One Embedding to Predict Them All: Visible and Thermal Universal Face Representations for Soft Biometric Estimation via Vision Transformers

Nelida Mirabet-Herranz, Chiara Galdi, Jean-Luc Dugelay
EURECOM
Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France
{ mirabet, galdi, dugelay } @eurecom.fr

Abstract

Human faces encode a vast amount of information including not only uniquely distinctive features of the individual but also demographic information such as a person's age, gender, and weight. Such information is referred to as soft-biometrics, which are physical, behavioral or adhered human characteristics, classifiable in pre–defined human compliant categories. As we often say 'one look is worth a thousand words'. Vision Transformers have emerged as a powerful deep learning architecture able to achieve accurate classifications for different computer vision tasks, but these models have not been yet applied to soft-biometrics. In this work, we propose the Bidirectional Encoder Face representation from image Transformers (BEFiT), a model that leverages the multi-attention mechanisms to capture local and global features and produce a multi-purpose face embedding. This unique embedding enables the estimation of different demographics without having to re-train the model for each soft-biometric trait, ensuring high efficiency without compromising accuracy. Our approach explores the use of visible and thermal images to achieve powerful face embedding in different light spectra. We demonstrate that the BEFiT embeddings can capture essential information for gender, age, and weight estimation, surpassing the performance of dedicated deep learning structures for the estimation of a single soft biometric trait. The code of BEFiT implementation is publicly available[1]

1. Introduction

Transformer models [24] have boosted the performance of deep learning across various domains in the last years. Traditionally employed in Natural Language Processing (NLP) tasks, attention-based neural networks such as the Vision Transformers (ViTs) are now making significant progress in image-based tasks attaining state-of-the-art (SotA) results on many computer vision benchmarks [26]. While Convolutional Neural Networks (CNNs) have achieved remarkable success in facial processing tasks, they face a fundamental challenge in capturing long-range relationships among different facial regions. To capture long-distance dependencies, the traditional convolution model should enlarge its receptive fields through the stacking of convolutional layers. However, Vision Transformers offers a natural solution to this problem by learning global token dependencies within images [23].

Soft biometric traits are human characteristics typically described using human-understandable labels and measurements. Soft biometrics, such as gender, age, height, weight, ethnicity, hair color, etc., are not unique to the individual but can be aggregated to provide discriminative biometric signatures. Indeed, their use has been proposed in the literature to enhance the performance of traditional biometric systems and enable identification based on human descriptions [6].

As other works before ours have formalized the process of face recognition [1], we provide below a mathematical formulation of soft biometric estimation from face images:

Let $\mathcal{D}$ be an electromagnetic spectral domain composed of a d-dimensional feature space $\mathcal{X} \subset \mathbb{R}^d$ with marginal distribution $\mathbb{P}(\mathcal{X})$ and a label space $\mathcal{Y} \subset \mathbb{N}$. Given a n-face database $X = \{x_i\}_{i=1}^{n}$, where $x_i \in \mathcal{X}$ and their corresponding n-value for each k-biometric trait $Y^k = \{y_j^k\}_{j=1}^{n}$ where $y_j^k \in \mathcal{Y}$ and $k = 1, ..., m$ with $n, m \in \mathbb{N}$. Then a *Face Processing Task* is defined as a parametric function $\mathfrak{F}_{k,\Theta}$ described by the deep learning model parameters Θ where

$$\mathfrak{F}_{k,\Theta} \colon X \times Y^k \longrightarrow [0,1]$$
$$(x_i, y_j^k) \longmapsto \mathbb{P}(Y = y_j^k | X = x_i, \Theta)$$

where $i, j \in [1, n]$ and $k \in [1, m]$. Thus, any facial

[1] https://github.com/nmirabeth/BEFiT/

processing model $\mathfrak{F}_{k,\Theta}$ aims to learn the optimal parameters Θ so that the probability of correctly estimating the trait k for all n identities is 1.

The number of soft biometric traits that can be extracted from a person's face is large, and traditionally, the estimation of each of them requires different networks trained specifically for each soft biometric to be estimated. In this work, we introduce BEFiT, a novel ViT-based facial image processing model that enables the extraction of a unique face embedding useful for the estimation of various soft biometrics without having to know in advance which trait is to be classified. Due to the ability of ViT to learn more nuanced and context-rich features from input images, we demonstrate that by employing BEFiT for extracting a unique embedding from a face, this same embedding can effectively be used for the estimation of various soft biometric traits – as opposed to specializing an end-to-end neural network classifier for the detection of a given soft biometric trait. Moreover, we test the ability of our proposed model to extract face features from both visible- and thermal-spectrum images. Because thermal imagery can provide advantages over visible imaging in challenging conditions such as occlusions, and variations in illumination [13], we explore the potential of BEFiT on visible and thermal face images for enhancing soft biometric estimation tasks, as well as on the fusion of visible and thermal imagery for improved estimation. The main contributions of this work are summarised in the following:

- We propose BEFiT, a vision transformer-based model designed to extract a general face embedding from which different soft biometric traits can be estimated;
- We effectively estimate three distinct traits using the general embedding provided by BEFiT: gender, age, and weight;
- We train BEFiT for visible and thermal face processing and compare their performance against SotA architectures, performing fusion at the score level.

The rest of this paper is organized as follows. In Section 2, we offer a comprehensive review of state-of-the-art methods for soft biometric estimation, along with an introduction to vision Transformers. Section 3 describes BEFiT and how it is employed for extracting the general face embeddings, as well as the fusion protocol, while in Section 4 the experimental setup is detailed. In Section 5 we present the performance analysis of our approach, including the utilization of both visible and thermal imagery, as well as the fusion of scores from both networks, for the estimation of gender, age, and weight. Finally, we conclude with future research directions in Section 6.

Table 1. Overview of soft biometric modalities and key human features within each category.

Permanent*			***Temporal***	
Global	**Face**	**Body**	**Biological**	**Clothing**
Gender Age Weight Height	Eye color Ethnicity Nose type Lip thickness	Arm lenght Wrist size Tattoos	Hair style Hair color Facial hair	Head coverage Clothing color Footwear type Eye glasses

*In this context, "*permanent*" refers to a trait unchangeable over a short period.

2. Related Work

In this section, we present an overview of Transformers applied to computer vision tasks and state-of-the-art methods for facial processing from a person's image, focusing in particular on the estimation of the three soft biometrics considered in this paper: gender, age, and weight.

2.1. Vision Transformers

Since their creation in 2016, Transformers [24] have proven superior to other architectures, such as Recurrent Neural Networks (RNNs), due to their ability to process data in parallel rather than sequentially. By leveraging self-attention mechanisms, Transformers can effectively capture relationships between different parts of input sequences, providing context that might not be discernible through sequential processing as the most relevant image patches for prediction may not necessarily be adjacent to the current one. This allows Transformers to process multiple sequences in parallel, speeding up the process thanks to the parallelization of attention mechanisms.

Vision Transformer [7] have excelled in different computer vision tasks, including image classification [3], object detection [4], and text-to-video translation [29]. Regarding facial processing tasks, some works have successfully applied ViTs in the context of Face Recognition (FR) proving their superiority over other architectures. A first approach was presented in 2021, where Zhong *et al.* proposed a modification to the patch generation process, enabling tokens with sliding patches to overlap with each other, thus enhancing the representation of facial features [28]. More recently, a novel Hybrid tOken Transformer (HOTformer) module was presented, which integrates seamlessly into the traditional ViT architecture, focusing on identifying crucial facial semantics to enhance the effectiveness of people recognition tasks [23]. Subsequently, Kim *et al.* proposed S-ViT, image Relative Positional Encoding as a customized positional encoding in the Transformer encoder [11].

2.2. Facial Soft Biometrics

Soft biometrics, such as gender, age, and weight, are intrinsic to conventional human descriptions [22], as we nat-

urally use these traits to identify and describe each other. Table 1 provides an overview of soft biometric modalities and the associated human traits within each modality. *Permanent* traits, unlike *Temporal* ones, are difficultly modified by a user. In this paper, we specifically put our focus on the estimation of *Permanent* features that can be inferred from face images.

Antipov *et al.* introduced one of the earliest deep learning-based methods for gender estimation [2]. They presented an ensemble model based on CNNs. D' Amelio *et al.* [8] achieved notable success in gender classification from real-world facial images by leveraging features extracted through the VGG-Face Deep Convolutional Neural Network. Age estimation through deep learning models gained traction in 2015 when Wang *et al.* introduced a methodology employing a CNN architecture, followed by linear Support Vector Regression for age estimation [25]. Inspired by them, other researchers explored various CNN architectures [19] as well as ensemble approaches combining multiple models [21] achieving improved performance over prior methods. Weight estimation from face images has been less explored in the literature. A Residual Neural Network (ResNet) with 50 layers and a final regression one, has been employed for this task [15].

Concerning thermal imagery, research has primarily focused on the estimation of gender, ethnicity, and weight from facial images. Deep learning structures began to be explored with a VGG-CNN trained on visible data and tested on thermal faces for gender and ethnicity classification [18]. Farooq *et al.* performed transfer learning from nine renowned architectures to estimate gender from thermal data [9], including ResNet-50, ResNet-101, Inception-V3, MobileNet-V2, VGG-19, AlexNet, DenseNet-121, DenseNet-20, and EfficientNet-B44. They also proposed GENNet for the same task. In a recent study, Mirabet-Herranz *et al.* conducted a comparative analysis of ResNet50's performance on visible and thermal input data for weight estimation [14].

All the approaches presented are end-to-end systems trained for a specific task. Soft biometric traits are numerous, and traditionally, dedicated networks need to be extensively trained for their estimation, sometimes originating from a similar base as transfer learning is typically performed from an analogous task such as face recognition. In this work, we introduce a multipurpose embedding that can be utilized for various tasks, and then we train lightweight CNNs to classify the proposed embedding.

3. Methodology

In this section, we describe vision transformers, which consist of multi-head attention and feed-forward neural networks. Following that, we detail our approach for computing different soft biometric traits from a single embedding and the fusion strategy employed. Additionally, we describe the existing networks adopted and the baselines re-implemented to compare our BEFiT model with other existing networks.

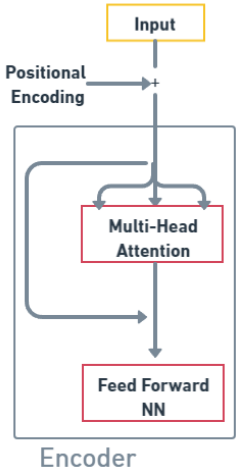

Figure 1. Vision Transformer Encoder

3.1. BEFiT

BEFiT is build upon the architecture of BEiT (Bidirectional Encoder representation from Image Transformers) [3] for image classification tasks. BEiT uses the traditional Transformer [24] as the backbone network. Vision Transformers [7] are a type of deep learning model that extends the Transformer architecture, originally designed for natural language processing tasks, to handle computer vision tasks such as image classification, object detection, and segmentation. BEiT enhances the performance of other vision transformers by introducing a masked image modeling task for pretraining.

In Figure 1 we present the basic architecture of vision transformers. To be processed by the vision transformer, each image is divided into fixed-size patches. Positional encodings are added to the patches to provide spatial information about the position of each patch in the image. Transformers operate through sequence-to-sequence learning, where the transformer takes a sequence of tokens (in our case, image patches) and predicts the next element in the output sequence. This process iterates through the encoder layers, with each layer generating encodings that define the relevance of each part of the input sequence to others, which are then passed to the next encoder layer.

The main advantage of transformers is the self-attention mechanism. The patch embeddings, along with their positional encodings, are fed into the self-attention mechanism where each patch embedding attends to all other patch embeddings, including itself, to compute a weighted sum representation of the entire image. In addition, vision

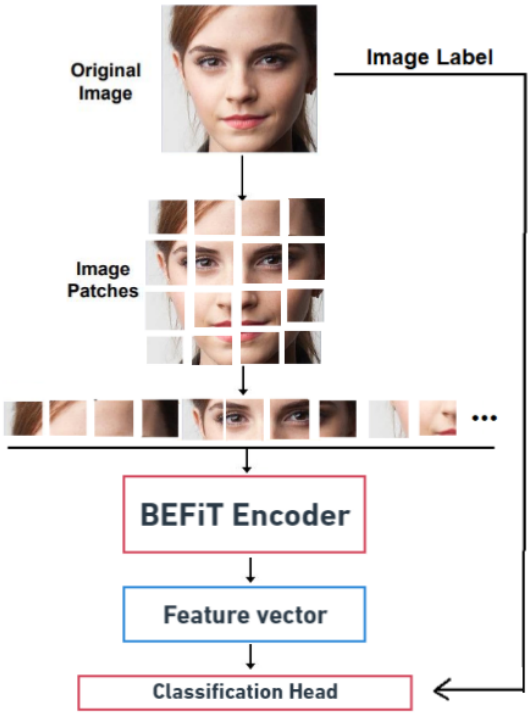

Figure 2. Overview of BEFiT training.

transformers employ multi-head attention, where the self-attention mechanism is performed multiple times in parallel with different sets of learned parameters. This allows the model to attend to different aspects of the input image simultaneously and learn diverse spatial relationships.

After the self-attention mechanism, the output is passed through position-wise Feed Forward Networks (FFNs). FFNs consist of two fully connected layers with a non-linear ReLU activation function applied in between. These layers help capture spatial features within individual patches, allowing the model to encode local information in the image such as edges, textures, and shapes.

In computer vision tasks like image classification where the goal is to predict a single output based on the input image, only the encoder is required. As depicted in Figure 2, by discarding the decoder, the input image patches are fed to the Transformer which produces a fixed-size embedding representing the entire image. This embedding is then fed into a classification head to make predictions.

3.2. Soft Biometric Estimation

Figure 3 depicts the pipeline for soft biometric estimation. Fine-tuning in deep learning refers to taking a pre-trained neural network and further training it on a new dataset for a different task. We fine-tune BEiT for FR on RGB face images obtaining BEFiT model. After that, BEFiT is fine-tuned once more with thermal faces to perform FR in thermal spectra. We will refer to the different versions of BEFiT as BEFiT-V and BEFiT-T depending on the spectra in which they work.

Once the models are trained, the classification head is removed and new face images are passed through BEFiT encoder to obtain the learned feature vector as presented in Figure 2. This general face embedding can be now used for estimating different facial traits without the need to re-training BEFiT.

Afterwards, customized CNNs are defined. These networks take as input the embeddings obtained with BEFiT and are trained to classify the three soft biometric traits studied in this paper. For binary traits such as gender, we define a CNN consisting of a dense layer with 64 units and ReLU activation followed by an output layer with 1 unit and sigmoid activation for binary classification. For regression traits, the CNNs architecture consists of a sequential stack of two fully connected (dense) layers: the first layer has 64 units and uses ReLU activation function, and the second layer has a single unit with ReLU activation function, which outputs the predictions.

Fusion is performed at the decision level. It is applied to scores for classification tasks and to predictions for regression tasks. A weighted average is computed in each case.

3.3. Baselines

In Section 5, we test BEFiT against other state-of-the-art networks for soft biometrics estimation. Unlike our universal face embedding, each of the other networks tested has been trained for a specific trait estimation.

Several baselines were defined by re-implementing SotA methods. VGG architecture has been proven in the literature as powerful for estimating gender and age from face images [10]. Moreover, in their comparative study of architectures for gender estimation from thermal data, Farooq *et al.* [9] revealed the high performance of VGGNet for this task. No study, to the authors' knowledge, has been conducted on the feasibility of thermal imagery for age estimation. Therefore, we select the VGG network with 16 weight layers, i.e., the VGG16 model, for our gender and age baseline estimation models. We use the VGG16 base model as a feature extractor and we add custom fully connected layers on top for binary classification and regression for gender and age prediction respectively. Regarding weight estimation from thermal faces, a ResNet50 architecture is selected [14].

In addition, we test three publicly available and largely trained SotA networks. Those networks estimate the different soft biometric traits from images in the visible domain. For gender classification, we adopt the open-source *DeepFace*[2] library. *DeepFace* provides the most popular pre-trained models for face detection and FR along with its

[2]https://github.com/serengil/deepface

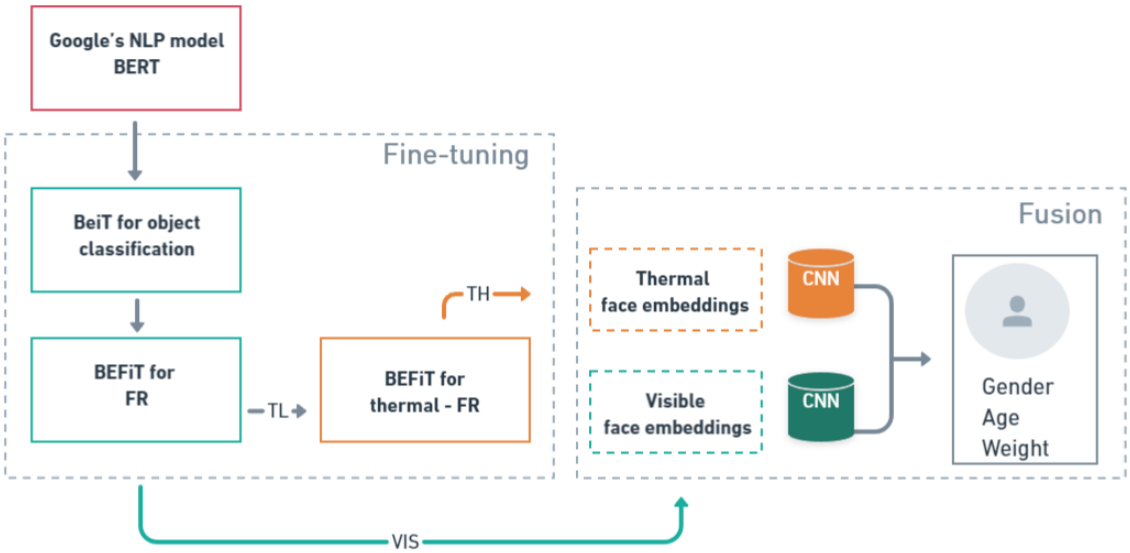

Figure 3. Fine-tuning pipeline for soft biometric estimation using BEFiT. Given any input face image, BEFiT-V and BEFiT-T compute a general face embedding. These embeddings serve as the foundation for estimating three key soft biometric traits: gender, age, and weight.

own models for gender classification. It returns the labels "man", "woman", and associated probabilities, once a human face is passed to the gender model. For age estimation, Deep EXpectation (DEX)[3] [21] model is adopted. DEX is a model for apparent age estimation based on an ensemble of CNNs with VGG-16 architecture pre-trained on ImageNet. Finally, for weight estimation, we use the ResNet50 implementation proposed by Mirabet-Herranz *et al.* [15].

4. Experimental setup

This section presents the databases employed in our experiments, as well as our training and testing protocols. Several metrics are reported in our experiments to facilitate comparability with future works. Finally, because reproducibility is essential for future studies, we provide all our model implementation details.

4.1. Databases

The training of BEFiT-V is performed using the CelebA database [12]. The CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images from more than 10K unique identities. The images in this dataset cover large pose variations and background clutter. BEFiT-T is fine-tuned on the TUFTS database [20], which was presented in 2018. The TUFTS database is composed of more than 10K images, including imagery from different modalities, namely visible, thermal, near-infrared, computerized facial sketch, and 3D images of each volunteer's face.

[3]https://github.com/siriusdemon/pytorch-DEX

Figure 4. Example images from the LVT Face Database. The three variations are displayed in visible (upper row) and thermal (bottom row) spectra, from left to right: N, O and A.

As described in Section 3.2, custom CNNs are defined to estimate each task from the BEFiT embeddings. Different databases are used to train the visible CNNs. The gender CNN is trained using the CALFW dataset [27]. The Cross-Age Labelled Faces in the Wild (CALFW) is an improved version of the LFW face dataset by adding face pairs with age gaps to incorporate the aging process intra-class variance while maintaining the same identities as in the LFW dataset. The CALFW dataset contains 4,025 individuals with 2, 3, or 4 images for each person. We use the AgeDB [17] database to train the age CNN. AgeDB is a manually collected database with a wide range of ages for each subject, comprising 568 identities with 29 images per subject. The weight CNN is trained on the VIP attribute dataset [5], which consists of facial images annotated for gender, height, weight, and Body Mass Index (BMI). This

dataset, collected from the web, includes 1,026 frontal face images of celebrities.

A limited number of publicly available databases contain thermal imagery, and among them, their annotation is minimal. To the authors' knowledge, one database exists with paired visible-thermal facial images annotated with gender, age, and weight: The LVT Face Database for face biometrics [14]. The LVT database is composed of a compendium of images, videos, soft biometrics, and health parameters recorded from 52 different subjects in two sessions. It contains 612 face and 416 shoulder images and videos, respectively, with three different conditions: Neutral, Ambient light, and Occlusion in the form of eyeglasses. Example images from the LVT database are shown in Figure 4.

To provide a fair comparison between the visible and thermal networks, we perform a subject-exclusive split of the LVT database: Training set (480 images from 40 subjects) and Testing set (120 images from the remaining 12 subjects). The thermal gender, age, and weight CNNs are trained on the LVT training set. The baselines are analogously trained on LVT training set. All the models are tested on the LVT test set.

A second test dataset is selected, for cross-database assessment, which is much more challenging than LVT, namely the VIS-TH database [13]. This database consists of 2100 paired visible-thermal images captured under challenging conditions, including variations in expressions, head poses, occlusions, and different illuminations. It encompasses 50 subjects of diverse age, sex, and ethnicity. However, only gender and age information is provided.

4.2. Metrics

Accuracy is used as a metric for gender classifier assessment. Regarding age and weight, we report the Mean Absolute Error (MAE) and Mean Root Square Error (MSRE) in years and kilograms (kg), respectively, and the Pearson's correlation coefficient (ρ). Additionally, for age, we provide the Standard Deviation (StD) of the difference between the predicted and the real age of the subjects. Finally, we include the Percentage of Acceptable Predictions (PAP) for the weight estimation network. This metric represents the percentage of predictions with an error smaller than 10% of the initial weight, indicating a reasonable error in medical applications.

4.3. Implementation Details

We initialize the training of BEFiT-V for face recognition with BEiT pre-trained values obtained from HuggingFace[4]. BEiT was pre-trained on ImageNet-22k, a collection of 14 million images and 22K classes. Each 224 $\times$ 224 image is divided into fixed-size patches of size 16 $\times$ 16. BEFiT-V and BEFiT-T were trained for 150 epochs in the CelebA and TUFTS databases respectively with a batch size of 32, a learning rate of 0.002 and weight decay set to 0.05. In our experiments, we used 2 Nvidia GeForce RTX 2080 Ti 11GB cards with CUDA 11.2. For training BEFiT-V, the training runtime was 3.5 days, with 69.5 samples trained per second and 1.087 train steps per second. For fine-tuning BEFiT-T, the training runtime was 0.8 hours, with 24.5 samples trained per second and 0.389 train steps per second.

For soft biometric estimation, once the embedding was extracted, they were passed to customized CNNs trained for 20 epochs with a batch size equal to 32, Adam optimizer, and a learning rate of 0.001. The loss function chosen was binary cross-entropy for gender classification and MAE for age and weight regression.

Fusion was performed at the decision level. A weighted average of the scores (classification) and estimations (regression) was computed, with weights α for the visible and $1-\alpha$ for the thermal. The values of α were set to 0.5, 0.2, and 0.7 respectively for gender, age, and weight networks.

The weights of the VGG16 and ResNet50 baselines were initialized with pre-trained weights obtained from the ImageNet and UTK dataset respectively. The input images were resized to 224 $\times$ 224 pixels. The VGG16 networks were trained for 20 epochs using the Adam optimizer with a learning rate of 0.001. Binary cross-entropy loss function was selected for gender classification, while mean squared error was employed for age estimation. Each ResNet50 model was re-trained during 10 epochs followed by an additional 10 epochs for training the final regression layer. Adam optimizer was used with a learning rate of 0.01, and Huber loss function was selected with $\delta = 1$. Regarding the fused VGG16 networks, analogously to BEFiT, a weighted average of the scores (classification) and estimations (regression) was computed, with weights α for the visible and $1-\alpha$ for the thermal. The values of α were set to 0.5, 0.1, and 0.7 respectively for gender, age, and weight networks.

5. Results

In this section, we present the experimental results of our proposed BEFiT model for the estimation of gender, age, and weight from face images in two spectra: visible and thermal. We compare the performance of a unique embedding for estimating soft biometrics in contrast to specialized architectures.

5.1. Gender, age, and weight estimation on LVT

Gender estimation: In Table 2, we present the accuracy of the different approaches for gender classification. When comparing VGG16 and BEFiT for both thermal and visible spectra, we observe an advantage of using RGB images for predicting this trait. However, the superiority of BEFiT for extracting gender is clear, with BEFiT-V correctly classifying 95% of the subjects in the LVT test set. Moreover,

[4]https://huggingface.co/docs/transformers/model_doc/beit

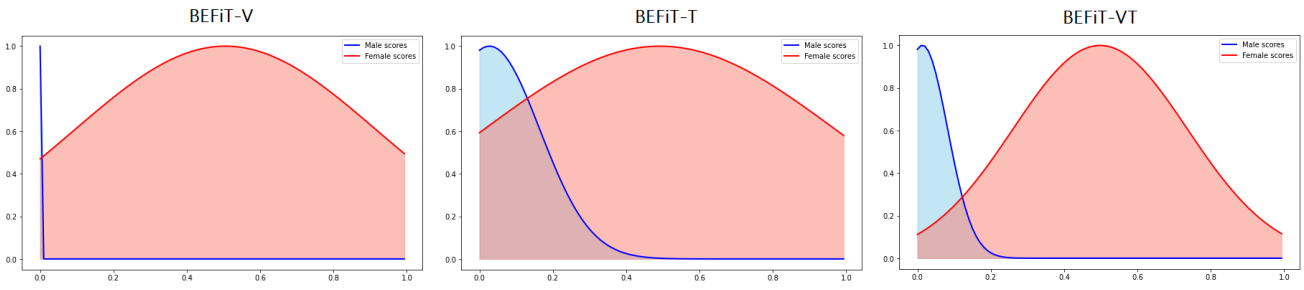

Figure 5. Scores distribution for male and female classes for gender classification via BEFiT model.

Table 2. Evaluation of the gender classification models in the LVT test set for different input data modalities.

GENDER	**Visible**			**Thermal**		**Fusion**	
	Deepface	**VGG16***	**BEFiT-V**	**VGG16***	**BEFiT-T***	**VGG16***	**BEFiT-VT***
Accuracy	0.79	0.81	0.95	0.77	0.86	0.82	**0.97**

by fusing the scores provided by BEFiT-V and BEFiT-T, BEFiT-VT achieves 97% correct classification.

In Figure 5, the distributions of male and female scores obtained with the different versions of BEFiT are presented. In our training, the male class was set to zero and the female to one. We can observe that the female class has scores spread along the entire interval, while the male class reports scores very close to zero, especially remarkable in the case of BEFiT-V. Fusing both spectra allows for greater separability between the classes, consistent with the results of Table 2, where the accuracy of BEFiT-VT for the classification task is higher.

Previous research has highlighted that male and female bodies have different bone mineral and muscle densities, which results in differing facial appearances even when individuals are of the same weight [15]. Therefore, although thermal imagery alone may not suffice for gender classification, it provides crucial information that complements visible input data.

Age estimation: Table 3 presents various metrics assessing the different age estimators. Contrary to the results for gender classification, thermal imagery surpasses RGB for both architectures, VGG16 and BEFiT. The fusion strategy seems especially beneficial for BEFiT-VT, able to gather information from both spectra achieving the lowest errors in the LVT test set. The MAE of BEFiT-VT, at 3.69, is half that of the one presented by the SotA estimator DEX. In contrast to this behavior, it can be observed that in the case of VGG16, thermal predictions are generally penalized by their visible counterpart, resulting in less accurate predictions.

The advantage of thermal data for facial processing is supported by the findings in medical research, where it has been shown that bone, muscle, and body fat do not conduct temperature equally [16]. Heat emission patterns can be utilized to characterize a person as they provide information about the location of major blood vessels, skeleton thickness, amount of tissues, and muscle and fat distribution [14].

Weight estimation: In Table 4, the results of the comparative study of different techniques for weight estimation are displayed. In this case, the results indicate that training a dedicated network delivers more accurate results than extracting weight from a general face embedding. The superiority of thermal data is also confirmed for this task.

When fusing the decisions in BEFiT-VT, the network achieves competitive performance with the state-of-the-art ResNet50. Indeed, BEFiT-VT has the lowest PAP in the LVT test set and competitive results in terms of MAE and RSME. The fusion strategy is also optimal for the ResNet50 networks, achieving the lowest MSRE and higher correlation coefficient in the LVT test set.

5.2. Gender and age estimation on VIS-TH

To assess the generalisation of BEFiT, we have tested gender and age estimation on a more challenging dataset, because of face pose, expression, and illumination variation. Table 5 presents the performance of BEFiT and the SotA methods on the VIS-TH database. Again, BEFiT-VT performs best for gender estimation despite the more complex conditions whereas VGG has a big drop in performance.

As for age, by observing the results, we can confirm that thermal data has an advantage over visible data. Moreover, the fusion strategy proves to be the most successful once more for both architectures, BEFiT and VGG16. BEFiT-VT and DEX have similar performances, with a drop compared

*Trained on LVT training set

Table 3. Evaluation of the age estimation models in the LVT test set for different input data modalities.

AGE	**Visible**			**Thermal**		**Fusion**	
	DEX [21]	**VGG16***	**BEFiT-V**	**VGG16***	**BEFiT-T***	**VGG16***	**BEFiT-VT***
StD	8.69	7.04	9.50	6.45	5.56	6.50	**5.21**
MAE	7.23	5.83	8.41	3.94	4.36	4.11	**3.69**
MSRE	9.12	8.82	10.70	7.33	6.33	7.45	**5.40**
Correlation	0.53	0.28	0.31	0.34	0.45	0.32	**0.55**

Table 4. Evaluation of the weight estimation models in the LVT test set for different input data modalities.

WEIGHT	**Visible**			**Thermal**		**Fusion**	
	ResNet50 [15]	**ResNet50***	**BEFiT-V**	**ResNet50***	**BEFiT-T***	**ResNet50***	**BEFiT-VT***
MAE	**8.13**	10.11	11.29	8.18	11.12	9.11	9.16
MSRE	11.26	12.97	13.48	12.36	16.03	**10.18**	10.76
Correlation	0.57	0.39	0.31	0.76	0.18	**0.74**	0.37
PAP	53%	35%	33%	36%	**60%**	36%	**60%**

Table 5. Evaluation of the gender and age estimation models in the VIS-TH database.

GENDER	**Visible**			**Thermal**		**Fusion**	
	Deepface	**VGG16**	**BEFiT-V**	**VGG16**	**BEFiT-T**	**VGG16**	**BEFiT-VT**
Accuracy	0.84	0.60	0.93	0.33	0.87	0.28	**0.97**
AGE	**Visible**			**Thermal**		**Fusion**	
	DEX [21]	**VGG16**	**BEFiT-V**	**VGG16**	**BEFiT-T**	**VGG16**	**BEFiT-VT**
StD	5.87	6.64	9.20	5.21	6.88	**5.19**	5.89
MAE	4.95	5.16	7.71	4.66	8.75	**4.55**	5.42
MSRE	6.17	6.67	9.44	5.49	10.73	**5.40**	6.89
Correlation	**0.47**	0.06	0.40	0.03	0.15	0.06	0.39

with the results on LVT. VGG16 achieves the best results in terms of MAE and MSRE, however, the low correlation coefficient obtained in each spectrum for the VGG16 architecture reflects that the predictions given by this architecture are always close to the dataset's average age, resulting in minimized age error without learning specific face features for age estimation.

6. Conclusion

Previous work on soft biometric estimation requires specialized networks for each soft biometric trait to be estimated. As an alternative, approaches such as multi-task learning are proposed, but their performance comes at the cost of network complexity. In addition, many soft biometric traits can be estimated from the face; consequently, multiple training sessions need to be done. To tackle this problem, in this paper, we introduce a novel structure for face embedding extraction: BEFiT. BEFiT is a vision transformer that can extract a unique face embedding from which different soft biometric traits can be estimated. Unlike other approaches for soft biometric estimation, the training of BEFiT for face feature vector extraction was not optimized for a specific soft biometric trait estimation, thus boosting embedding generalization. We train two different versions of BEFiT (BEFiT-V and BEFiT-T) in the visible and thermal spectra, and we compare their performance with state-of-the-art networks and baselines. Additionally, we perform fusion at the decision level, enhancing the performance of the soft biometric estimators by gaining insights from both visible and thermal spectra. Our experimental results demonstrate that the embeddings extracted from BEFiT-V and BEFiT-T achieve competitive performance with the ones extracted from dedicated architectures for gender, age, and weight estimation. Furthermore, our fusion strategy successfully estimates the three traits considered, achieving state-of-the-art performance for gender and age on the LVT database.

Acknowledgment

This work has been partially supported by the European CHIST-ERA program (grant agreement CHIST-ERA-19-XAI-011).

References

[1] David Anghelone, Cunjian Chen, Arun Ross, and Antitza Dantcheva. Beyond the visible: A survey on cross-spectral face recognition. *arXiv preprint arXiv:2201.04435*, 2022. 1

[2] Grigory Antipov, Sid-Ahmed Berrani, and Jean-Luc Dugelay. Minimalistic cnn-based ensemble model for gender prediction from face images. *Pattern recognition letters*, 70:59–65, 2016. 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 2, 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[5] Antitza Dantcheva, Francois Bremond, and Piotr Bilinski. Show me your face and i will tell you your height, weight and body mass index. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3555–3560. IEEE, 2018. 5

[6] Antitza Dantcheva, Carmelo Velardo, Angela D'angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification: New trends and challenges. *Multimedia Tools and Applications*, 51:739–777, 2011. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[8] Alessandro D'Amelio, Vittorio Cuculo, and Sathya Bursic. Gender recognition in the wild with small sample size-a dictionary learning approach. In *International Symposium on Formal Methods*, pages 162–169. Springer, 2019. 3

[9] Muhammad Ali Farooq, Hossein Javidnia, and Peter Corcoran. Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system. *Journal of Electronic Imaging*, 29(6):063004–063004, 2020. 3, 4

[10] Dipesh Gyawali, Prashanga Pokharel, Ashutosh Chauhan, and Subodh Chandra Shakya. Age range estimation using mtcnn and vgg-face model. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2020. 4

[11] Geunsu Kim, Gyudo Park, Soohyeok Kang, and Simon S Woo. S-vit: Sparse vision transformer for accurate face recognition. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 1130–1138, 2023. 2

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[13] Khawla Mallat and Jean-Luc Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In *International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September*, LNI, pages 199 – 206. GI / IEEE, 2018. 2, 6

[14] Nelida Mirabet-Herranz and Jean-Luc Dugelay. Lvt face database: A benchmark database for visible and hidden face biometrics. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2023. 3, 4, 6, 7

[15] Nelida Mirabet-Herranz, Khawla Mallat, and Jean-Luc Dugelay. New insights on weight estimation from face images. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023. 3, 5, 7, 8

[16] MJ Morley. Thermal conductivities of muscles, fats and bones. *International Journal of Food Science & Technology*, 1(4):303–311, 1966. 7

[17] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 5

[18] Neeru Narang and Thirimachos Bourlai. Gender and ethnicity classification using deep learning in heterogeneous face recognition. In *2016 International Conference on biometrics (ICB)*, pages 1–8. IEEE, 2016. 3

[19] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5285–5294, 2018. 3

[20] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A Taylor, Arash Samani, et al. A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):509–520, 2018. 5

[21] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 3, 5, 8

[22] Sina Samangooei, Baofeng Guo, and Mark S Nixon. The use of semantic human description as a soft biometric. In *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7. IEEE, 2008. 2

[23] Weicong Su, Yali Wang, Kunchang Li, Peng Gao, and Yu Qiao. Hybrid token transformer for deep face recognition. *Pattern Recognition*, 139:109443, 2023. 1, 2

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3

[25] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. Deeply-learned feature for age estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 534–541. IEEE, 2015. 3

[26] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 1

[27] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 5

[28] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021. 2

[29] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 2