ADAPTIVE MEAN-RESIDUE LOSS FOR ROBUST FACIAL AGE ESTIMATION

Ziyuan Zhao[†‡], *Peisheng Qian*[†], *Yubo Hou and Zeng Zeng*[†‡]

Institute of Infocomm Research, A*STAR, Singapore
{zhaoz, qianp, houy, zengz}@i2r.a-star.edu.sg
[‡]School of Computer Science and Engineering, Nanyang Technological University, Singapore

ABSTRACT

Automated facial age estimation has diverse real-world applications in multimedia analysis, *e.g.*, video surveillance, and human-computer interaction. However, due to the randomness and ambiguity of the aging process, age assessment is challenging. Most research work over the topic regards the task as one of age regression, classification, and ranking problems, and cannot well leverage age distribution in representing labels with age ambiguity. In this work, we propose a simple yet effective loss function for robust facial age estimation via distribution learning, *i.e.*, adaptive mean-residue loss, in which, the mean loss penalizes the difference between the estimated age distribution's mean and the ground-truth age, whereas the residue loss penalizes the entropy of age probability out of dynamic top-K in the distribution. Experimental results in the datasets FG-NET and CLAP2016 have validated the effectiveness of the proposed loss.

Index Terms— Age estimation, label distribution learning, deep learning, neural networks.

1. INTRODUCTION

Automated facial age estimation has been widely applied to different multimedia application scenarios but remains a very challenging task. Human face aging is a complicated and random process affected by various internal factors, *e.g.*, genes, makeup, living environment. As a result, there are noticeable variances in facial appearance among different subjects with similar ages. Furthermore, the aging process of each subject is lasting, making it hard to perceive the variances in its facial appearance among neighbouring ages.

In general, the research work for facial age estimation in the literature can be classified into three categories, *i.e.*, regression-based [1, 2], classification-based [3, 4], and ranking-based [5, 6]. But these methods fail to address the age distribution effectively, as there are no clear boundaries between adjacent ages, it is relatively easy for humans to guess the age with some confidence. For example, one person could guess a girl is around 25 years old, or she may be in her mid-20s. In real life, the guessing process follows some certain probability distribution, which means that we could

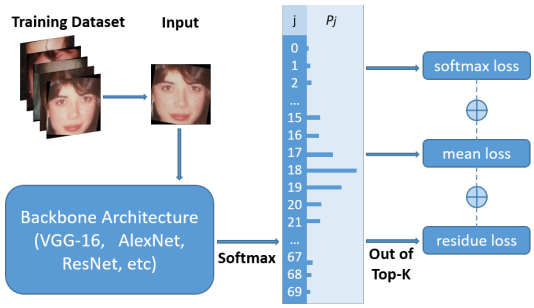

Fig. 1: Overview of the proposed approach

convert the mission of age estimation into a process of age distribution learning. In [7], the authors assumed that the age distribution follows a Gaussian distribution with a particular mean age, m_i, and a standard deviation, σ_i, for i-th image, and proposed a novel mean-variance loss. The mean-variance loss attempts to penalize the variance between m_i and the ground-truth age, y_i, by mean loss, and concentrate more on the classes around m_i by variance loss. However, it is possible that the 2 loss functions suppress each other in part of the age distribution, achieving undesired solutions, and even worse, if y_i happens to be out of the range $[m_i - \sigma_i,\ m_i + \sigma_i]$, variance loss penalizes the softmax loss that attempts to increase the probability of y_i. Furthermore, the age distributions of different subjects vary a lot, and it is not true that smaller σ_i leads to a more accurate m_i.

On the other hand, the top-K accuracy of deep learning models has achieved a very high level, *e.g.*, the best *top-5* accuracy on the ImageNet dataset is 98% [8]. In age estimation, we can assume that y_i is included in top-K classes and the ages out of top-K have a very limited correlation with y_i, and then the sum of out-of-top-K probabilities can be treated as *residue*. However, how to determine K remains a problem. These motivate us to propose a hypothesis: *If it is hard to extract deeper facial features, why not suppress uncorrelated features and dynamically penalize the residue to strengthen the correlation among the top-K classes indirectly?*

Following the idea, we designed an adaptive entropy-based residue loss, which can penalize the age probabilities out of dynamic top-K. By combining mean loss with residue

loss, we proposed a simple, yet very efficient loss, adaptive mean-residue loss, for facial age estimation. The proposed mean-residue loss is simple to incorporate into other network architectures, such as ResNet. Experimental results are superior to the existing state-of-the-art benchmarks, *e.g.*, mean-variance loss.

2. RELATED WORK

Early age estimation work from facial portraits was carried out in [9], where ages were simply split into the following categories, *i.e.*, babies, young and senior individuals. Since then, more and more research interests have been attracted to age estimation from facial images [6, 1]. Traditional methods for facial age estimation consist of two separate stages: feature extraction and one of age regression, classification, and ranking-based methods. In [10], an active appearance model (AAM) was proposed to use shape landmarks and textural features for age estimation. In [11], age estimation is achieved by multi-direction and multi-scale Gabor filters with the feature pooling function were used for identification of biologically inspired features (BIF). Robustness of BIF in age estimation has been verified, but it still relied on feature representations crafted by hand, which is not an optimal solution to facial age estimation.

Deep learning based approaches have succeeded in many computer vision tasks, *e.g.*, object detection [12], image classification [13]. and medical image analysis [14, 15]. They also have remarkable advance facial age estimation recently. In [16], deep convolutional neural networks (DCNNs) were proposed to extract features from different regions on facial images and a square loss was utilized for age prediction. In [17], a multi-task deep learning model was proposed for joint estimation based on numerous attributes, including shared feature extraction and attribute group feature extraction. To encode both the ordinal information between adjacent ages and their correlation, soft-ranking label encoding was proposed in [18], which encourages deep learning models to learn more robust facial features for age estimation.

On the other hand, label distribution learning (LDL) was proposed to distinguish label ambiguity, which is challenging. [19, 20]. In LDL, a label distribution can be assigned to an sample. Moreover, the correlation among values in the label space can be leveraged, from which a more robust estimation can be obtained. Gang *et al.* proposed several LDL based methods for age estimation and presented the robustness of them, *e.g.*, maximum-entropy modeling [21]. It has been argued that a single facial image facilitates not only the estimation of a singular age, but is also informative for adjacent ages [22]. In [7], novel loss functions for model training were proposed and the mean-variance loss was proposed, in which the mean loss aims at minimizing the distance between the predicted age and ground-truth, while the variance loss attempts to lower the variance in the predicted age distribution. Consequently, the age distribution could be sharpened. Meanwhile, a sharper age distribution does not necessarily lead to better age estimation. Under particular circumstances, mean loss and variance loss contradict each other and prevent the accurate prediction to be achieved. More details of mean-variance loss will be discussed in Section 4.

3. PRELIMINARY

In an age estimation dataset, $y_i \in \{1, 2, \ldots, L\}$ represents the corresponding age of i-th sample, x_i stands for the facial feature, and $f(x_i) \in \mathbb{R}^{N\times M}$ represents the output from the layer, followed by a last fully connected (FC) layer. Let $z \in \mathbb{R}^{N\times L}$ be the output vector from the last FC layer, and $p \in \mathbb{R}^{N\times L}$ is the softmax probability defined in Equ.1:

$$z = f(x_i)\cdot\theta^T; \quad p_{i,j} = \frac{e^{z_{i,j}}}{\sum_{l=1}^{L} e^{z_{i,l}}}, \tag{1}$$

in which x_i denotes the feature vector, $\theta \in \mathbb{R}^{L\times M}$ contains trainable parameters in the FC layer, $j \in \{1, 2, \ldots, L\}$ stands for an age, $z_{i,j}$ is an element of z in the i-th sample with age j. $p_{i,j}$ represents the probability that the age of sample i is j. Hence, the estimated or mean age m_i of the sample i is calculated in Equ. 2:

$$E(age_i) = m_i = \sum_{j=1}^{L} j \cdot p_{i,j}. \tag{2}$$

4. METHODOLOGY

As illustrated in Fig. 1, the proposed adaptive mean-residue loss can be embedded into a deep convolutional neural network (DCNN) for end-to-end learning. As illustrated in Fig. 2, the proposed adaptive mean-residue loss penalizes (a) the difference between the mean of the estimated age distribution and the ground-truth age, and (b) residue errors in the two long tails of the age distribution.

4.1. Mean Loss

The mean loss suppresses the variance between an estimated age distribution's mean and the true age. According to Equ. 2, we define the mean loss as

$$L_m = \frac{1}{2N}\sum_{i=1}^{N}(m_i - y_i)^2 = \frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=1}^{L} j\cdot p_{i,j} - y_i)^2, \tag{3}$$

in which N, m_i, and y_i are the training batch size, the estimated age, and and ground-truth age, respectively. Unlike the widely used softmax loss, the mean loss is presented in many regression problems. L_2 distance can be used to evaluate the variances between the mean of an estimated age distribution and the ground-truth age. Therefore, the proposed mean loss complements the softmax loss during training.

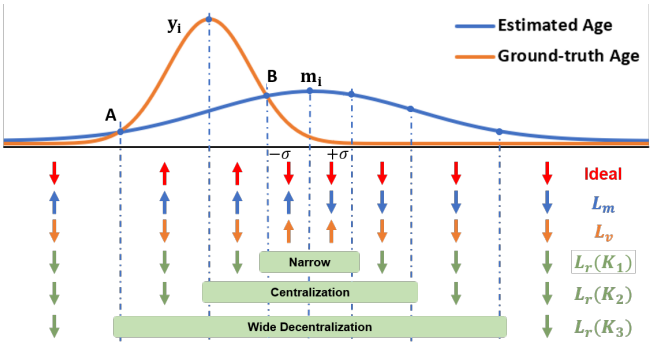

Fig. 2: Gradient analysis on different components of the loss functions. (L_m: the mean loss, L_v: the variance loss, L_r: the residue loss, where K_1, K_2 and K_3 denote three different situations.) $\uparrow$ and $\downarrow$ indicate the direction in which the loss function optimizes the distribution, while the ideal direction is indicated (red). For simplicity, $-\sigma$, B, and K_1 overlap; y_i and K_2 overlap; A and K_3 overlap.

4.2. Residue Loss

The residue loss penalizes the residue errors in the tails that exist in an estimated age distribution after the top-K pooling operation. The residue entropy is presented as follows,

$$L_r = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1, j\notin top-K}^{L} p_{i,j}\cdot \log p_{i,j}. \tag{4}$$

The mean-variance loss [7] tries to suppress the probabilities within the range $m_i \in (j \in [1, m_i-\sigma]) \cup [m_i+\sigma, L])$, where σ is the standard deviation of the age distribution in the i-th sample. However, inevitably, y_i may probably fall inside this range. In this regard, p_{i,y_i} is incorrectly suppressed, which could induce errors, while our residue loss L_r guarantees that y_i falls in the top-K classes and avoids such errors that occur with the mean-variance loss.

To further optimize the residue loss, we propose a dynamic top-K pooling method. More specifically, let K_i denote the number of top-K. Let R_{y_i} stands for the ranking of y_i, for i-th image in a training batch. Subsequently, we can set K_i to be adaptive as:

$$K_i = max\{2, R_{y_i}\}. \tag{5}$$

With the dynamic top-K pooling, y_i is always included during the optimization process, and will not be incorrectly penalized by the residue loss.

4.3. Adaptive Mean-Residue Loss

Combining with the softmax loss L_s, the adaptive mean-residue loss is showed in Equ. 6:

$$\begin{aligned} L &= L_s + \lambda_1 L_m + \lambda_2 L_r \\ &= \frac{1}{N}\sum_{i=1}^{N}[-\log p_{i,y_i} + \frac{\lambda_1}{2}(m_i - y_i)^2 + \lambda_2 r_i], \end{aligned} \tag{6}$$

where λ_1 and λ_2 are hyperparameters that attempt to balance the influences of mean and residue sub-losses in the combined loss function. We use SGD [13] to optimize parameters in the network. At inference, the estimated age of the i-th test image can be calculated based on Equ. 2.

4.4. Gradient Analysis

4.4.1. *Comparison with Mean-Variance Loss*

The mean-variance loss and the proposed mean-residue loss share the same mean loss function. The key difference exists in the variance loss and the residue loss, and their joint effect with the mean loss, as plotted in Fig. 2. In the mean-variance loss, the variance loss attempts to enhance the probabilities of classes within the range of $[m_i-\sigma, m_i+\sigma]$ while suppressing the probabilities of classes out of the range, **no matter** y_i **is in the range or not**. When y_i happens to be out of the range, variance loss will decrease the probability of y_i, leading to a worse effect on the model performance. In the mean-residue loss, the residue entropy loss function attempts to suppress the probabilities of the classes out of top-K and help the network to focus more on the top-K classes indirectly. Adaptive top-K guarantees that y_i is within top-K and residue loss function can help the network to increase the probability of classes within top-K as a whole. Hence, the performance of softmax loss can be enhanced. Effects of different K values for top-K are analyzed in the following.

4.4.2. *Situations of Different K*

The joint effect of the mean and residue loss is dependent on the choice of top-K in the residue loss. A proper selection of K is the key to ensuring the correct optimization. The analysis of top-K is illustrated in Fig. 2, and different situations of top-K are summarized as follows:

Over Centralization: When top-K is too **narrow**, *e.g.*, K_1 in Fig. 2, the ground-truth y_i is **excluded** from the **over-centralized** top-K classes. The probability at y_i is penalized by the residue loss in the wrong direction, and the residue loss conflicts with the mean loss at y_i, which is undesired.

Centralization: When top-K is **appropriate**, *e.g.*, K_2 in Fig. 2, the ground-truth y_i is **included** in the **centralized** top-K classes. The residue loss at y_i is 0, and the joint effect of both mean and residue loss are the same as the ideal direction, which is desired.

Decentralization: When top-K is too **wide**, *e.g.*, K_3 in Fig. 2, the ground-truth y_i is **included** in the **decentralized**

top-K classes. The joint effect of the loss functions is therefore following the ideal optimization direction. However, too many classes are covered by the top-K, in which the residue loss is not calculated. It lessens the effect of the residue loss and makes it more difficult to optimize the model.

Compared with a fixed K, the proposed dynamic top-K pooling can obtain a proper range for centralizing the age distribution. As the training proceeds, the distributions of the prediction converge towards the ground-truth, during which the optimal K also decreases. Empirically, a fixed K is either too small at the start of training, excluding y_i from top-K pooling, or too large towards the end of the training, slowing down the optimization process. In contrast, the adaptive K, as shown in Equ. 5, always includes y_i without over centralization or decentralization. Therefore it does not suffer from the drawbacks of a fixed K. From our theoretical analysis, we can anticipate that the adaptive mean-residue loss shall outperform the mean-variance loss in accuracy and convergence.

5. EXPERIMENTS

5.1. Datasets and Evaluation Metrics

Extensive experiments have been carried out using the proposed loss on two popular facial image datasets, *i.e.*, FG-NET [23] and CLAP 2016 [24]. The basic statistics of age distribution refers to Table 1. **FG-NET** [23] includes 1002 facial images collected from 82 subjects. The ages of subjects lie between 0 to 69 years old. We adopt the leave-one-person-out (LOPO) protocol in the experiments [7]. **CLAP2016** [24] is a competition dataset released in 2016 at the ChaLearn Looking at people challenges. There are 4113 training subjects, 1500 validation subjects and 1979 test subjects. In CLAP 2016, An apparent mean age and standard deviation is labeled to each image. For evaluation, we use the mean absolute error (MAE) between the ground-truth age μ_i and the prediction y_i in FG-NET. The ε-error is adopted from [24] for CLAP2016, which is formulated as:

$$\varepsilon = 1 - \frac{1}{N} \cdot \sum_{i=1}^{N} \exp - \frac{(y_i - \mu_i)^2}{2 \cdot \sigma_i^2}, \tag{7}$$

where μ_i stands for the ground-truth mean age of sample i, and σ_i $(1 \leq i \leq N)$ is its standard deviation.

Table 1: Age distribution of the face images.

Age	$0 \sim 19$	$20 \sim 39$	$40 \sim 59$	$60 \sim 69$	≥ 80
FG-NET	710	223	61	8	0
CLAP 2016	1394	4362	1423	366	46

Table 2: Comparisons of different losses.

Method	FG-NET		CLAP2016	
	VGG-16	ResNet-50	VGG-16	ResNet-50
Softmax Loss	7.19	6.99	0.4926	0.4756
Mean Loss + Softmax Loss	4.25	3.95	0.4687	0.4532
Variance Loss + Softmax Loss	9.78	7.63	0.5516	0.5697
Mean-Variance Loss	4.10	3.95	0.4552	0.4018
Residue Loss + Softmax Loss	6.39	6.55	0.4921	0.4699
Adaptive Mean-Residue Loss	**3.79**	**3.61**	**0.4511**	**0.3882**

5.2. Experiment Settings

To reduce the influence of noises, *e.g.*, bodies, environments, all face images from different datasets were cropped with the cascaded classifier in OpenCV [25] and resized into $256 \times 256 \times 3$. We also employed data augmentation with rotation, flipping, color jittering, and affine transformations to reduce the overfitting. We adopted VGG-16 [26] and ResNet-50 [13] as our deep learning architectures for age estimation. The models are initialized with weights pre-trained using ImageNet [27]. The models are implemented using PyTorch. The initial learning rate and batch size are set to 0.001 and 64 respectively. The model is trained for 100 epochs. Furthermore, for every 15 epoch, the learning rate is reduced by a multiplication factor of 0.1.

5.3. Comparisons with Different Losses

We compare the proposed mean-residue loss with the mean-variance loss proposed in [7]. Both two losses have components softmax loss (i.e., L_s) and mean loss (i.e., L_m). Besides, we testify the effect of each component from both losses in an incremental manner. As shown in Table 2, we notice that the mean component always plays a core role in the prediction under both VGG-16 and ResNet-50. In comparison between variance loss and residue loss, the residue loss beats the variance loss in either case of the combination with the softmax loss or the combination with both the mean and softmax loss, which demonstrates the effectiveness of the proposed residue loss. Finally, our adaptive mean-residue loss outperforms all the other combinations, including mean-variance loss using VGG-16 and ResNet-50.

5.4. Influences of the Parameter λ_1 and λ_2

Since hyperparameters λ_1 and λ_2 in Equ. 6 control the strengths of three components (softmax, mean, and residue) in the proposed loss during network training, we evaluate their influences of λ_1 and λ_2 on CLAP2016. In the rest of this section, we aim to pick out the best portfolio for a pair of hyperparameters λ_1 and λ_2 in the proposed loss function. Following [7], we set λ_1 to 0.2 empirically and change λ_2 from 0 to 0.2 at interval of ever 0.025. The $\epsilon-$error of the performance with different portfolios with different architectures, *i.e.,* VGG-16 and ResNet-50 are shown in Fig. 3. We

can see that VGG-16 is not sensitive to the changes of λ_2, while the best portfolio on CLAP 2016 is $\lambda_1 = 0.2$ and $\lambda_2 = 0.75$ for ResNet-50, which yields the lowest $\epsilon-$error.

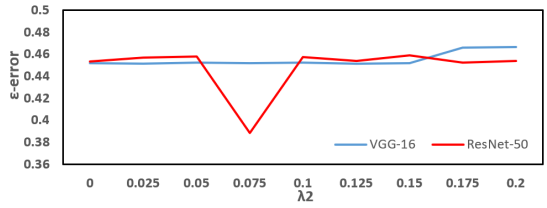

Fig. 3: The $\epsilon-$error w.r.t. $\lambda_1 = 0.2$ and λ_2 from 0 to 0.2.

5.5. Influences of the Parameter K

In Section 4, the influence of the choice of K on the model optimization is explained. Fig. 4 illustrates such impact of K values on the model performance. For a fixed K, it can be observed that the test MAE is lowest when $K = 5$. Moreover, it is noted that the model with an adaptive K value consistently outperforms fixed K values. Fig. 4 describes the adaptive K values during training, which gradually converges to the best fixed K value of 5, proving the capability of the algorithm to find the optimal K during training.

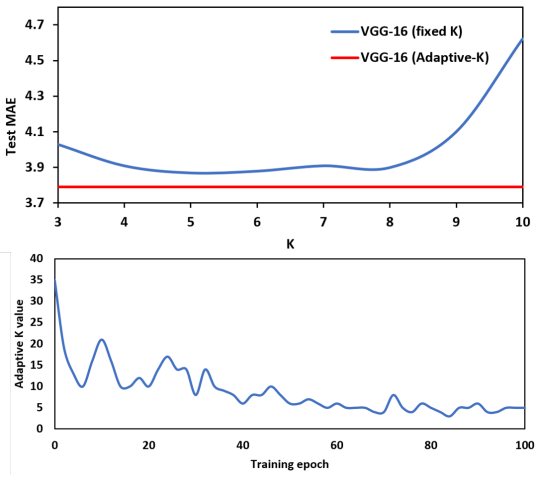

Fig. 4: Top: The MAE on FG-NET achieved by VGG-16 using different K (blue line) and Adaptive K (red line). Bottom: The adaptive K value during training.

5.6. Comparisons with the State-of-the-art

Experiments have been carried out for comparison between the proposed method and a number of state-of-the-art benchmarks on FG-NET and CLAP2016 respectively. As indicated in Table 3, the proposed mean-residue loss achieves the lowest MAE error among these approaches in FG-NET. Experimental results show that methods with distribution learning, such as mean-variance loss [7] can outperform ranking, regression, or classification based methods [28, 5, 29]. It is noted that, in DHAA [30], a hybrid structure with multiple branches was utilized to achieve the best performance among the rest approaches. We can explore the combination of DHAA with mean-residue loss in our future work. In Table 4, we directly quote the results of DeepAge and MIPAL_SNU from [24] as a complementary comparison. Our proposed loss outperforms mean-variance loss, which proves that residue loss with *adaptive-K* pooling is helpful to concentrate more on top-K ages indirectly. In Fig. 5, we present some examples (3 good cases and 3 poor cases) predicted by the proposed method on CLAP 2016. The proposed approach performs well for different age groups. But when the images have poor qualities, *e.g.*, bad illumination, blurring, the age estimation accuracy decreases dramatically. In addition, good makeup and extreme values would influence the results.

Table 3: Comparisons on FG-NET dataset.

Method	MAE	Protocol
RED-SVM [28]	5.24	LOPO
OHRank [5]	4.48	LOPO
DEX [29]	4.63	LOPO
Mean-Variance Loss [7]	3.95	LOPO
DRFs [31]	3.85	LOPO
DHAA [30]	3.72	LOPO
Adaptive Mean-Residue Loss	**3.61**	LOPO

Table 4: Comparisons on CLAP2016 dataset.

Method	ε-error	Single Model?
DeepAge [24]	0.4573	YES
MIPAL_SNU [24]	0.4565	NO
Mean-Variance Loss [7]	0.4018	YES
Adaptive Mean-Residue Loss	**0.3882**	YES

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a simple, yet very efficient mean-residue loss for robust facial age estimation. We verify the superiority of our proposed method over state-of-the-art benchmarks through theoretical analysis and experiments. In the future, we would extend our method to other domains for continuous value estimation, such as survival year estimation in healthcare, sales prediction in e-business, etc.

7. ACKNOWLEDGEMENT

We are grateful for the help and support of Xiaohao Lin, Xi Fu and Ce Ju. The work is supported by Institute for Info-

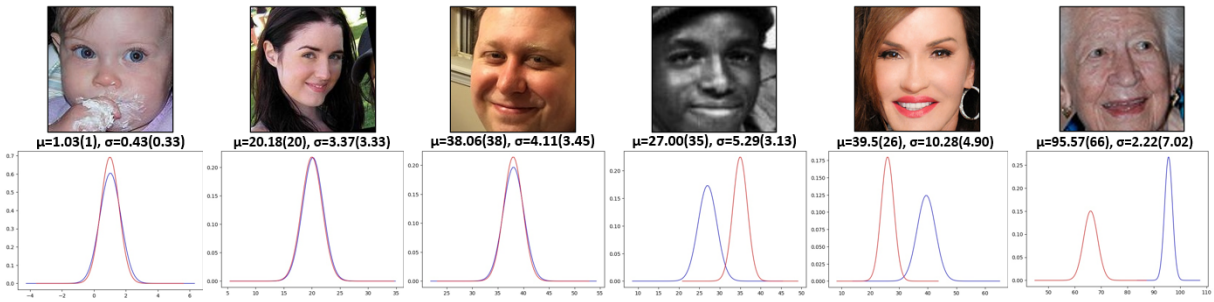

Fig. 5: Examples with age distributions (μ = ground-truth age (estimated age), σ = ground-truth std (estimated std)) estimated by the proposed loss. The red and blue curves are the estimated and ground-truth distributions, respectively.

comm Research (I2R) and Artificial Intelligence, Analytics and Informatics (AI3), Agency for Science, Technology and Research (A*STAR), Singapore.

8. REFERENCES

[1] Yun Fu and Thomas Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Trans. Multimedia*, 2008. 1, 2

[2] Xi Yang, Jianyi Liu, Yao Ma, and Jianru Xue, "Facial age estimation from web photos using multiple-instance learning," in *IEEE ICME*, 2014. 1

[3] Yixin Zhu, Jun-Yong Zhu, and Wei-Shi Zheng, "Part-based convolutional network for imbalanced age estimation," in *IEEE ICME*, 2019. 1

[4] Chao Zhang, Ce Zhu, Jimin Xiao, Xun Xu, and Yipeng Liu, "Image ordinal classification and understanding: Grid dropout with masking label," in *IEEE ICME*, 2018. 1

[5] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *IEEE CVPR*, 2011. 1, 5

[6] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *IEEE CVPR*, 2017. 1, 2

[7] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen, "Mean-variance loss for deep age estimation from a face," in *IEEE CVPR*, 2018. 1, 2, 3, 4, 5

[8] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Herve Jegou, "Fixing the train-test resolution discrepancy," *Advances in Neural Information Processing Systems*, 2019. 1

[9] Young Kwon and Niels Lobo, "Age classification from facial images," *Computer Vision and Image Understanding*, 1994. 2

[10] Gabriel Panis, Andreas Lanitis, Nicolas Tsapatsoulis, and Timothy Cootes, "Overview of research on facial aging using the fg-net aging database," *IET Biometrics*, 2015. 2

[11] G. Guo, Guowang Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *IEEE CVPR*, 2009. 2

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2015. 2, 3, 4

[14] Ziyuan Zhao, Kerui Zhang, Xuejie Hao, Jing Tian, Matthew Chin Heng Chua, Li Chen, and Xin Xu, "Bira-net: Bilinear attention net for diabetic retinopathy grading," in *IEEE ICIP*, 2019. 2

[15] Ziyuan Zhao, Kaixin Xu, Shumeng Li, Zeng Zeng, and Cuntai Guan, "Mt-uda: Towards unsupervised cross-modality medical image segmentation with limited source labels," in *MICCAI*. Springer, 2021. 2

[16] Dong Yi, Zhen Lei, and Stan Z. Li, "Age estimation by multi-scale convolutional network," in *ACCV*, 2015. 2

[17] F. Wang, H. Han, S. Shan, and X. Chen, "Deep multi-task learning for joint prediction of heterogeneous face attributes," in *IEEE International Conference on Automatic Face Gesture Recognition*, 2017. 2

[18] Xusheng Zeng, Junyang Huang, and Changxing Ding, "Soft-ranking label encoding for robust facial age estimation," *IEEE Access*, 2020. 2

[19] X. Geng, "Label distribution learning," *IEEE Trans. on Knowledge and Data Engineering*, 2016. 2

[20] Zeng Zeng, Wei Zhao, Peisheng Qian, Yingjie Zhou, Ziyuan Zhao, Cen Chen, and Cuntai Guan, "Robust traffic prediction from spatial-temporal data based on conditional distribution learning," *IEEE Trans. on Cybernetics*, 2021. 2

[21] Zhouzhou He, Xi Li, Zhongfei Zhang, Fei Wu, Xin Geng, Yaqing Zhang, Ming-Hsuan Yang, and Yueting Zhuang, "Data-dependent label distribution learning for age estimation," *IEEE Tran. on Image Processing*, 2017. 2

[22] Xin Geng, Chao Yin, and Zhi-Hua Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013. 2

[23] Gabriel Panis and Andreas Lanitis, "An overview of research activities in facial age estimation using the fg-net aging database," in *ECCV*. Springer, 2014. 4

[24] Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al., "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *IEEE CVPRW*, 2016. 4, 5

[25] Paul Viola and Michael J Jones, "Robust real-time face detection," *International journal of computer vision*, 2004. 4

[26] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 4

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, 2015. 4

[28] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung, "A ranking approach for human ages estimation based on face images," in *IEEE CVPR*, 2010. 5

[29] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *IEEE CVPR*, 2016. 5

[30] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li, "Deeply-learned hybrid representations for facial age estimation," in *IJCAI*, 2019. 5

[31] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep regression forests for age estimation," in *IEEE CVPR*, 2018, pp. 2304–2313. 5