Joint Person Identity, Gender and Age Estimation from Hand Images using Deep Multi-Task Representation Learning

Nathanael L. Baisa

***Abstract*—In this paper, we propose a multi-task representation learning framework to jointly estimate the identity, gender and age of individuals from their hand images for the purpose of criminal investigations since the hand images are often the only available information in cases of serious crime such as sexual abuse. We investigate different up-to-date deep learning architectures and compare their performance for joint estimation of identity, gender and age from hand images of perpetrators of serious crime. To simplify the age prediction, we create age groups for the age estimation. We make extensive evaluations and comparisons of both convolution-based and transformer-based deep learning architectures on a publicly available 11k hands dataset. Our experimental analysis shows that it is possible to efficiently estimate not only identity but also other attributes such as gender and age of suspects jointly from hand images for criminal investigations, which is crucial in assisting international police forces in the court to identify and convict abusers. The source code is available at https://github.com/nathanlem1/IGAE-Net.**

***Index Terms*—Person identification, Gender estimation, Age estimation, Deep representation learning, Multi-task learning.**

I. Introduction

Recognition of individuals using body parts or behavioural characteristics, often termed as biometric identitification [1], [2], has recently attracted remarkable attention for numerous applications. The primary biometric traits, also called biometric modalities, used for the identification of an individual include face [3], body [4], hand [5], fingerprint [6], iris [7], gait [8] and voice [9]. Personal attributes or ancillary information such as gender, age, (hair, skin, eye) color, ethnicity, height, weight, etc. which are also referred to as soft biometrics can be inexpensively obtained from the primary biometric traits [10]. The recognition of an individual can be improved by fusing different biometric modalities or by fusing biometric modalities with soft biometrics [11]. Hand images are often the only available information in cases of serious crime such as sexual abuse. This is because abusers usually hide their identity by covering, for instance, their faces; however, their hands often remain visible. Hand images also deliver discriminative features for biometric person recognition. They not only have less variability when compared to other biometric modalities but also have strong and diverse features which remain relatively stable after adulthood [12]–[14]. Because of this, there is a strong potential to investigate hand images captured by digital cameras for persons recognition (preferably with their gender and age predictions), especially for criminal investigations in uncontrolled environments, which is crucial in assisting international police forces in the court to identify and convict abusers.

Full hand images have been used in [5] to identify a person using both traditional and deep learning methods. Similar appraoch has also been used in [13] with additional data type for fusion, near-infrared (NIR) images, in addition to the RGB hand images. However, these methods are not an end-to-end. An end-to-end approaches considering both horizontal and vertical uniform partitioning [12] and multi-branch network with attention mechanism [15] have been proposed using full hand images as input. Knuckle patterns of dorsal hand images have been used in [14], [16], [17] for person identification. The work in [18] has combined fingernails and dorsal knuckle patterns for identifying individuals. Full hand images have been combined with knuckle patterns in [19] for person recognition. Age estimation from hand images is also proposed in [20]. Facial images-based age and gender prediction methods are proposed in [21], [22]; also with additional ethnicity prediction [23] or identity recognition [24]. However, all these hand-based methods recognize only the identity of an individual or only age of an individual from hand images. Unlike these methods, our proposed deep multi-task representation learning framework jointly estimates the identity, gender and age of individuals from their hand images for the purpose of criminal investigations.

In this work, we propose a multi-task representation learning framework to jointly estimate the identity, gender and age of individuals from their hand images for the purpose of criminal investigations. We investigate the performance of different deep learning architectures and compare their performance for joint estimation of identity, gender and age from hand images of perpetrators of serious crime. To simplify the age prediction, we create age groups for the age estimation. Our contributions can be summarized as follows.

1) We propose a multi-task representation learning framework to jointly estimate the identity, gender and age of individuals from their hand images.
2) We investigate different up-to-date deep learning architectures and compare their performance for joint estimation of identity, gender and age from hand images.
3) We make extensive evaluations and comparisons on a publicly available 11k [5] hands dataset.

The rest of the paper is organized as follows. The pro-

Nathanael L. Baisa is with the School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, UK. Email: nathanael.baisa@dmu.ac.uk.

posed method is described in Section II including the overall architecture and the loss function. The experimental results are analyzed and compared in Section III followed by the main conclusion along with suggestion for future work in Section IV.

II. Proposed Method

In this section, we introduce the overall architecture of the identity, gender and age estimation network (IGAE-Net) and the used loss function.

A. Network Architecture Overview

The overall architecture of the proposed IGAE-Net is given in Fig. 1. We use different deep learning architectures as backbone network. We keep the original structure of each deep learning architecture used as backbone before its classifier layer remain the same when we modify each backbone network to produce the IGAE-Net. We create 3 independent branches just before the classifier layer of each backbone network to produce three heads: identity head, gender head and age head. Each head follows the original classifier layer structure of the backbone network. We use both convolution-based and transformer-based up-to-date deep learning architectures as backbone network for detailed performance comparisons.

CNN-based Backbone: For the case of the convolution-based architectures, we use ResNet50 [25], DenseNet121 [26] and ConvNeXt-Tiny [27].

Transformer-based Backbone: For the case of the transformer-based architectures, we use ViT-B-16 [28], Swin-T [29] and MaxViT-T [30].

Each of the identity head, gender head and age head, which acts as classification layer, is implemented independently using a fully-connected (FC) layer (with additional normalization layer(s) for the ConvNeXt-Tiny and MaxViT-T) followed by a softmax function to jointly predict the identity, gender and age group of each input hand image, respectively.

B. Loss Function

The IGAE-Net is optimized during training by minimizing the loss function $\mathcal{L}$ consisting of the sum of cross-entropy losses over the predictions from the 3 heads i.e. the identity head, the gender head and the age head for multi-output classification, as given in Eq. 1. Given the input hand image, classifier from each head predicts its corresponding output i.e. identity, gender and age group of the input image as shown in Fig. 1.

$$\mathcal{L} = \sum_{l=1}^{3} \mathcal{L}_{l,xent} \tag{1}$$

For the learned features $\mathbf{f}_i$, the cross-entropy loss (softmax loss) with label smoothing [31] is given as:

$$\mathcal{L}_{l,xent} = -\sum_{i=1}^{N} q_{y_i} \log \frac{e^{\mathbf{W}_{y_i}^T \mathbf{f}_i + b_{y_i}}}{\sum_{c=1}^{C} e^{\mathbf{W}_c^T \mathbf{f}_i + b_c}} \tag{2}$$

where N is the batch-size, C is the number of classes (identities) in the training dataset, and $\mathbf{W}_c$ and b_c are weight vector and bias for class c, respectively. Note that $z_c = \mathbf{W}_c^T \mathbf{f}_i + b_c$ are the logits or unnormalized probabilities. The ground-truth distribution over labels q_{y_i} by including label smoothing can be given as

$$q_{y_i} = \begin{cases} 1 - \frac{C-1}{C}\epsilon, & \text{if } y_i = y \\ \frac{1}{C}\epsilon, & \text{otherwise} \end{cases} \tag{3}$$

where y is ground-truth label and ϵ is a smoothing value.

III. Experiments

A. Dataset

We use the 11k hands dataset[1] [5] which has 190 subjects (identities) with genders and varying ages between 18 - 75 years old. Any hand image containing accessories is excluded from the data to avoid any potential bias as in [12]. The dataset is then divided into right dorsal, left dorsal, right palmar and left palmar sub-datasets. After excluding accessories and dividing the dataset into the sub-datasets, the right dorsal has 143 identities, the left dorsal has 146, the right palmar has 143 and the left palmar has 151 identities. We take the samples in random order before dividing each sub-dataset into training and test sets equally based on the number of samples (in each class or subject).

The age distribution of the 11k hands dataset (right dorsal as an example) is shown in Fig. 2a. As can be seen in this figure, the largest portion of the images belongs to individuals with age between 20 and 23. The number of images for persons with other ages is very low. This poses a problem to properly learn the ages of individuals, especially by treating the prediction task as a classification task. To alleviate this problem and simplify the age prediction, we group images of persons whose age lies in some range as shown in Fig. 2b. The grouping is done into 6 classes (age groups) as can be observed in Fig. 2b. In this manner, we can predict the age group of the persons instead of their exact age. Accordingly, the identity statistics, gender statistics and age groups statistics of each sub-dataset used in this paper are given in Table I, Table II and Table III, respectively.

TABLE I
Identity statistics of the 11k hands dataset [5] for right dorsal (D-r), left dorsal (D-l), right palmar (P-r) and left palmar (P-l) sub-datasets. Number of identities (ids) and number of images (samples) are shown for each sub-dataset.

	D-r of 11k	D-l of 11k	P-r of 11k	P-l of 11k
# ids	143	146	143	151
# images	2004	1869	1965	2027

B. Implementation Details

We implemented the IGAE-Net using PyTorch deep learning framework and trained it on NVIDIA GeForce RTX 2080 Ti GPU. The input images are resized to 256×256 and then randomly cropped to 224×224, augmented by random horizontal flip, color jittering and normalization during training.

[1] https://sites.google.com/view/11khands

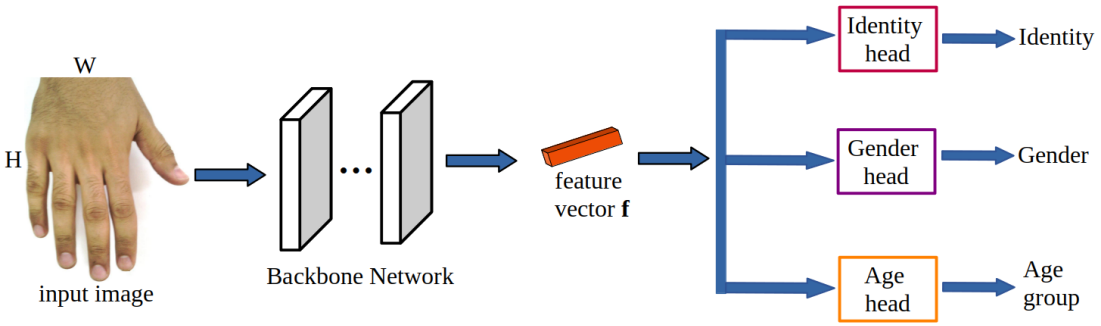

Fig. 1. Structure of IGAE-Net. Given an input hand image, feature vector **f** is obtained by passing it through the backbone network and then is fed into each head (identity head, gender head and age head). The identity head, gender head and age head predict the identity, gender and age group of the input image, respectively.

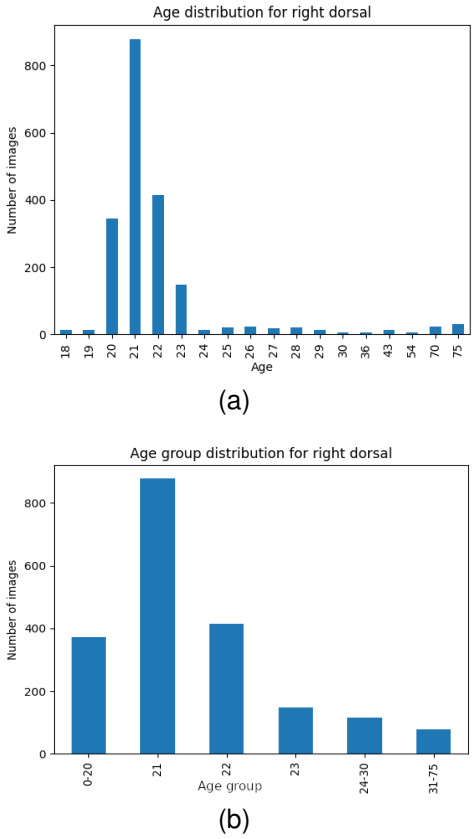

Fig. 2. Age statistics for right dorsal of the 11k hands dataset [5]: (a) Age distribution, (b) Age group distribution. The number of images per age or age group is shown.

TABLE II
Gender statistics of the 11k hands dataset [5] for right dorsal (D-r), left dorsal (D-l), right palmar (P-r) and left palmar (P-l) sub-datasets. Number of images (samples) for male and female are shown.

	D-r of 11k	D-l of 11k	P-r of 11k	P-l of 11k
Male	909	846	948	949
Female	1095	1023	1017	1078

TABLE III
Age groups statistics of the 11k hands dataset [5] for right dorsal (D-r), left dorsal (D-l), right palmar (P-r) and left palmar (P-l) sub-datasets. Number of images (samples) for each age group is shown.

	D-r of 11k	D-l of 11k	P-r of 11k	P-l of 11k
0-20	372	328	371	381
21	878	852	861	890
22	414	362	418	401
23	147	162	148	180
24-30	114	87	95	111
31-75	79	78	72	64

However, only normalization is utilized during testing with the test images resized to 224×224, without a random crop. A random order of images is used by reshuffling the dataset. We use a sum of cross-entropy losses over the predictions of the 3 heads as in Eq. (1) to train the IGAE-Net. To prevent over-fitting and over-confidence, label smoothing [31] with smoothing value (ϵ) of 0.1 is also used with the cross-entropy loss. We train the model for 50 epochs with mini-batch size of 20 and Adam optimizer with the weight decay factor for L2 regulization of 5×10^{-4}. For the first 10 epochs, we use a warmup strategy [32], increasing a learning rate linearly from 8×10^{-6} to 8×10^{-4}, and then it is decayed to 4×10^{-4} after 30 epochs. The learning rate is divided by 10 for the existing layers of the backbone network i.e. ten times bigger learning rate is given to the newly added layers (FC layers and layer normalization(s) for the 3 heads) with default weight and bias initializations.

C. Performance Evaluation

We compare the performance of our proposed IGAE-Net with different up-to-date deep learning architectures as backbone netwowrk. We use both convolution-based and transformer-based deep learning architectures on a publicly available 11k hands dataset [5] for comprehensive analysis using accuracy evaluation metric. The quantitative performance comparison of the IGAE-Net with different backbone networks is given in Table IV. All the backbone networks are pre-trained on ImageNet-1K [33].

CNN-based Backbone: For the case of the convolution-based architectures, we use ResNet50 [25], DenseNet121 [26]

and ConvNeXt-Tiny [27]. As shown in Table IV, ConvNeXt-Tiny-based IGAE-Net averagely outperforms all other CNN-based methods across all the sub-datasets in accuracy evaluation metric.

Transformer-based Backbone: For the case of the transformer-based architectures, we use ViT-B-16 [28], Swin-T [29] and MaxViT-T [30]. As can be observed in Table IV, the best performing transformer-based IGAE-Net is with the Swin-T backbone network when considering efficiency and accuracy. Though ViT-B-16-based IGAE-Net is performing comparably as the Swin-T-based IGAE-Net, the number of parameters of the ViT-B-16 is almost 3 times that of the Swin-T i.e. 86.6M vs 28.3M (M for million).

The majority of the models perform better in gender prediction when compared to the identity and age group predictions. For instance, ConvNeXt-Tiny-based IGAE-Net has scored 100% gender prediction on right dorsal (D-r) of 11k, left dorsal (D-l) of 11k and right palmar (P-r) of 11k hands sub-datasets.

We also showed the confusion matrices for identity outputs, gender outputs and age groups outputs in Fig. 3 on right dorsal of the 11k hands dataset using Swin-T-based IGAE-Net for more detailed analysis of the model's performance. As can be seen on the main diagonal of each confusion matrix, all the cases are predicted as the true label of their class with few exceptions on the identity and age predictions. It is important to note that very few false positive percentage in the age group classification model corresponds to the images belonging to the 21 age group which are mistaken for 0-20 and 22. Similarly, few images belonging to the 0-20 age group are mistaken for 21. The main diagonal of the gender confusion matrix in Fig. 3b confirms the 100% accuracy of our model on the right dorsal of the 11k hands dataset using Swin-T backbone network given in Table IV.

The qualitative results of our proposed method on right dorsal of the 11k hands dataset using Swin-T-based IGAE-Net are also shown in Fig. 4. As can be observed in Fig. 4, the IGAE-Net model predicts correctly for all of the sample images of our test set in terms of identity, gender and age group. The model's predictions (PR) are provided next to the ground-truth labels (GT) of the hand images. For instance, for the first hand image in this figure, GT labels are 0001533 (identity), female (gender) and 22 (age group). The model predicted the correct outputs i.e. 0001533 (identity), female (gender) and 22 (age group).

In this evaluation, we assume a closed setting where the training classes are available in the test set with different images from the training set. In open setting such as in real criminal investigation applications in the court, the learned feature representations just after a fully-connected layer and before a softmax function of each head (shown in Fig. 1) can be used for feature representations comparison using, for instance, cosine similarity metric.

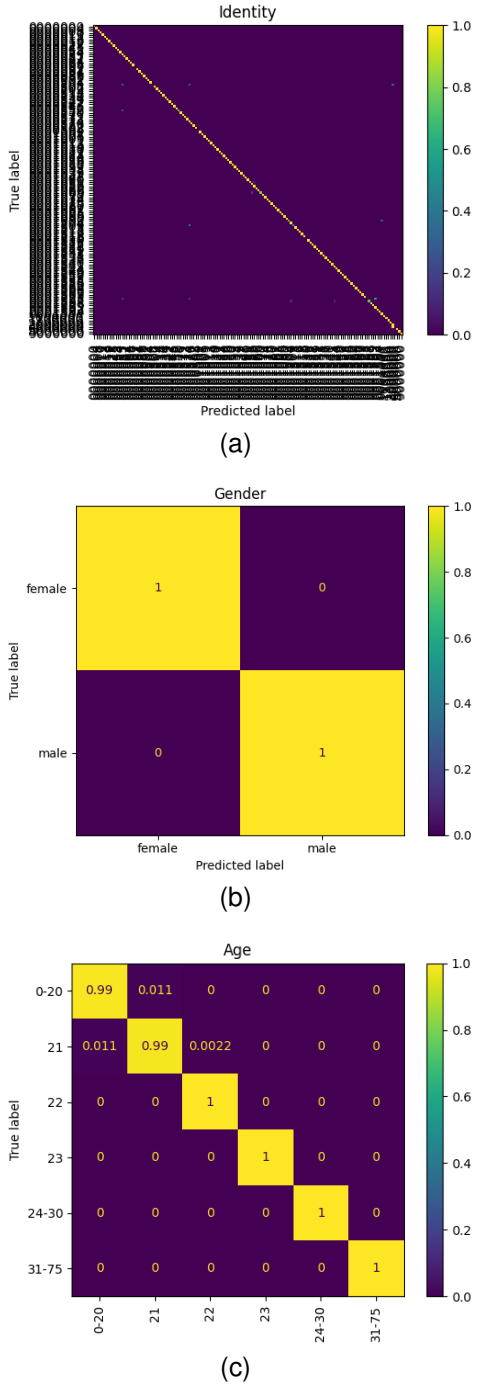

Fig. 3. Confusion matrices on right dorsal of 11k hands dataset [5] using Swin-T [29]-based IGAE-Net: (a) confusion matrix for identity outputs, (b) confusion matrix for gender outputs, (c) confusion matrix for outputs of age groups.

IV. Conclusion

In this work, we introduce a multi-task representation learning framework to jointly estimate the identity, gender and age of individuals from their hand images for the purpose of

TABLE IV
Quantitative performance comparison of our method (IGAE-Net) with different up-to-date deep learning backbone architectures on right dorsal (D-r) of 11k, left dorsal (D-l) of 11k, right palmar (P-r) of 11k and left palmar (P-l) of 11k hands sub-datasets [5]. The results are shown in accuracy (%) with the best results in **bold**. The models are pre-trained on ImageNet1k [33].

	D-r of 11k			D-l of 11k			P-r of 11k			P-l of 11k		
Method	Identity	Gender	Age	Identity	Gender	Age	Identity	Gender	Age	Identity	Gender	Age
ResNet50 [25]	95.82	99.78	95.39	98.83	99.46	99.63	97.45	99.00	98.82	96.61	99.27	94.71
DenseNet121 [26]	98.75	99.61	99.12	99.61	**100.00**	99.82	98.50	97.94	99.33	98.16	98.04	98.40
ConvNeXt-Tiny [27]	98.61	**100.00**	99.78	99.64	**100.00**	99.75	98.99	**100.00**	99.55	98.34	99.81	98.39
ViT-B-16 [28]	98.11	99.80	**99.81**	99.47	**100.00**	99.67	**99.47**	**100.00**	99.67	98.17	99.12	97.36
Swin-T [29]	98.08	**100.00**	99.54	99.47	**100.00**	99.89	99.10	99.92	99.67	97.18	99.84	97.66
MaxViT-T [30]	99.04	**100.00**	98.95	99.40	99.37	98.80	98.21	99.80	98.94	96.13	99.35	93.19
Vmamba	99.50	**100.00**	99.75	99.47	**100.00**	**100.00**	99.24	**100.00**	**99.75**	**99.51**	99.75	**100.00**
Vmamba$_{schedule}$	**100.00**	99.75	99.75	**100.00**	**100.00**	99.73	99.24	**100.00**	99.24	**99.51**	**100.00**	99.75

Fig. 4. Some qualitative results of our proposed method on right dorsal of 11k hands dataset [5] using Swin-T [29]-based IGAE-Net. The ground truth labels (GT) vs the predicted labels (PR) of identity, gender and age group of each hand image, respectively, are shown.

criminal investigations. We investigate our proposed identity, gender and age estimation network (IGAE-Net) with different up-to-date deep learning architectures as backbone network and compare their performance for joint estimation of identity, gender and age from hand images of perpetrators of serious crime such as sexual abuse. To simplify the age prediction, we create age groups for the age estimation. We make extensive evaluations and comparisons of both convolution-based and transformer-based deep learning architectures on a publicly available 11k hands dataset, which demonstrates that it is possible to efficiently estimate the identity as well as other attributes such as gender and age of suspects jointly from their hand images for criminal investigations. From the convolution-based architectures, ConvNeXt-Tiny-based IGAE-Net performs best whereas Swin-T-based IGAE-Net is the best performing transformer-based architecture. These two networks, ConvNeXt-Tiny-based IGAE-Net and Swin-T-based IGAE-Net, give comparable performance, which confirms the competitive nature of the convolution-based and transformer-based architectures for computer vision tasks. Our proposed method is crucial in assisting international police forces in the court to identify and convict abusers.

REFERENCES

[1] Shervin Minaee, AmirAli Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang, "Biometric recognition using deep learning: A survey," *CoRR*, vol. abs/1912.00271, 2019.

[2] Anil K. Jain, Debayan Deb, and Joshua J. Engelsma, "Biometrics: Trust, but verify," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 303–323, 2022.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[4] Nathanael L. Baisa, "Local-aware global attention network for person re-identification," *CoRR*, vol. abs/2209.04821, 2022.

[5] Mahmoud Afifi, "11k hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, 2019.

[6] J. J. Engelsma, K. Cao, and A. K. Jain, "Learning a fixed-length fingerprint representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[7] Z. Zhao and A. Kumar, "Towards more accurate iris recognition using deeply learned spatially corresponding features," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3829–3838.

[8] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 140–154, 2015.

[9] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech 2018*, Sep 2018.

[10] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? A survey on soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 441–467, 2016.

[11] Maneet Singh, Richa Singh, and Arun Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187 – 205, 2019.

[12] Nathanael L. Baisa, Bryan Williams, Hossein Rahmani, Plamen Angelov, and Sue Black, "Hand-based person identification using global and part-aware deep feature representation learning," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2022, pp. 1–6.

[13] Yimin Yuan, Chaoying Tang, Shuhang Xia, Zhou Chen, and Tong Qi, "HandNet: Identification based on hand images using deep learning methods," in *Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing*, New York, NY, USA, 2020, ICVISP 2020, Association for Computing Machinery.

[14] Abdelouahab Attia, Zahid Akhtar, and Youssef Chahir, "Feature-level fusion of major and minor dorsal finger knuckle patterns for person authentication," *Signal, Image and Video Processing*, Feb. 2021.

[15] Nathanael L. Baisa, Bryan Williams, Hossein Rahmani, Plamen Angelov, and Sue Black, "Multi-branch with attention network for hand-based person recognition," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022.

[16] A. Kumar and Z. Xu, "Personal identification using minor knuckle patterns from palm dorsal surface," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2338–2348, 2016.

[17] Ritesh Vyas, Hossein Rahmani, Ricki Boswell-Challand, Plamen Angelov, Sue Black, and Bryan M. Williams, "Robust end-to-end hand identification via holistic multi-unit knuckle recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–8.

[18] Mona Alghamdi, Plamen Angelov, and Bryan Williams, "Automated person identification framework based on fingernails and dorsal knuckle patterns," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 01–08.

[19] Zahra Ebrahimian, Seyed Ali Mirsharji, Ramin Toosi, and Mohammad Ali Akhaee, "Automated person identification from hand images using hierarchical vision transformer network," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2022, pp. 398–403.

[20] Mohamed Ait Abderrahmane, Ibrahim Guelzim, and Abdelkaher Ait Abdelouahad, "Hand image-based human age estimation using a time distributed CNN-GRU," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 2020, pp. 1–5.

[21] AmirAli Abdolrashidi, Mehdi Minaei, Elham Azimi, and Shervin Minaee, "Age and gender prediction from face images using attentional convolutional network," *CoRR*, vol. abs/2010.03791, 2020.

[22] Avishek Garain, Biswarup Ray, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, and Ram Sarkar, "GRA-Net: A deep learning model for classification of age and gender from facial images," *IEEE Access*, vol. 9, pp. 85672–85689, 2021.

[23] Hu Han and Anil K. Jain, "Age, gender and race estimation from unconstrained face images," 2014.

[24] Andrey V. Savchenko, "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output convnet," *PeerJ Comput. Sci.*, vol. 5, pp. e197, 2019.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, jul 2017, pp. 2261–2269, IEEE Computer Society.

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. 2022, pp. 11966–11976, IEEE.

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

[30] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li, "Maxvit: Multi-axis vision transformer," in *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds., Cham, 2022, pp. 459–479, Springer Nature Switzerland.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[32] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2020.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.