

Age Prediction From Face Images Via Contrastive Learning

Yeongnam Chae

Poulami Raha

Mijung Kim

Björn Stenger

Rakuten Institute of Technology, Rakuten Group, Inc.

{yeongnam.chae, poulami.raha, mijung.a.kim, bjorn.stenger}@rakuten.com

Abstract

This paper presents a novel approach for accurately estimating age from face images, which overcomes the challenge of collecting a large dataset of individuals with the same identity at different ages. Instead, we leverage readily available face datasets of different people at different ages and aim to extract age-related features using contrastive learning. Our method emphasizes these relevant features while suppressing identity-related features using a combination of cosine similarity and triplet margin losses. We demonstrate the effectiveness of our proposed approach by achieving state-of-the-art performance on two public datasets, FG-NET and MORPH II.

1 Introduction

Age estimation from facial images has found application in various fields. However, the different attributes in a face image present distinct visual features that can be challenging to disentangle. While facial feature extractors have been pre-trained for specific tasks such as face recognition, they are not well-suited for tasks like age estimation. Collecting a large dataset of face images of the same individuals at different ages is more difficult than collecting a dataset of different individuals. Therefore, the key question is how to develop a method for learning aging-related features that are not influenced by identity-related features.

Face age estimation methods based on convolutional neural networks have made significant progress and can be grouped into three categories: classification [9, 20, 34, 42], regression [12, 13, 14, 25, 30] and ranking [5, 6, 7, 8] approaches. Recently, self-supervised [4] and attention-based approaches [39] have been proposed. However, most of these techniques have relied on information from individual face images, causing the model to be biased towards features associated with attributes like identity, which hinders the model from focusing on relevant but sparse features related to age, such as small wrinkles or skin texture. Moreover, some studies have explored comparative approaches [1, 18, 22, 32], which aim to learn ranking or transformation information based on relative age differences. In contrast, the proposed approach mainly focuses on how to emphasize sparser features related to age while penalizing non-relevant features related to identity through contrasting the features extracted

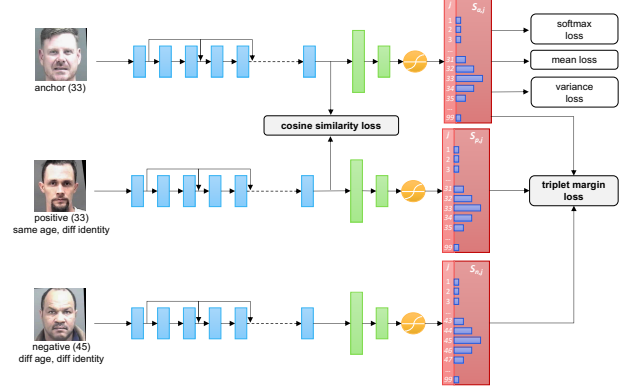


Figure 1: Contrastive Learning for Age Estimation using cosine similarity loss for positive samples with the same age as the anchor image and triplet margin loss for negative samples with a different age.

from images from the same age group.

We present a new approach for face age estimation that leverages contrastive learning. Our method aims to suppress identity-related features while emphasizing age-related ones. To learn identity-independent age features, we use triplets of images. Given an anchor face image, we sample two images - one of the same age (positive sample) but with a different identity, and one of a different age (negative sample). By comparing the anchor image to both the positive and negative samples, we jointly minimize a cosine similarity loss and a triplet margin loss, see Figure 1. Owing to the large number of possible triplet samples, the method is data-efficient and able to learn age prediction from small datasets without resorting to additional data sources, as shown in the experiments on public datasets.

2 Method

We use contrastive learning to suppress identity-related features in order to compare face images with the same age but different identity. We extract the facial feature vector from an anchor image, \mathbf{x}_a , and a positive sample, \mathbf{x}_p . To ensure that the positive sample \mathbf{x}_p of the given anchor \mathbf{x}_a is free from identity-related features, we select it from a set of face images $\mathcal{P}(\mathbf{x}_a)$ that have the same age but different identity labels:

$$\mathcal{P}(\mathbf{x}_a) = \{\mathbf{x}_p \mid y_p = y_a \text{ and } i_p \neq i_a\}, \quad (1)$$

where i_a and i_p are the identity labels and y_a and $y_p \in \{1, 2, \dots, A\}$ are age labels of the anchor and the positive sample, respectively. A is the maximum age label in a dataset. We extract ResNet-18 features $\mathbf{f}_a = R(\mathbf{x}_a)$, $\mathbf{f}_p = R(\mathbf{x}_p)$, where $\mathbf{f} \in \mathbb{R}^D$ are features before the last fully connected (FC) layer. $D = 512$ is the dimension of the feature vector. We compute the cosine similarity loss L_c [29] between features \mathbf{f}_a and \mathbf{f}_p :

$$L_c(\mathbf{f}_a, \mathbf{f}_p) = \frac{\mathbf{f}_a^T \mathbf{f}_p}{\|\mathbf{f}_a\| \|\mathbf{f}_p\|}. \quad (2)$$

The cosine distance has been widely used to compare face images or features [10, 17, 19, 24, 27, 40].

The probabilities computed in the final softmax output of the network for the anchor image define a distribution over age values. Let $s_{a,j}$ denote the probability that sample \mathbf{x}_a has age label j . We compute a mean loss, L_m , and a variance loss, L_v [31]:

$$L_m(\mathbf{s}_a) = \sum_{j=1}^A j s_{a,j} - y_a, \quad (3)$$

$$L_v(\mathbf{s}_a) = \sum_{j=1}^A s_{a,j} \left(j - \sum_{k=1}^A k s_{a,k} \right)^2. \quad (4)$$

Ternary contrast via triplet margin loss. To use contrastive learning for different ages, we adopt a triplet margin loss [37]. The triplet margin loss is a loss function where the anchor image is compared with a positive and a negative sample. The distance from the anchor to the positive sample is minimized, while the distance to the negative sample is maximized. We use a negative sample \mathbf{x}_n with a different age and identity to contrast with the positive sample and the anchor.

$$\mathcal{N}(\mathbf{x}_a) = \{\mathbf{x}_n \mid y_n \neq y_a \text{ and } i_n \neq i_a\} \quad (5)$$

The triplet loss is applied to the softmax probability, \mathbf{s} as follows:

$$L_t(\mathbf{s}_a, \mathbf{s}_p, \mathbf{s}_n) = \max(\|\mathbf{s}_a - \mathbf{s}_p\|^2 - \|\mathbf{s}_a - \mathbf{s}_n\|^2 + \alpha, 0), \quad (6)$$

where \mathbf{s}_n is softmax probability of the negative image and α is a margin between positive and negative samples. The overall loss function is defined as:

$$L_{\text{total}} = L_s + \lambda_m L_m + \lambda_v L_v + \lambda_c L_c + \lambda_t L_t, \quad (7)$$

where L_s denotes the softmax loss, λ denote hyperparameters that balance the influence of each loss term. We empirically set λ_m to 0.2 and λ_v to 0.05. We experimentally evaluate the values of the binary and ternary loss terms, λ_c and λ_t .

Table 1: Age distributions in MORPH II and FG-NET.

Dataset	0-19	20-39	40-59	>60
MORPH II	7,469	31,682	15,649	334
FG-NET	710	223	61	8

3 Experiments

Datasets & protocols. The *MORPH II* dataset is a face dataset, containing 55,134 images of 13,618 individuals. Ages range from 16 to 77 with a median age of 33. To be consistent with prior work, the five-fold random split (RS) and five-fold subject exclusive (SE) protocols are used in the experiments [35].

The *FG-NET* dataset contains 1,002 face images from 82 individuals with ages ranging from 0 to 69 years [33]. We evaluate using the commonly used leave-one-person-out (LOPO) protocol. Table 1 shows the age distributions in the MORPH II and FG-NET datasets.

Implementation details & evaluation metric. We first perform facial alignment of all images using five landmarks detected with MTCNN [44]. These aligned face images are then normalized to a size of $256 \times 256 \times 3$. Subsequently, we extract features from the normalized face images by utilizing a ResNet-18 model that has been pre-trained on ImageNet. We apply a series of augmentation techniques, including random affine transformations with slight variations, random vertical flips, and random crops to 224×224 during the training phase. We optimize the model parameters using the Adam optimizer with an initial learning rate of 0.001 and train for 100 epochs, using a batch size of 64. As error metric we use the mean absolute error (MAE), which is defined as the L1 distance between the predicted age, \hat{y}_i , of image \mathbf{x}_i and its ground-truth age, y_i :

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|. \quad (8)$$

Results on MORPH II. Table 2 shows the results following the RS protocol. The proposed model achieves an MAE of 2.14 when using binary contrast ($\lambda_c = 10$, $\lambda_t = 0$). It performs best among all approaches that do not use external data for pre-training. Several approaches employ model pre-trained on large face datasets like IMDB-WIKI [36], MS Celeb 1M [15], or FaceAugmentation [26]. Results following the SE evaluation protocol on the MORPH II dataset are shown in Table 3. The proposed model achieves an MAE of 2.37 when using ternary contrast ($\lambda_c = 10$, $\lambda_t = 1$). In the SE protocol, images of individuals who appear in the training set are excluded from the test set. We observe that in this case the ternary loss

Table 2: Evaluation on MORPH II. Mean Absolute Error (MAE) comparison following the RS evaluation protocol. RS(I) indicates pre-training on ImageNet, RS(F) on other large face datasets, including IMDB-WIKI [36], MSCeleb(†) [15], and FaceAugmentation(‡) [26]. Parameters: (binary) $\lambda_c = 10$, $\lambda_t = 0$ (ternary) $\lambda_c = 10$, $\lambda_t = 1$.

Method	RS(I)	RS(F)	Year
CRCNN[1]	3.74		2016
OR-CNN[30]	3.27		2016
DEX[36]	3.25	2.68	2016
ODFL[23]	3.12		2017
ARN[3]	3.00		2017
Ranking-CNN[8]	2.96		2017
AP[45]	2.87	2.52	2017
M-LSDML[22]		2.89 [‡]	2018
RCL[32]		2.46	2018
SVRT[18]		2.38 [†]	2018
MV[31]	2.41	2.16	2018
C3AE[43]	2.78	2.75	2019
BridgeNet[21]		2.38	2019
ADVL[41]		1.94	2020
NRLD[11]	2.35	1.81	2020
OCCO[4]	2.29		2021
ADPF[39]	2.54		2022
Ours (binary loss)	2.14		
Ours (ternary loss)	2.20		

Table 3: Evaluation on MORPH II. Mean Absolute Error (MAE) comparison following the SE evaluation protocol.

Methods	SE(I)	SE(F)	Year
DIF[16]	3.00		2018
RCL[32]		2.88	2018
SVRT[18]		2.87 [†]	2018
MV[31]	2.80	2.79	2018
Ours (binary loss)	2.43		
Ours (ternary loss)	2.37		

improves the performance. The coefficients of the individual loss terms are discussed in the ablation study.

Results on FG-NET. As shown in Table 4, our method shows good performance on the FG-NET dataset. We observe that our model shows competitive performance with models that use an external dataset, IMDB-WIKI [36]. By sampling triplets of face images, the proposed method achieves excellent performance on FG-NET without any additional data.

Examples of age estimation results on FG-NET and MORPH II are shown in Fig. 2. The proposed method performs robustly for various age ranges. Poor estimates are typically caused by poor image quality.

Table 4: Evaluation on FG-NET in terms of Mean Absolute Error (MAE) following the LOPO evaluation protocol. LOPO(I) indicates pre-training on ImageNet, LOPO(F) on the large IMDB-WIKI [36] dataset. Parameters: (binary) $\lambda_c = 1$, $\lambda_t = 0$ (ternary) $\lambda_c = 1$, $\lambda_t = 1$

Methods	LOPO(I)	LOPO(F)	Year
DEX[36]	4.63	3.09	2016
CRCNN[1]	4.13		2016
RCL[32]		4.21	2018
MV[31]	4.10	2.68	2018
C3AE[43]	4.09	2.95	2019
BridgeNet[21]		2.56	2019
NRLD[11]	3.23	2.55	2020
ADVL[41]		2.32	2020
ADPF[39]	2.86		2022
Ours (binary loss)	2.42		
Ours (ternary loss)	2.31		

Table 5: Mean variance for fixed identity.

Dataset	Methods	MV	Ours
FG-NET	$\mu(V_f)$	0.18	6.16
FG-NET	$\mu(V_s)$	36.03	43.71
MORPH II	$\mu(V_f)$	0.18	7.22
MORPH II	$\mu(V_s)$	35.87	44.44

Identity invariance. To measure the dependence on face identity, we compare the feature variance when keeping the identity fixed. A low variance indicates a larger dependency on identity and vice versa. We calculate the mean variance of the extracted feature \mathbf{f} and \mathbf{s} by identity and compare it with the Mean Variance (MV) method [31] on FG-NET and MORPH II datasets. As shown in Table 5, the feature extracted by the proposed method show higher variance than the MV method for the same identity.

For the qualitative study, a Grad-CAM [38] comparison between the MV method [31] and the proposed method on the MORPH II dataset is shown in Fig. 2. MV focuses on the entire face area, whereas the proposed model concentrates mainly on forehead regions for people in their 30s and 40s. In case of teens and 20s, it focuses on areas around the nose and mouth. The eye related features like eye shape, eye size and eyebrow shape are reported as distinctive feature for face identification than other facial features [2]. On the other hand, features related to wrinkles around the forehead, nose and mouth region are important features for age estimation [28]. As shown in Fig. 2, the proposed method makes model less dependent on the identity related features and emphasize the features related to the age.

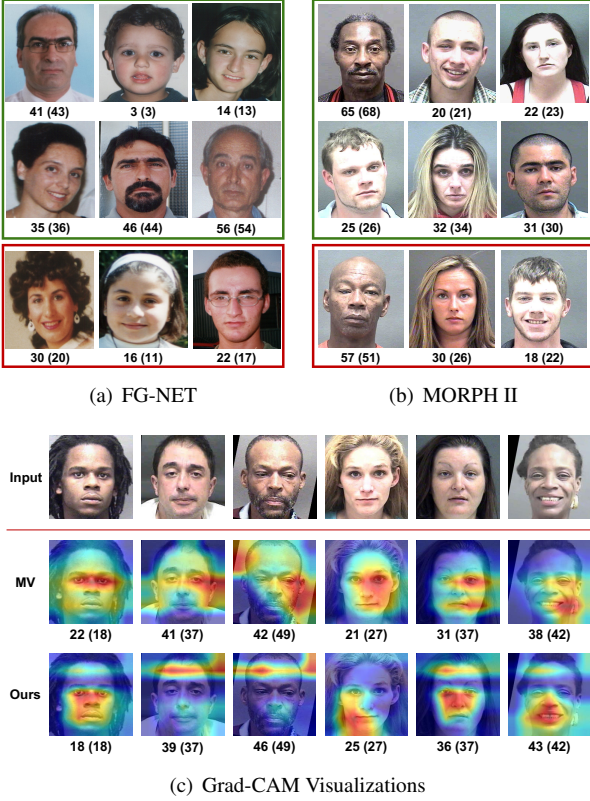


Figure 2: Example age prediction results on (a) *FG-NET* and (b) *MORPH II*. Numbers below each image show predicted and ground truth age. Top two rows show accurate estimation results, the bottom row shows poor estimates. (c) A Grad-CAM analysis on *MORPH II* samples shows that our method better focuses on features extracted from the face region compared to the Mean Variance method [31].

Ablation study on loss functions. To evaluate the combination of different loss functions we study binary and ternary contrast by including or excluding the triplet margin loss. In addition, we compare cosine similarity with a Kullback-Leibler (KL) divergence loss. Cosine similarity is a common to measure of feature similarity while the KLD loss is used for continuous distributions. We adopt the KLD loss on the softmax probabilities for anchor and positive samples:

$$L_{\text{KLD}}(\mathbf{s}_a, \mathbf{s}_p) = \frac{1}{A} \sum_{j=1}^A \mathbf{s}_{p,j} (\log \mathbf{s}_{p,j} - \mathbf{s}_{a,j}) . \quad (9)$$

The MAE results for different losses are shown in Table 6. For both binary and ternary contrast, the KLD loss performs worse than the cosine similarity loss. The triplet margin loss improves the performance on small datasets such as *FG-NET*. The cosine similarity loss improves the accuracy when there is sufficient training data such as in the *MORPH II* dataset. Overall, cosine similarity loss with triplet margin loss shows the best performance for both small and large datasets.

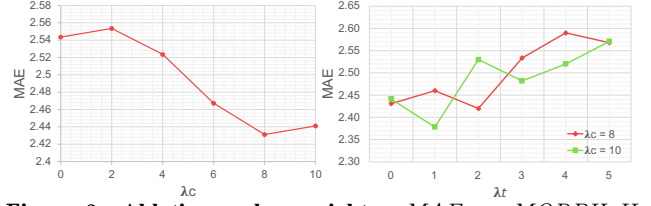


Figure 3: Ablation on loss weights. MAE on *MORPH II* (SE) for varying λ_c and λ_t . (left) varying λ_c with $\lambda_t = 0$, (right) varying λ_t with $\lambda_c = 8$ or 10.

Table 6: Ablation study. MAE for different loss terms. Best and second-best results are shown in bold and underlined.

Loss	FG-NET (LOPO)	MORPH II (RS)	MORPH II (SE)
MV	4.10	2.41	2.80
MV + KLD	2.43	2.34	2.85
MV + Cosine	2.42	2.19	<u>2.54</u>
MV + Triplet	2.30	2.49	2.88
MV + KLD + Triplet	2.33	3.13	2.85
MV + Cosine + Triplet	<u>2.31</u>	<u>2.24</u>	2.50

Loss coefficients. In order to select weights for the individual loss terms, we evaluate different weight combinations. We fix coefficients λ_m , λ_v for the mean and variance terms to 0.2 and 0.05, respectively. We change the coefficient of the cosine similarity loss, λ_c without negative channel and the triplet margin loss (i.e. $\lambda_t = 0$). We change it from 0 to 10 because an absolute scale of the cosine similarity loss is ten times smaller than either mean loss or variance loss. We measure the MAE with the SE protocol on the *MORPH II* dataset, see Fig. 3 (a). We observe that cosine similarity shows good performance for values $\lambda_c = 8$ and $\lambda_c = 10$. We vary the triplet loss coefficient, λ_t from 0 to 5. We measure the MAE on *MORPH II* (SE protocol). As shown in Fig. 3 (b), the cosine similarity loss is a poor choice ($\lambda_t = 0$). Compared with the triplet margin loss, the cosine similarity loss has a relatively larger influence on the performance of the model. The minimum MAE is obtained for $\lambda_t = 1$ and $\lambda_c = 10$.

4 Conclusion

We introduced a method for age estimation from face images via contrastive learning from triplets of face images. Our proposed approach focuses on sparser features that are more relevant to age by penalizing non-relevant features that are associated with identity. To encourage the similarity of positive samples, we leverage cosine similarity, and we employ a ternary loss to increase the distance to negative samples. Experiments on the *MORPH II* and *FG-NET* datasets demonstrated the effectiveness of our proposed method, which achieved state-of-the-art results.

References

- [1] Fatma S Abousaleh, Tekoing Lim, Wen-Huang Cheng, Neng-Hao Yu, M Anwar Hossain, and Mohammed F Alhamid. A novel comparative deep learning framework for facial age estimation. *EURASIP Journal on Image and Video Processing*, 2016(1):1–13, 2016.
- [2] Naphtali Abudarham and Galit Yovel. Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, 16(3), 2016.
- [3] Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Anchored regression networks applied to age estimation and super resolution. In *ICCV*, pages 1652–1661, 2017.
- [4] Weiwei Cai and Hao Liu. Occlusion contrasts for self-supervised facial age estimation. In *Multimedia Understanding with Less Labeling on Multimedia Understanding with Less Labeling*, pages 1–7. 2021.
- [5] Kuang-Yu Chang and Chu-Song Chen. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, 24(3):785–798, 2015.
- [6] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. A ranking approach for human ages estimation based on face images. In *2010 20th International Conference on Pattern Recognition*, pages 3396–3399, 2010.
- [7] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR 2011*, pages 585–592, 2011.
- [8] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *CVPR*, pages 742–751, 2017.
- [9] Mohammad Mahdi Dehshibi and Azam Bastanfard. A new algorithm for age recognition from facial images. *Signal Processing*, 90(8):2431–2444, 2010.
- [10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [11] Zongyong Deng, Mo Zhao, Hao Liu, Zhenhua Yu, and Feng Feng. Learning neighborhood-reasoning label distribution (nrld) for facial age estimation. In *IEEE Int. Conf. Multimedia and Expo (ICME)*, pages 1–6, 2020.
- [12] F. Dornaika, SE. Bekhouche, and I. Arganda-Carreras. Robust regression with deep cnns for facial age estimation: An empirical study. *Expert Systems with Applications*, 141:112942, 2020.
- [13] Yun Fu and Thomas S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008.
- [14] Guodong Guo, Yun Fu, Charles R. Dyer, and Thomas S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE TIP*, 17(7):1178–1188, 2008.
- [15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.
- [16] Hu Han, Anil K. Jain, Fang Wang, S. Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE TPAMI*, 40:2597–2609, 2018.
- [17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.
- [18] Woobin Im, Sungeun Hong, Sung-Eui Yoon, and Hyun S Yang. Scale-varying triplet ranking with classification loss for facial age estimation. In *Asian Conference on Computer Vision*, pages 247–259. Springer, 2018.
- [19] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022.
- [20] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.
- [21] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgnet: A continuity-aware probabilistic network for age estimation. In *CVPR*, pages 1145–1154, 2019.
- [22] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, 13(2):292–305, 2017.
- [23] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Ordinal deep feature learning for facial age estimation. In *IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 157–164, 2017.
- [24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [25] Yangjing Long. Human age estimation by metric learning for regression problems. In *Int. Conf. Computer Graphics, Imaging and Visualization*, pages 343–348, 2009.
- [26] Iacopo Masi, Anh Tu an Tr  n, Tal Hassner, Jatuporn Toy Leksut, and G  rard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision (ECCV)*, October 2016.
- [27] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.
- [28] Choon-Ching Ng, Moi Hoon Yap, Nicholas Costen, and Baihua Li. Will wrinkle estimate the face age? In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2418–2423, 2015.
- [29] Hieu V Nguyen and Li Bai. Cosine similarity metric

- learning for face verification. In *ACCV*, pages 709–720. Springer, 2010.
- [30] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928, 2016.
 - [31] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, pages 5285–5294, 2018.
 - [32] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Revised contrastive loss for robust age estimation from face. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3586–3591. IEEE, 2018.
 - [33] Gabriel Panis, Andreas Lanitis, Nicolas Tsapatsoulis, and Timothy Cootes. An overview of research on facial aging using the fg-net aging database. *IET Biometrics*, 5, May 2015.
 - [34] KBRK Ramesha, KB Raja, KR Venugopal, and LM Patnaik. Feature extraction based face recognition, gender and age classification. *International Journal on Computer Science and Engineering*, 2:14–23, 2010.
 - [35] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006.
 - [36] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126:144–157, 2016.
 - [37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.
 - [38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.
 - [39] Haoyi Wang, Victor Sanchez, and Chang-Tsun Li. Improving face-based age estimation with attention-based dynamic patch fusion. *IEEE Transactions on Image Processing*, 2022.
 - [40] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
 - [41] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. Berlin, Heidelberg, 2020. Springer-Verlag.
 - [42] Zhiguang Yang and Haizhou Ai. Demographic classification with local binary patterns. In Seong-Wan Lee and Stan Z. Li, editors, *Advances in Biometrics*, pages 464–473, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.
 - [43] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3ae: Exploring the limits of compact model for age estimation. In *CVPR*, pages 12579–12588, 2019.
 - [44] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.
 - [45] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. Quantifying facial age by posterior of age comparisons. In *BMVC*, 2017.