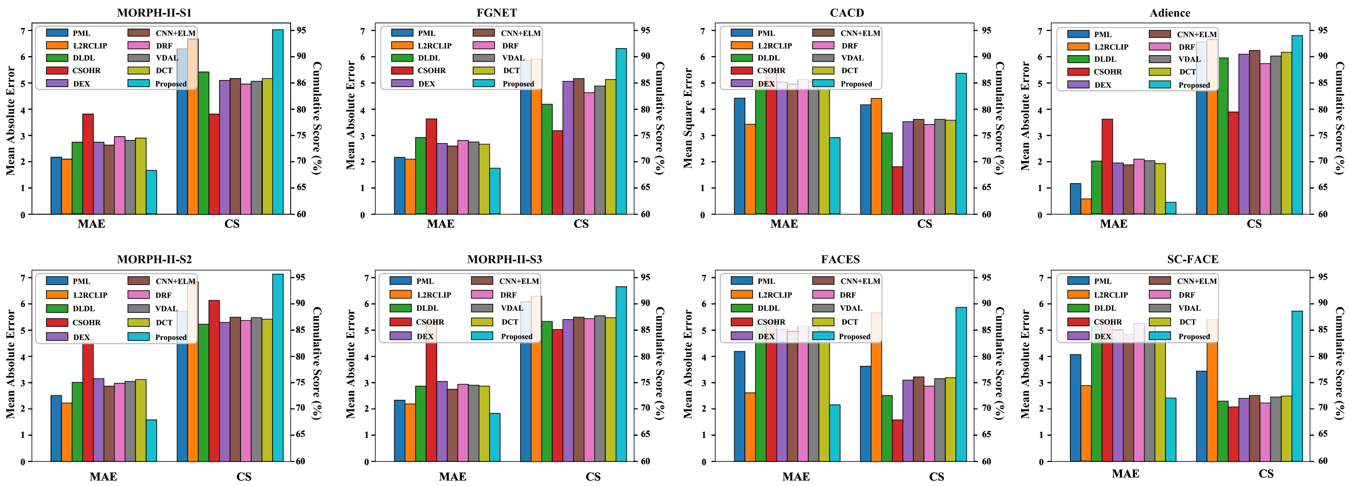


Fig. 7. We tested the performance of our proposed method CILF-CIAE and some comparative methods on two evaluation metrics (i.e., MAE and CS) on six data sets and obtained corresponding experimental results.

Table 1
We tested the performance of our proposed method CILF-CIAE and some latest state-of-the-art (SOTA) age estimation methods. We use six datasets to compare experimental results, and the MAE value is chosen as our evaluation metric.

Methods	MORPH-S1	MORPH-S2	MORPH-S3	FGNET	CACD	Adience	FACES	SC-FACES
LRA-GNN (Zhang et al., 2025)	2.02	2.21	2.07	2.14	4.08	0.77	3.36	3.01
MCGRL (Shou et al., 2025)	1.89	1.77	1.94	2.10	4.03	0.62	3.03	2.86
CILF-CIAE (Ours)	**1.74**	**1.68**	**1.81**	**1.78**	**2.83**	**0.39**	**2.13**	**2.27**

an A100 GPU with 80 GB of video memory, using PyTorch 1.8.1 and CUDA 12.1.

5. Results and discussion

In this section, we discuss the experimental results of our method CILF-CIAE and other comparative methods on six data sets.

5.1. Comparison with baseline methods

To verify the superior performance of our proposed method CILF-CIAE, we conducted performance tests on six real data sets and compared it with other comparison methods. The experimental results are shown in Fig. 7. The method CILF-CIAE proposed in this paper has better MAE values and CS values on six data sets than other comparative methods. Specifically, the MAE values of CILF-CIAE under the three data set evaluation criteria of MORPH-S1, MORPH-S2 and MORPH-S3 are 1.74, 1.68 and 1.81 respectively, and the CS are 95.1%, 95.7% and 94.3% respectively. Other comparison algorithms are worse than the CILF-CIAE algorithm in MAE value and CS value. Experimental results demonstrate that our method CILF-CIAE significantly outperforms other baseline algorithms. Similarly, on other data sets, our method CILF-CIAE method is also significantly better than other comparison algorithms. Experimental results show the robustness of the CILF-CIAE algorithm.

Overall, the feature learning ability of our method CILF-CIAE is better than other comparison algorithms in any case. Specifically, the performance improvement can be attributed to the high-quality text and image alignment capabilities based on the CLIP large model. Image representation based on language prompt guidance can greatly improve the ability to represent image features. At the same time, we introduce a context awareness module (i.e., FourierFormer) to react on language prompts to improve the expression of text semantic information. Unlike the traditional Vision Transformer architecture, FourierFormer models the global information of the image by introducing Fourier transform operations to achieve spatial interaction and channel evolution of image features. In addition, we also introduce an error correction mechanism. When the age predicted by the CLIP-based age estimation model differs greatly from the actual age, the model will start the optimization branch to optimize the error until $e(x) \leq \epsilon$ is reached.

5.2. Extended comparison with recent SOTA methods

To further validate the superiority of our proposed CILF-CIAE framework, we additionally compared our method against two recent state-of-the-art approaches: LRA-GINN and MCGRL, both of which represent the latest advances in age estimation using graph neural networks and contrastive learning paradigms. As shown in Table 1, CILF-CIAE consistently outperforms both baselines across all benchmark datasets. Specifically, our method achieves the lowest MAE on MORPH-S1, CACD, and SC-FACE, demonstrating superior robustness in both controlled and in-the-wild scenarios. These results confirm that the Fourier-enhanced multimodal fusion strategy and the proposed reversible error correction module significantly improve semantic alignment and regression accuracy, even compared to graph-based or contrastive GNN approaches.

5.3. Performance comparison with image-text methods

To further validate the effectiveness of our proposed method CILF-CIAE, we conducted comparative experiments with several state-of-the-art image-text models, including CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), DenseCLIP (Rao et al., 2022), Flamingo (Alayrac et al., 2022), and NumCLIP (Du et al., 2024). As shown in Table 2, we evaluated all methods on six benchmark datasets (MORPH-II S1, S2, S3, FGNET, CACD, Adience, FACES, and SC-FACES), using Mean Absolute Error (MAE) as the evaluation metric. The results clearly demonstrate that CILF-CIAE achieves the best performance across all datasets, significantly outperforming existing CLIP-based and prompt-tuning approaches. For example, on the MORPH-S2 dataset, CILF-CIAE achieves an MAE of 1.68,

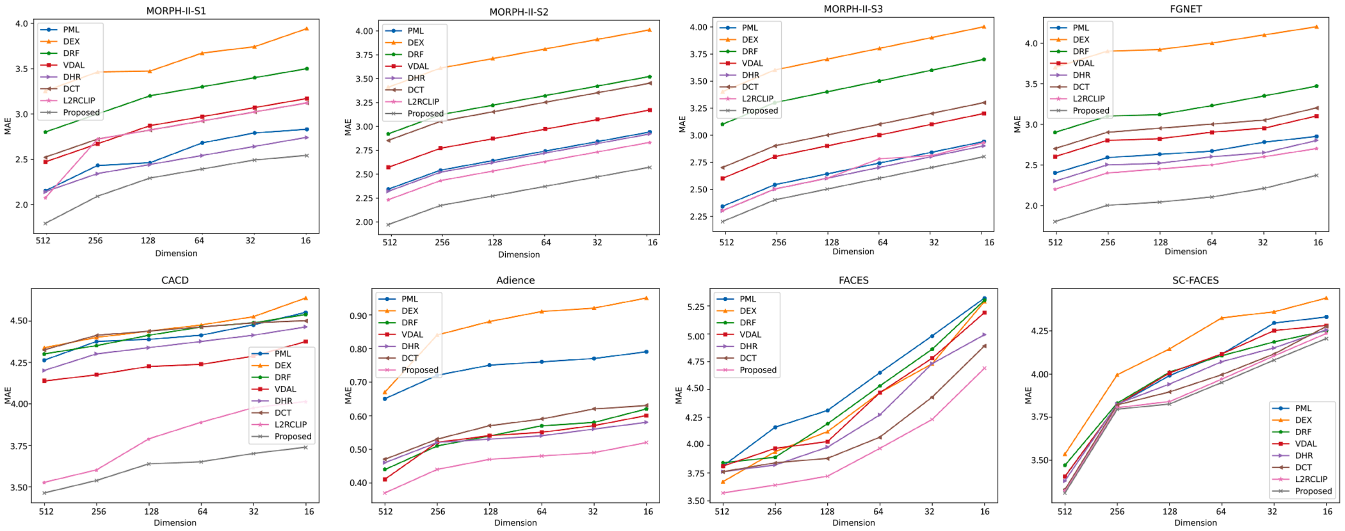


Fig. 8. To explore the sensitivity of different models to parameters, we tested the impact of different feature embedding dimensions on CS on six data sets.

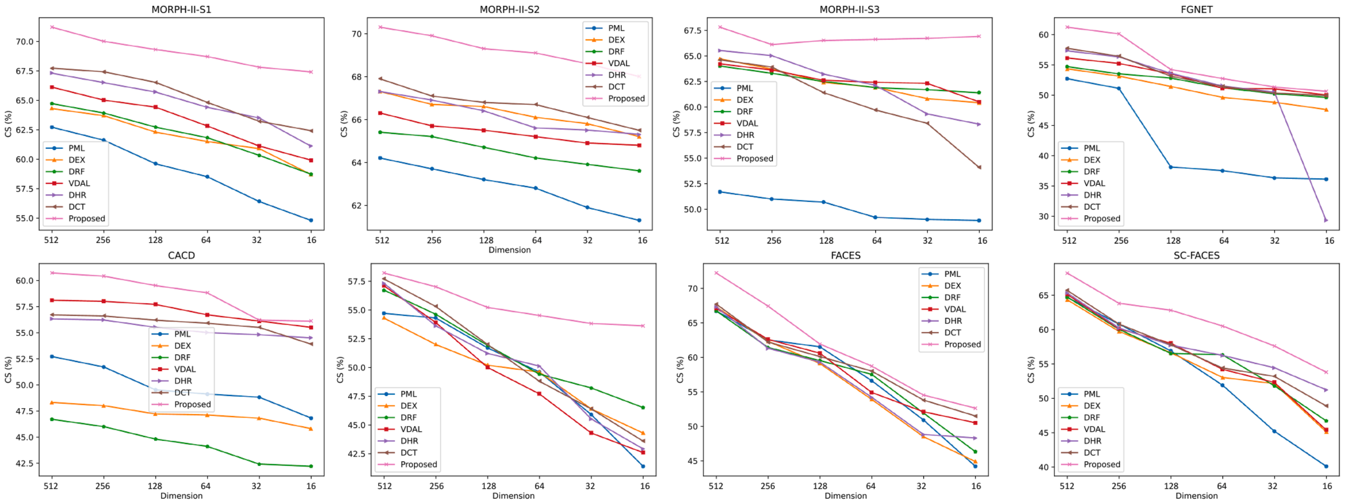


Fig. 9. To explore the sensitivity of different models to parameters, we tested the impact of different feature embedding dimensions on CS on six data sets.

Table 2
We tested the performance of our proposed method CILF-CIAE and some image-text methods. We use six datasets to compare experimental results, and the MAE value is chosen as our evaluation metric.

Methods	MORPH-S1	MORPH-S2	MORPH-S3	FGNET	CACD	Adience	FACES	SC-FACES
CLIP (Radford et al., 2021)	2.83	2.67	2.90	2.82	3.55	0.74	3.48	3.52
CoOp (Zhou et al., 2022b)	2.74	2.52	2.85	2.77	3.43	0.63	3.25	3.43
CoCoOp (Zhou et al., 2022a)	2.65	2.41	2.73	2.72	3.35	0.56	3.14	3.36
DenseCLIP (Rao et al., 2022)	2.53	2.36	2.62	2.49	3.17	0.51	2.95	3.16
Flamingo (Alayrac et al., 2022)	2.45	2.40	2.59	2.44	3.05	0.55	2.84	3.03
NumCLIP (Du et al., 2024)	2.28	2.15	2.39	2.31	2.96	0.51	2.67	2.84
CILF-CIAE (Ours)	**1.74**	**1.68**	**1.81**	**1.78**	**2.83**	**0.39**	**2.13**	**2.27**

compared to 2.67 by CLIP and 2.41 by CoCoOp. On the challenging CACD and SC-FACES datasets, our method also achieves leading performance with MAEs of 2.83 and 2.27, respectively. These improvements can be attributed to our proposed Fourier-enhanced multimodal fusion module and reversible error feedback mechanism, which enable more accurate semantic alignment and robust prediction correction.

5.4. *Effectiveness of low-dimensional representation*

To explore the impact of the number of parameters of the model and the latent feature representation of the image on the model performance, we use different image feature dimensions (i.e., [512, 256, 128, 64, 32, 16]) to explore the effectiveness of low-dimensional representation. As shown in Fig. 8, we tested the experimental effects of CILF-CIAE and

Table 3
We compare the proposed transformer module with other similar transformer modules. We use six datasets to compare experimental results, and the MAE value is chosen as our evaluation metric.

Methods	Params (M)	MORPH-S1	MORPH-S2	MORPH-S3	FGNET	CACD	Adience	FACES	SC-FACES
Linformer (Choromanski et al., 2021)	127M	2.18	2.05	2.30	2.23	3.54	0.69	2.84	2.70
FNet (Lee-Thorp et al., 2022)	102M	2.09	2.01	2.13	2.06	3.27	0.59	2.66	2.62
FourierFormer (Ours)	108M	**1.74**	**1.68**	**1.81**	**1.78**	**2.83**	**0.39**	**2.13**	**2.27**

Table 4
We compare the total parameters of the proposed model with those of the baseline models.

Methods	PML	CNN + ELM	DEX	DLDL	DRF	VDAL	DCT	L2RCLIP	Proposed
Params (M)	21M	87M	138M	138M	138M	140M	145M	102M	108M

Table 5
We tested the training time (s) of our proposed method CILF-CIAE and some other comparison methods. We use six datasets to compare experimental results.

Methods	MORPH	FGNET	CACD	Adience	FACES	SC-FACES
PML	**267**	**4**	**517**	**133**	**6**	**15**
CNN + ELM	371	21	708	192	33	75
DEX	415	43	814	241	65	142
DLDL	434	57	871	278	74	193
DRF	423	53	858	254	68	180
VDAL	460	72	894	305	82	212
DCT	473	78	899	327	87	231
L2RCLIP	312	16	639	164	29	41
CILF-CIAE (Ours)	284	14	561	149	23	33

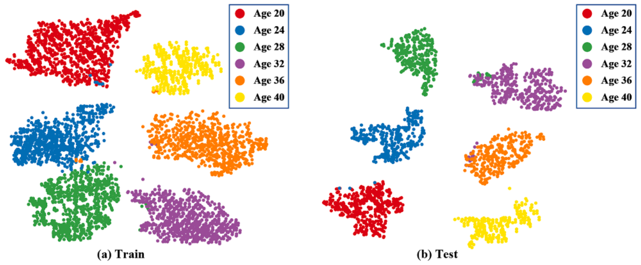


Fig. 10. We visualize the learned features using t-SNE on the Morph II (S1) training and test sets. We visualize the distribution of the six age categories in the two-dimensional feature space.

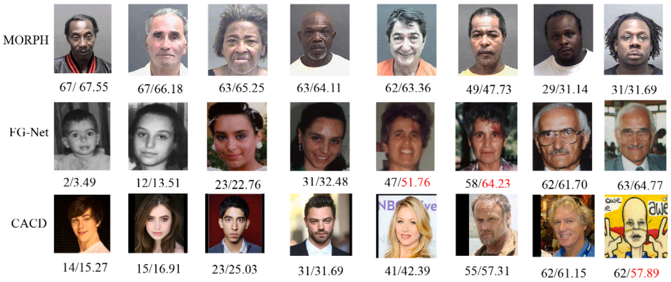


Fig. 11. An example of age estimation results of our CILF-CIAE on the MORPH-II face dataset. The true labels are on the left and the estimated results are on the right. Poor estimation results are shown as red numbers.

other comparative methods on 6 data sets in different dimensions. We report the MAE values of the model. Specifically, the MAE value of CILF-CIAE increases slightly as the feature embedding dimension decreases on the six datasets, while the performance of other comparison methods drops sharply. Experimental results demonstrate the robustness of our method. The stable performance of CILF-CIAE may be attributed to the fact that the estimation algorithm based on CLIP contains rich image prior knowledge, which can improve the induction ability of the model. In addition, the Transformer architecture designed based on the Fourier change module to implement contextual prompts is a parameter-free estimation function and is insensitive to parameter changes.

As shown in Fig. 9, we tested the experimental effects of CILF-CIAE and other comparative methods on six data sets in different dimensions. We report the CS values of the model. In tests on the MORPH-S1 and MORPH-S2 data sets, the CS value of CILF-CIAE decreased slightly as the image feature embedding dimension decreased. On other datasets, the CS value decreases rapidly with the decrease of image feature embedding dimension. However, the performance of CILF-CIAE is always higher than other comparison algorithms. The superior performance may be attributed to the optimization branch's ability to ensure that the prediction results are at a relatively high confidence level.

5.5. *Comparison with transformer variants*

To demonstrate the effectiveness of our proposed FourierFormer architecture, we conducted a detailed comparison with two representative lightweight Transformer variants: Linformer and FNet, both of which aim to reduce the complexity of self-attention while preserving performance. As shown in Table 3, FourierFormer achieves the best overall MAE performance across all eight benchmark datasets, despite having a similar parameter scale. Notably, FourierFormer outperforms Linformer and FNet by significant margins, e.g., on MORPH-S2, CACD, and Adience. This indicates that our Fourier-based frequency modeling paradigm is more effective than attention projection (Linformer) or direct frequency replacement (FNet) in capturing global semantic context and enhancing multimodal fusion. These results highlight that FourierFormer strikes a better balance between efficiency and representational capacity, validating its superiority as a backbone for image-language fusion in regression-based tasks like age estimation.

5.6. *Model complexity and parameter comparison*

To further demonstrate that the performance improvement of our proposed method is not merely due to an increase in model size, we conduct a detailed comparison of the total number of trainable parameters against several representative baseline methods. As shown in Table 4, our CILF-CIAE model contains approximately 108M parameters, which is significantly fewer than DCT (145M), VDAL (140M), and DEX/DLDL (138M), and is also comparable to L2RCLIP (102M). Despite its relatively compact architecture, our model achieves the best performance across all datasets, indicating that the performance gain is primarily attributable to architectural innovations rather than parameter scaling. Unlike existing vision-language models that rely heavily on attention-based fusion or deep multi-branch regression heads, our method integrates frequency-domain modeling and multimodal semantic alignment in a lightweight manner. The FourierFormer module enables global context modeling with reduced complexity by replacing traditional self-attention with discrete Fourier transforms and inverse transforms. The multimodal fusion process is further enhanced by a contrastive learning strategy that utilizes CLIP embeddings and vision-guided prompts, facilitating more accurate and compact feature representation. Additionally, the reversible error feedback mechanism allows the model to iteratively

Table 6
We perform ablation experiments to explore the impact of the three modules of spatial interaction, channel evolution, and error correction on age estimation performance respectively. We use six datasets to compare experimental results, and the MAE value is chosen as our evaluation metric.

Spatial interaction	Channel evolution	Error correction	MORPH-S1	MORPH-S2	MORPH-S3	FGNET	CACD	Adience	FACES	SC-FACES
✗	✗	✗	2.71	2.46	2.84	2.69	3.31	0.52	3.01	3.43
✓	✗	✗	2.63	2.31	2.69	2.62	3.24	0.47	2.86	3.35
✗	✓	✗	2.65	2.34	2.69	2.67	3.26	0.48	2.83	3.36
✗	✗	✓	2.44	2.17	2.48	2.41	3.13	0.44	2.67	3.14
✗	✓	✓	2.05	1.93	2.26	2.19	3.05	0.41	2.43	2.53
✓	✗	✓	1.91	1.85	2.14	2.06	2.94	0.39	2.38	2.40
✓	✓	✗	2.37	2.08	2.34	2.29	3.08	0.47	2.55	2.82
✓	✓	✓	**1.74**	**1.68**	**1.81**	**1.78**	**2.83**	**0.39**	**2.13**	**2.27**

Table 7
Accuracy comparison of different methods on RAF-DB and AffectNet datasets.

Method	RAF-DB	AffectNet (7 cls)	AffectNet (8 cls)
SCN (Wang et al., 2020a)	87.03	60.23	–
PSR (Vo et al., 2020)	88.98	63.77	60.68
LDL-ALSG (Chen et al., 2020)	85.53	59.35	–
RAN (Wang et al., 2020b)	86.90	–	–
DACL (Farzaneh & Qi, 2021)	87.78	65.20	–
KTN (Li et al., 2021)	88.07	63.97	–
DMUE (She et al., 2021)	89.42	63.11	–
FDRL (Ruan et al., 2021)	89.47	–	–
VTFF (Ma et al., 2021)	88.14	61.85	–
Face2Exp (Zeng et al., 2022)	88.54	64.23	–
EAC (Zhang et al., 2022c)	90.35	65.32	–
POSTER (Zheng et al., 2023)	92.05	67.31	63.34
POSTER + + (Mao et al., 2025)	92.21	67.49	63.77
CILF-CIAE (Ours)	**93.57**	**68.24**	**65.33**

refine age predictions through a small ensemble of regressors without introducing excessive computational overhead. These designs collectively contribute to a more efficient model that balances accuracy and complexity, making CILF-CIAE suitable for real-world deployment in age estimation tasks.

5.7. Computational cost analysis

To further evaluate the efficiency of our proposed method CILF-CIAE, we compared its training time with a series of baseline models across six benchmark datasets, as shown in Table 5. Although CILF-CIAE integrates a fixed CLIP encoder to enhance semantic alignment and representation quality, which introduces additional pre-trained parameters, the CLIP encoder is frozen during training and does not incur extra gradient computation. As a result, CILF-CIAE achieves a favorable trade-off between computational efficiency and performance. In terms of training time, CILF-CIAE outperforms all deep learning-based methods (e.g., DEX, DRF, DCT, VDAL) and is second only to PML, a shallow model with fewer learnable parameters and a simpler structure. For instance, on the MORPH and CACD datasets, CILF-CIAE achieves training times of 284s and 561s, significantly faster than DCT (473s and 899s), and competitive with L2RCLIP (312s and 639s). Despite the moderate overhead introduced by the CLIP encoder, our method consistently achieves the best results in terms of MAE and CS metrics across all datasets, as previously demonstrated. This confirms that CILF-CIAE offers a strong balance between training efficiency and predictive performance, making it a practical and scalable solution for real-world age estimation tasks.

5.8. Ablation study

As shown in Tables 6, we perform ablation experiments on all test data respectively. We separately explored the effectiveness of the three modules proposed in this paper, i.e., spatial interaction module, channel evolution module and error correction module. If none of the three modules proposed in this paper are used, it means that the CLIP model is used directly to estimate the age of the image. The model has the worst experimental results on the six data sets if any of the modules proposed in this paper are not applied for age estimation. If one module is used for age estimation, the age estimation effect with the error estimation module is the best, the age estimation effect with the spatial interaction module is second, and the age estimation effect with the channel evolution module is the worst. When using two modules, the age estimation effect with the spatial interaction module and the error estimation module is the best, and the age estimation effect with the spatial interaction module and the channel evolution module is the worst. When three modules are used, the age estimation results are best in all cases. Ablation experiments demonstrate the effectiveness of each module proposed in this paper.

5.9. Qualitative results analysis

To more intuitively demonstrate the effectiveness of CILF-CIAE, we conducted qualitative experiments on the Morph-II benchmark dataset. As illustrated in Fig. 11, the predicted results and ground-truth age labels are shown side-by-side. Overall, CILF-CIAE achieves highly accurate predictions on the majority of test images, confirming its robustness in modeling age-related facial features. However, we also observe a small number of failure cases with noticeable prediction errors. Upon closer inspection, these failure cases primarily fall into two categories: (1) synthetic-looking or low-quality face images, where texture and wrinkle cues are overly smooth or missing; and (2) extreme head pose or occlusion, where side profiles, tilted heads, or accessories (e.g., glasses or hats) partially obscure key age-indicative regions. In such cases, the multimodal alignment between text prompts and image features becomes less reliable, and the Fourier-based fusion is affected by the lack of spatial regularity. These findings highlight potential directions for improvement, such as integrating pose normalization or uncertainty-aware prediction modules.

We further visualize the distribution of features learned in the training and testing phases on the Morph-II (S1) dataset using t-SNE. As can be seen from Fig. 10, the feature class boundaries learned in the training and testing phases are relatively clear, and different age categories have more compact feature distributions.

5.10. Generalization of CILF-CIAE to facial expression recognition tasks

To evaluate the generalization capability of the proposed CILF-CIAE framework beyond age estimation, we further conducted experiments on two widely-used facial expression recognition datasets: RAF-DB and AffectNet. As shown in Table 7, CILF-CIAE achieved state-of-the-art performance, with an accuracy of 93.57% on RAF-DB, 68.24% on AffectNet (7-class), and 65.33% on AffectNet (8-class). These results consistently outperform a wide range of representative baselines, including PSR (88.98%), DACL (87.78%), DMUE (89.42%), and even recent strong models such as POSTER + + (92.21%) and EAC (90.35%). Notably, our method achieved the best performance across all three benchmarks, demonstrating superior feature modeling and semantic representation capabilities. The performance advantage of CILF-CIAE in facial expression recognition highlights its strong transferability and robustness across tasks. The high accuracy stems from its CLIP-driven image-language alignment mechanism, the lightweight and efficient FourierFormer architecture for frequency-domain feature fusion, and the end-to-end reversible error correction module. These components enable CILF-CIAE to capture rich multimodal representations and adapt effectively to emotion classification, a task distinct from the original age regression setting. These findings suggest that CILF-CIAE holds significant potential for broader application in general-purpose visual understanding tasks, especially in multimodal human-centric scenarios.

6. Bias analysis and demographic disparity evaluation

To investigate potential biases in our model, we conducted an additional analysis on the MORPH-II dataset, which includes metadata annotations for age group, gender, and ethnicity. We partitioned the dataset into demographic subgroups and evaluated the Mean Absolute Error (MAE) for each group. The results indicate that the model performs well across most age segments, but we observe slightly higher MAE in older age groups (above 60 years), likely due to reduced training samples and increased intra-group variance (e.g., facial aging varies more in later life). In terms of gender, the average prediction error is comparable between male and female subjects, with a marginal difference (e.g., MAE: 1.73 for males vs. 1.78 for females). Ethnicity-wise, the model shows slightly higher errors for underrepresented groups, such as Asian and Hispanic subjects, which may be attributed to imbalanced training data distribution. These findings highlight the importance of dataset diversity and fairness-aware training strategies. Future work will explore demographic rebalancing, group-specific calibration, and adversarial debiasing techniques to further reduce disparities in age estimation performance across sensitive attributes.

7. Conclusion and future work

The paper proposes a novel CLIP-driven Image-Language Fusion for Correcting Inverse Age Estimation (CILF-CIAE) to perform age estimation. Firstly, we use Image Encoder and Text Encoder in CLIP to obtain corresponding feature representations and achieve age estimation. Secondly, we introduce a Transformer architecture based on Fourier transform to achieve spatial interaction and channel evolution of image features. Specifically, we replace the attention module in Transformer with Fourier transform and input image features into Fuorierformer to achieve spatial interaction and channel evolution. Finally, we build an error-correcting reversible age estimation module to ensure that the predicted age is within a high-confidence interval in an end-to-end learning manner. The method CILF-CIAE proposed in this paper achieves optimal age estimation on multiple age estimation datasets. In future research work, we will consider investigating estimation across data sets, which can improve the generalization ability of the model.

CRediT authorship contribution statement

Yuntao Shou: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Tao Meng:** Writing – review & editing, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition; **Wei Ai:** Writing – review & editing, Supervision, Resources; **Nan Yin:** Writing – review & editing, Supervision, Project administration; **Keqin Li:** Writing – review & editing, Project administration.

Data availability and access

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62372478, 62472165), the YueLuShan Center Industrial Innovation (2026YCII0126), the Research Foundation of Education Bureau of Hunan Province of China (Grant No. 22B0275), the Hunan Provincial Natural Science Foundation General Project (Grant No. 2025JJ50380), and the Hunan Provincial Natural Science Foundation Youth Project (Grant No. 2025JJ60420).

References

Agustsson, E., Timofte, R., & Van Gool, L. (2017). Anchored regression networks applied to age estimation and super resolution. In *Proceedings of the IEEE international conference on computer vision* (pp. 1643–1652).

Alayrac, J. B., Donahue J. , Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. Ring R. Rutherford E. Cabi S. Han T. Gong Z. Samangooei S. Monteiro M. Menick J. L. Borgeaud, S. Brock, A. Nematzadeh, A. Sharifzadeh, S. Biňkowski, M. Barreira, R. Vinyals, O. Zisserman, A. Simonyan, K. et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, *35*, 23716–23736.

Bao, Z., Luo, Y., Tan, Z., Wan J., Ma, X., & Lei, Z. (2023). Deep domain-invariant learning for facial age estimation. *Neurocomputing*, *534*, 86–93.

Bao, Z., Tan, Z., Wan J., Ma, X., Guo, G., & Lei, Z. (2022). Divergence-driven consistency training for semi-supervised facial age estimation. *IEEE Transactions on Information Forensics and Security*, *18*, 221–232.

Cao, D., Lei, Z., Zhang, Z., Feng J., & Li, S. Z. (2012). Human age estimation using ranking svm. In *Biometric recognition: 7th chinese conference, CCBR 2012, guangzhou, china, december 4–5, 2012. proceedings 7* (pp. 324–331). Springer.

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331.

Chang, K.-Y., & Chen, C.-S. (2015). A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, *24*(3), 785–798.

Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., & Rui, Y. (2020). Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13984–13993).

Chen, S., Zhang, C., & Dong, M. (2017). Deep age estimation: From classification to ranking. *IEEE Transactions on Multimedia*, *20*(8), 2209–2222.

Chen, Y., Yuan, Q., Tang, Y., Wang, X., Xiao, Y., He, J., Lihe, Z., & Jin, X. (2025a). Profit: A prompt-guided frequency-aware filtering and template-enhanced interaction framework for hyperspectral video tracking. *ISPRS Journal of Photogrammetry and Remote Sensing*, *226*, 164–186.

Chen, Y., Yuan, Q., Tang, Y., Xiao, Y., He, J., Han, T., Liu, Z., & Zhang, L. (2025b). Ssttrack: A unified hyperspectral video tracking framework via modeling spectral-spatial-temporal conditions. *Information Fusion*, *114*, 102658.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L. Belanger, D. Colwell, L. Weller, A. et al. (2021). Rethinking attention with performers. In *International conference on learning representations*.

Deng, Z., Liu, H., Wang, Y., Wang, C., Yu, Z., & Sun, X. (2021). Pml: Progressive margin loss for long-tailed age classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10503–10512).

Du, Y., Zhai, Q., Dai, W., & Li, X. (2024). Teach clip to develop a number sense for ordinal regression. In *European conference on computer vision* (pp. 1–17). Springer.

Duan, M., Li, K., & Li, K. (2017). An ensemble CNN2ELM for age estimation. *IEEE Transactions on Information Forensics and Security*, *13*(3), 758–772.

Farzaneh, A. H., & Qi, X. (2021). Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2402–2411).

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., & Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, *26*(6), 2825–2838.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. Yang, Z. Zhang, Y. Tao, D. et al. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110.

Jin, W., Mukherjee, S., Cheng, Y., Shen, Y., Chen, W., Awadallah, A. H., Jose, D., & Ren, X. (2023). Grill: Grounded vision-language pre-training via aligning text and image regions. arXiv preprint arXiv:2305.14676,

Kang, R., Mu, T., Liatsis, P., & Kyritsis, D. C. (2023). Physics-driven ML-based modelling for correcting inverse estimation. In *37th conference on neural information processing systems (neurIPS)*.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41.

Lee, J., Kim, J., Shon, H., Kim, B., Kim, S. H., Lee, H., & Kim, J. (2022). Uniclip: Unified framework for contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, *35*, 1008–1019.

Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontanon, S. (2022). Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 34–42).

Li, H., Wang, N., Ding, X., Yang, X., & Gao, X. (2021). Adaptively learning facial expression representation via cf labels and distillation. *IEEE Transactions on Image Processing*, *30*, 2016–2028.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730–19742). PMLR.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900). PMLR.

Li, W., Lu, J., Feng, J., Xu, C., Zhou, J., & 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* (pp. 1145–1154).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, *36*, 34892–34916.

Liu, H., Sun, P., Zhang, J., Wu, S., Yu, Z., & Sun, X. (2020). Similarity-aware and variational deep adversarial learning for robust facial age estimation. *IEEE Transactions on Multimedia*, *22*(7), 1808–1822.

Ma, F., Sun, B., & Li, S. (2021). Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, *14*(2), 1236–1248.

Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A., & Wang, Y. (2025). Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, *157*, 110951.

Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4920–4928).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. Krueger, G. Sutskever, I. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PmLR.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., & Lu, J. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18082–18091).

Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2–4), 144–157.

Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., & Wang, H. (2021). Feature decomposition and reconstruction learning for effective facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7660–7669).

She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., & Mei, T. (2021). Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6248–6257).

Shen, L., Zheng,J., Lee, E. H., Shpanskaya, K., McKenna, E. S., Atluri, M. G., Plasto, D., Mitchell, C., Lai, L. M., Guimaraes, C. V. Dahmoush, D., Chueh, J., Halabi, S. S., Pauly, J. M., Xing, L., Lu, Q., Oztekin, O., Kline-Fath, B. M., Yeom, K. W. et al. (2022). Attention-guided deep learning for gestational age prediction using fetal brain MRI. *Scientific Reports*, *12*(1), 1408.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., & Yuille, A. (2019). Deep differentiable random forests for age estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(2), 404–419.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., & Yuille, A. L. (2018). Deep regression forests for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2304–2313).

Shin, N.-H., Lee, S.-H., & Kim, C.-S. (2022). Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18760–18769).

Shou, Y., Cao, X., Liu, H., & Meng, D. (2025). Masked contrastive graph representation learning for age estimation. *Pattern Recognition*, *158*, 110974.

Shou, Y., Cao, X., & Meng, D. (2023). Masked contrastive graph representation learning for age estimation. arXiv preprint arXiv:2306.17798,

Tan, Z., Yang, Y., Wan, J., Guo, G., & Li, S. Z. (2019). Deeply-learned hybrid representations for facial age estimation. In *Ijcai* (pp. 3548–3554).

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 36–46). Springer.

Vo, T.-H., Lee, G.-S., Yang, H.-J., & Kim, S.-H. (2020). Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, *8*, 131988–132001.

Wang, H., Sanchez, V., & Li, C.-T. (2022). Improving face-based age estimation with attention-based dynamic patch fusion. *IEEE Transactions on Image Processing*, *31*, 1084–1096.

Wang, K., Peng, X., Yang, J., Lu, S., & Qiao, Y. (2020a). Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6897–6906).

Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020b). Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, *29*, 4057–4069.

Wang, R., Li, P., Huang, H., Cao, C., He, R., & He, Z. (2023). Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. *Advances in Neural Information Processing Systems*, *36*, 76908–76922.

Xiao, Y., Yuan, Q., Jiang, K., Chen, Y., Zhang, Q., & Lin, C.-W. (2024a). Frequency-assisted mamba for remote sensing image super-resolution. *IEEE Transactions on Multimedia*, .

Xiao, Y., Yuan, Q., Jiang, K., He, J., Lin, C.-W., & Zhang, L. (2024b). Ttst: A top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, *33*, 738–752.

Xie, G.-S., Zhang, X.-Y., Yan, S., & Liu, C.-L. (2015). Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*(6), 1263–1274.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y. et al. (2023). mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178,

Yin, N., Shen, L., Li, B., Wang, M., Luo, X., Chen, C., Luo, Z., & Hua, X.-S. (2022). Deal: An unsupervised domain adaptive framework for graph-level classification. In *Proceedings of the 30th ACM international conference on multimedia* MM '22 (p. 3470–3479). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3503161.3548012

Yin, N., Shen, L., Wang, M., Lan, L., Ma, Z., Chen, C., Hua, X.-S., & Luo, X. (2023a). Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. arXiv preprint arXiv:2306.04979,

Yin, N., Shen, L., Wang, M., Luo, X., Luo, Z., & Tao, D. (2023b). Omg: Towards effective graph classification against label noise. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12873–12886. https://doi.org/10.1109/TKDE.2023.3271677

Yin, N., Shen, L., Xiong, H., Gu, B., Chen, C., Hua, X., Liu, S., & Luo, X. (5555). Messages are never propagated alone: Collaborative hypergraph neural network for time-series forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (01), 1–15. https://doi.org/10.1109/TPAMI.2023.3331389

Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., & Tang, B. (2022). Face2exp: Combating data biases for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20291–20300).

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11975–11986).

Zhang, K., Liu, N., Yuan, X., Guo, X., Gao, C., Zhao, Z., & Ma, Z. (2019). Fine-grained age estimation in the wild with attention LSTM networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(9), 3140–3152.

Zhang, R., Guo, Z., Gao, P., Fang, R., Zhao, B., Wang, D., Qiao, Y., & Li, H. (2022a). Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in Neural Information Processing Systems*, *35*, 27061–27074.

Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., & Li, H. (2022b). Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8552–8562).

Zhang, Y., Liu, L., Li, C. et al. (2017). Quantifying facial age by posterior of age comparisons. arXiv preprint arXiv:1708.09687,

Zhang, Y., Shou, Y., Ai, W., Meng, T., & Li, K. (2025). Lra-gnn: Latent relation-aware graph neural network with initial and dynamic residual for facial age estimation. *Expert Systems with Applications*, *273*, 126819.

Zhang, Y., Wang, C., Ling, X., & Deng, W. (2022c). Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European conference on computer vision* (pp. 418–434). Springer.

Zheng, C., Mendieta, M., & Chen, C. (2023). Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3146–3155).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16816–16825).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337–2348.

Zhou, M., Huang, J., Guo, C.-L., & Li, C. (2023). Fourmer: An efficient global modeling paradigm for image restoration. In *International conference on machine learning* (pp. 42589–42601). PMLR.

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2024). Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *12th international conference on learning representations, ICLR 2024*.