Vision Transformers for Age Prediction from Gait Energy Image Data

Chengzong Zhao[1] **Jack Wang**[1] **Zara Cook**[1] **Grant Sinha**[1] **Simon Li**[1]
Chang Shu[2] **Pengcheng Xi**[1,2]
[1]University of Waterloo, Canada
[2]National Research Council Canada
{c233zhao,j88wang,zncook,gsinha,simon.li2}@uwaterloo.ca
{chang.shu,pengcheng.xi}@nrc-cnrc.gc.ca

Abstract

Gait age estimation aims to predict a person's age using visual surveillance information. One popular approach involves using Gait Energy Images (GEIs), which capture the essence of an individual's gait for analysis. Nonetheless, training a model from scratch demands considerable computational resources and extensive data. In contrast to the traditional approaches, we utilized pre-trained vision transformer (ViT) models to enhance the performance. We froze the backbone of the pre-trained transformers and assessed their capabilities in zero-shot tasks by training regression heads on a compact dataset. Our approach yielded an optimal model with the best Mean Average Error (MAE) of 10. The findings suggest that the advanced ViT models can effectively carry out zero-shot predictions in gait recognition tasks while maintaining low computational demands and utilizing minimal datasets. We expect that the research findings will provide an insight into vision transformer-based gait recognition for future research and applications.

1 Introduction

Gait recognition, in general, refers to a biometric application that aims to identify pedestrians by their walking patterns [1]. This paper focuses on the application of gait recognition for the purpose of age estimation. Accurately estimating the ages of individuals within a crowd has diverse applications, including enhancing public security, refining marketing strategies, improving aging healthcare, and informing urban planning initiatives. Gait recognition as the medium for estimating age offers two outstanding advantages: it is relatively low-cost and it requires less cooperation from individuals, especially compared to other methods such as surveys or facial recognition.

For several years, strategies employing deep convolutional neural networks (CNNs) with Gait Energy Image (GEI) [2] as input have been prominent in addressing the challenge of age prediction from gait [3–7]. However, all CNN-based methodologies rely on deep learning techniques that necessitate training on substantial datasets for more than 100 epochs [6, 7], resulting in a training process that is notably time-intensive. This also becomes problematic when there is not enough data to train a model from scratch.

The Vision Transformer (ViT) [8], which utilizes self-attention mechanism [9], has recently achieved remarkable success in computer vision. Building on this advancement, the concept of self-supervised pre-trained models has been introduced [10]. These models are pre-trained on extensive datasets for general applications and can be subsequently fine-tuned for a variety of downstream tasks. Capitalizing on this attribute, we propose an approach to gait recognition tasks. Our method involves training downstream regressors atop different pre-trained ViT models, offering a cost-effective and high-performing solution.

Our research investigates the zero-shot capabilities of pre-trained ViT models in the context of gait age prediction tasks using GEI. We establish the baselines for transformer-based gait recognition models and examine the efficacy of zero-shot ViT models within this domain. Our contributions are as follows:

- Zero-shot cost-effective approaches using pre-trained models.
- Investigating the capacity of pre-trained vision transformer models on the gait age estimation task.
- Establishing baseline transformer-based models for future studies.

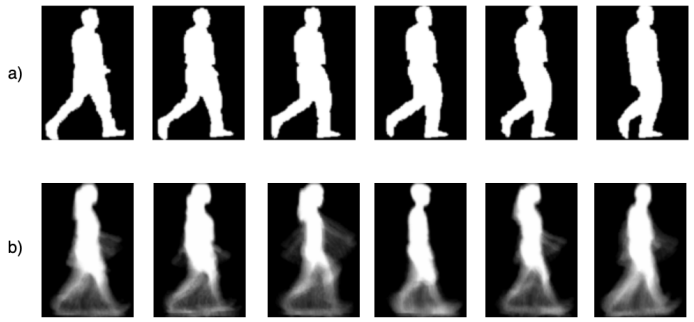

Fig. 1: (a) A segment of a binary silhouette gait cycle, which comprises a series of continuous human motions. (b) Gait Energy Images (GEIs) [2]

2 Literature Review

Gait Energy Image Originally introduced by [2], the Gait Energy Image (GEI) is an effective method that condenses a sequence of a gait cycle into a singular gait template through a weighted average. This process of averaging effectively eliminates a significant amount of noise and compacts time-related information into a singular dimension, yet preserves a comparable level of information [5]. It is possible to calculate the GEI in the following manner when the Gait Cycle image sequence is $B_t(x,y)$:

$$G(x,y) = \frac{1}{N} B_t(x,y) \tag{1}$$

where $B_t(x,y)$ is the context of a series at time t. x and y describe the coordinates of each frame B or image B, and N is the total number of images taken in a Gait Cycle [2]. Fig. 1 demonstrates the relationship between binary silhouette gait cycle and the GEI.

Gait-based Age Estimation In the early stage of machine learning, age estimation is established on classification tasks using support vector machines (SVM). Makihara et al. [11] tried to classify gaits into four classes, namely children, adult males, adult females, and the elderly using SVM. There are also other studies that tried to classify gaits into children or adults [12].

With the maturation of deep learning methodologies, researchers have begun utilizing CNN models as frameworks for conducting gait-based age estimation tasks. Sakata et al. [13] employed DenseNet [14] as the backbone, utilizing GEI as inputs, and achieved significant results. Subsequently, Xu et al. achieved state-of-the-art by inducing the uncertainty of the estimation using a label distribution framework on the CNN-based GEISet [3].

Currently, researchers are leveraging the mechanisms of ViT in gait recognition tasks. [15] had trained end-to-end ViT models in performing the GEI classification tasks and had achieved great results in comparison to the CNN methods. Nevertheless, all of these models are end-to-end and need to be trained from scratch. In our work, we aim to find a way to train simpler models with similar performance.

Vision Transformers Transformers [8] have recently achieved tremendous success in the field of Natural Language Processing

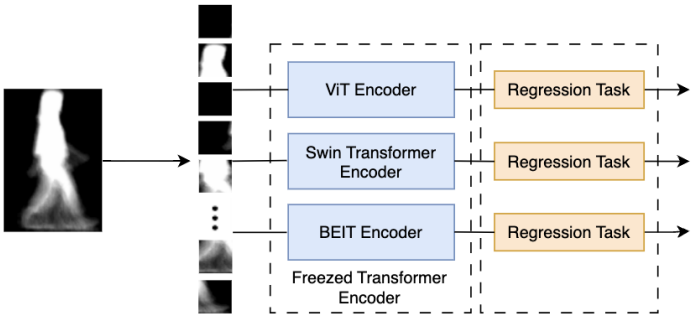

Fig. 2: We use the freeze pre-trained ViT backbones to encode the GEI and train a fully connected layer for age estimations.

(NLP). Inheriting the ideas of self-attention and tokenization, vision transformer [9] (ViT) and its variants have also achieved great results in the field of image classification. One outstanding variant is the Swin Transformer [16]. It brings greater efficiency by limiting self-attention computation to non-overlapping local windows, while also allowing for cross-window connection by shifting the attention windows.

The concept of self-supervised learning (SSL) has been leveraged to exploit large amounts of unlabelled or weakly-labeled training sets. The BEiT v2 [17], following the idea of BERT [18], can perform self-supervised learning using image token masking and a teacher model.

In our study, we train regression heads on the pre-trained ViT model and its variant models, leveraging the assertion that self-attention mechanisms are more adept at capturing abstract features compared to CNNs [9].

3 Method

3.1 Overview

We implemented three distinct ViT architectures as our backbones: Vanilla Vision Transformer, Swin Transformer, and the BEiT v2, each configured to the Base size. Our objective was to devise the most straightforward downstream structure possible, enabling us to evaluate the zero-shot performance of various ViT models effectively.

Following this, we remove the decoder part of the ViT and freeze the pre-trained backbone that was previously pre-trained or fine-tuned on ImageNet21k [19]. The ViT and Swin Transformer were pre-trained using ImageNet1k and BEiT v2 was pre-trained using ImageNet21k, which is a larger dataset, and then further fine-tuned on ImageNet1k. Then we trained a fully connected layer as the regression head that contains no activation function. The output of the regression head is the resulting age. Fig. 2 demonstrates the different backbones and the linear regression. Since the ViT models have not been previously trained on GEI data, this represents a zero-shot learning scenario for the ViT backbones.

3.2 Training

To train the model, we first re-scale the GEI images to the size of 224×224, which is the standard input dimension for the ViT models, and normalize the pixel value of GEI between 0 and 1. Subsequently, we replicate the channels to form a three-channel input and feed this into the ViT models, treating it as a standard RGB image input.

Then we perform batch gradient descent on the last layer of the model by using a Mean Square Error:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (2)$$

N is the batch size where we used 64, y is the predicted age and $\hat{y}$ is the ground truth. During the training, we used an Adam optimizer with a learning rate of 0.0001 and performed the training process on an NVIDIA RTX 3060 graphic card.

Method	MAE	CS(1)	CS(5)	CS(10)
Conventional methods				
MLG[22]	10.98	16.7	43.4	60.8
OPLDA[23]	8.45	7.7	37.9	64.1
OPMFA[23]	9.08	7.0	34.9	64.1
Deep learning methods using CNN				
DenseNet[13]	5.79	22.5	55.9	80.4
GEINet[12]	5.43	23.5	61.7	82.5
Zero-shot using pre-trained vision transformers				
ViT Base[9]	13.6	4.9	25.44	51.6
Swin Base[16]	13.7	5.0	25.4	51.6
BEiTv2 Base[17]	10.0	6.4	32.2	71.0

Table 1: MAE [Years] and CSs [%] at 1-, 5-, and 10-Year Absolute Errors. The first 3 methods are the state-of-art methods of conventional approaches. The two methods in the middle are state-of-art methods for deep CNN approaches and the last three methods are our approaches using zero-shot backbones.

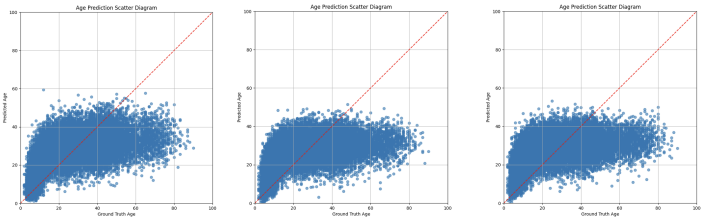

Fig. 3: The scatter plot of predicted age vs. ground truth. From left to right: ViT, Swin Transformer, BEiT v2.

4 Experiment

4.1 Dataset

The dataset we used to train and test our model is the *OU-ISIR Gait Database, Large Population Dataset with Age* [20], also known as the OU-ISIR Age dataset. The dataset contains a total of 63,846 subjects, each labeled with gender and age, with the age range spanning from 2 to 90 years old. This is also the most commonly used dataset in gait age prediction. The dataset was pre-divided into training and testing sets by the authors, adhering to a 5:5 ratio. Additionally, we partitioned the training set into separate training and validation sets by random sampling, adhering to a 4:1 ratio.

4.2 Result

Since only one layer is required to train, it only takes a few epochs for the model to achieve full convergence. Following the evaluations of [3, 13], we also evaluated the performance of all our models by using mean absolute error MAE$= \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$ [Years] and CS$(y) = N(y)/N$ [%], where $N(y)$ is the number of samples whose estimation absolute error is within y years[21]. The MAEs and CSs for 1-, 5-, and 10-year tolerances comparing with different state-of-art methods are summarized in Table 1.

5 Discussion

In Table 1, we compared the proposed methods against the state-of-the-art CNN methods, along with a comparison to the conventional SVR models [22, 23] as baseline models. The result indicates that the pre-trained ViT model approach is not yet able to outperform the existing deep CNN approaches. Furthermore, even the most refined model merely matches the performance of the conventional methods. This is to be expected, as the competing models are specifically designed and trained for GEI age estimation tasks, while ViT models are zero-shot to GEIs.

In order to study why there is such a big gap with the CNN method, we conducted the examination of age-predicted vs. the ground truth,

shown in Fig. 3. It was observed that the model exhibits enhanced precision in identifying individuals below 20 years of age. However, for subjects over 20, the model predominantly forecasts ages within the 20 to 40-year range. Regarding the result, we posit that the discernible discrepancy in body size between children and adults accounts for the pre-trained model's ability to identify these larger gaps effectively, and therefore, a simple head is sufficient to distinguish them. However, the variation in body size and posture among adults across different ages is subtle. Consequently, fine-tuning a small head is insufficient for capturing these finer details.

In comparing various transformer models, we established the baseline by the vanilla ViT base model and assessed the performance of its variants. This comparison aimed to determine whether enhancements designed to bolster the feature extraction capability of ViT could yield improved results in GEI-based age estimation tasks. The comparison shows that there was a significant improvement from the baseline model ViT, to the BEiT model. This implies that despite these pre-trained models not having prior exposure to GEI, they still have the capability to extract relevant features from GEI, which can assist in performing age estimation tasks. It is also worth mentioning that BEiT v2 is pre-trained on ImageNet21k using a self-supervised approach, while the others are pre-trained on the smaller ImageNet1k. This suggests that exposure to a more extensive dataset may contribute to the enhanced performance of BEiT v2. Furthermore, we intend to investigate the influence of the training dataset's size on zero-shot learning outcomes in the context of self-supervised training.

6 Conclusion and Future Works

In this paper, instead of performing gait-based age estimation by building a dedicated CNN deep learning or conventional algorithm, we focused on pre-trained vision transformers. We used a straightforward way to investigate the zero-shot performance of different ViT models, and hence established baselines for fine-tuning on ViT models. In the near future, we will try to leverage more pre-trained models and investigate the application of ViTs in gait recognition.

Acknowledgments

The authors would like to acknowledge support from Aging in Place challenge program at National Research Council Canada (AiP-006).

References

[1] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[2] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2005.

[3] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *2016 International Conference on Biometrics (ICB)*, 2016, pp. 1–8.

[4] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8126–8133.

[5] C. Shen, S. Yu, J. Wang, G. Q. Huang, and L. Wang, "A comprehensive survey on deep gait recognition: algorithms, datasets and challenges," *arXiv preprint arXiv:2206.13732*, 2022.

[6] S. Zhang, Y. Wang, and A. Li, "Gait-based age estimation with deep convolutional neural network," in *2019 International Conference on Biometrics (ICB)*, 2019, pp. 1–8.

[7] C. Xu, A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, Y. Yagi, and J. Lu, "Uncertainty-aware gait-based age estimation and its applications," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 479–494, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[10] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," 2022.

[11] Y. Makihara, H. Mannami, and Y. Yagi, "Gait analysis of gender and age using a large-scale multi-view gait database," in *Computer Vision – ACCV 2010*, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 440–451.

[12] J. W. Davis, "Visual categorization of children and adult walking styles," in *International conference on audio-and video-based biometric person authentication*. Springer, 2001, pp. 295–300.

[13] A. Sakata, Y. Makihara, N. Takemura, D. Muramatsu, and Y. Yagi, "Gait-based age estimation using a densenet," in *Computer Vision – ACCV 2018 Workshops*, G. Carneiro and S. You, Eds. Cham: Springer International Publishing, 2019, pp. 55–63.

[14] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[15] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu, "Gait-vit: Gait recognition with vision transformer," *Sensors*, vol. 22, no. 19, 2022.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[17] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," 2022.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[19] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," 2021.

[20] C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, and J. Lu, "The ouisir gait database comprising the large population dataset with age and performance evaluation of age estimation," *IPSJ Transactions on Computer Vision and Applications*, vol. 9, no. 1, pp. 1–14, 2017.

[21] C. Xu, Y. Makihara, R. Liao, H. Niitsuma, X. Li, Y. Yagi, and J. Lu, "Real-time gait-based age estimation and gender classification from a single image," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3459–3469.

[22] J. Lu and Y.-P. Tan, "Gait-based human age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 761–770, 2010.

[23] J. Lu and Y.-P. Tan, "Ordinary preserving manifold analysis for human age and head pose estimation," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 249–258, 2012.