– Supplemental Document –

Moving Window Regression: A Novel Approach to Ordinal Regression

Nyeong-Ho Shin
Korea University
nhshin@mcl.korea.ac.kr

Seon-Ho Lee
Korea University
seonholee@mcl.korea.ac.kr

Chang-Su Kim
Korea University
changsukim@korea.ac.kr

1. Facial Age Estimation Datasets

To assess the performance of the proposed algorithm in facial age estimation, we use seven existing datasets: MORPH II [32], CLAP2015 [12], FG-NET [21], CACD [7], UTK [40], Adience [23] and IMDB-WIKI [33]. MORPH II has the Institutional Review Board approval. The other datasets contain facial images obtained by web crawling, and they have been made available for academic research purpose only. Among the images in the datasets, any will be removed if there are delete requests from the original owners. Except for the CACD and IMDB-WIKI datasets, which contain celebrity name labels, there are no name labels. We use the seven datasets only for evaluating the ordinal regression performance of the proposed algorithm. Details about the datasets and experimental settings are as follows.

MORPH II [32]: This is the most widely used dataset for age estimation, containing about 55,000 facial images of 13,617 subjects in the age range [16, 77]. It provides gender and race labels as well. Using these labels, various evaluation protocols have been proposed. We employ the four evaluation settings A, B, C, and D [22, 26], as described in Section 2.1 in this supplemental document.

CLAP2015 [12]: It is for apparent age estimation. The apparent age of each image was rated by at least 10 annotators, and the mean rating was set to be the ground-truth. CLAP2015 also provides the standard deviation of ratings for each image. It contains 4,691 facial images in total, which are split into 2,476 for training, 1,136 for validation, and 1,079 for testing. The age range is [3, 85].

FG-NET [21]: It provides 1,002 color or grayscale images of 82 people in the age range from 0 to 69. As in [25, 38], we use the leave-one-person-out (LOPO) protocol.

CACD [7]: It contains about 160K images of 2,000 celebrities, which are divided into three subsets by celebrities: 1,800 for training, 80 for validation, and 120 for testing. As in [33–35], we train the MWR algorithm using the train set and the validation set, respectively. The age range is [14, 62].

UTK [40]: It provides about 20,000 facial images in the age range [0, 116]. For fair comparison, we employ the same evaluation protocol as in [2, 16] — 13,147 for training, 3,287 for testing.

Adience [23]: It is for age group estimation. It contains 26,580 facial images of 2,284 subjects, which are grouped into 8 classes: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60-year-olds. As in [11, 24, 27, 28], we adopt the 5-fold subject-exclusive (SE) cross-validation evaluation setting.

IMDB-WIKI [33]: It contains about 500,000 celebrity images, crawled from IMDB and Wikipeida. It has been used to pre-train recent age estimators [22, 25, 26, 30, 33, 36, 38]. We also pre-train the proposed ρ-regressors using 175,000 clean images from IMDB-WIKI, unless specified otherwise.

2. More Experimental Results

2.1. Performance Comparison on MORPH II

Four evaluation settings [22, 26] are adopted for the performance comparison on MORPH II [32].

- Setting A: 5,492 images of Caucasians are sampled and then randomly split into train and test sets with ratio 8:2.
- Setting B: About 21K images of Caucasians and Africans are randomly chosen so that the ratio between Caucasians and Africans is 1:1 and that between females and males is 1:3. Then, it is divided into three subsets (S1, S2, S3). The training and testing are repeated twice — 1) training on S1, testing on S2+S3, and 2) training on S2, testing on S1+S3.
- Setting C: The whole dataset is randomly divided into five folds, satisfying the constraint that images of the same person should belong to only one fold. Then, the 5-fold cross-validation is performed.
- Setting D: The whole dataset is randomly divided into five folds without any restriction. Then, the 5-fold cross-validation is performed.

Table S-1 provides more comparison results with conventional algorithms on MORPH II.

Table S-1. Extended table of the performance comparison on the MORPH II dataset.

	Setting A		Setting B		Setting C		Setting D	
	MAE	CS(%)	MAE	CS(%)	MAE	CS(%)	MAE	CS(%)
RED-SVM [4]	-	-	-	-	-	-	6.49	49.0
OHRank [5]	-	-	-	-	-	-	6.07	56.3
KPLS [14]	-	-	4.18	-	-	-	-	-
CPLF [39]	-	-	3.63	-	-	-	-	-
Huerta *et al.* [19]	-	-	-	-	3.88	-	-	-
OR-CNN [29]	-	-	-	-	-	-	3.27	73.0
Tan *et el.* [41]	-	-	3.03	-	-	-	-	-
Ranking-CNN [8]	-	-	-	-	-	-	2.96	85.0
DEX [33]	2.68	-	-	-	-	-	-	-
DMTL [18]	-	-	-	-	3.00	85.3	-	-
CMT [3]	-	-	-	-	2.91	-	-	-
DRFs [34]	2.91	82.9	2.98	-	-	-	2.17	91.3
AGEn [36]	2.52	85.0	2.70	83.0	-	-	-	-
MV [30]	-	-	-	-	2.79	-	2.16	-
C3AE [6]	-	-	-	-	-	-	2.75	-
BridgeNet [25]	2.38	91.0	2.63	86.0	-	-	-	-
AVDL [38]	2.37	-	2.53	-	-	-	**1.94**	-
OL [26]	2.41	91.7	2.75	88.2	2.68	88.8	2.22	93.3
DRC-ORID [22]	2.26	93.8	**2.51**	89.7	2.58	89.5	2.16	93.5
Proposed	**2.13**	**94.2**	2.53	**90.4**	**2.53**	**90.5**	2.00	**95.0**

2.2. Global vs. Local Regression

Table S-2 compares the performances of global and local ρ-regressors on various facial age estimation datasets. In all tests, the local ρ-regressors improve the performances by refining the global regression results, for they are capable of learning diverse patterns more effectively. For example, in MORPH II setting D, the local regression lowers MAE by 7.4%. Moreover, in MORPH II setting C, which is the most challenging task, the local regression improves the MAE performance by 3.1%. Also, on the train split in CACD, it improves the MAE performance meaningfully by 7.4%.

Table S-2. Comparison of global and local regression on facial age estimation datasets.

	MORPH II (MAE/CS)				CLAP2015 (MAE/ϵ-error)		FG-NET (MAE/CS)	CACD (MAE)		UTK (MAE)	Adience (MAE/Accuracy)
	Setting A	Setting B	Setting C	Setting D	Validation	Test	LOPO	Train	Validation	Coral	SE
Global ρ-regressor	2.24/93.5	2.55/90.1	2.61/89.5	2.16/93.0	3.12/0.27	2.82/0.26	2.24/90.6	4.76	5.75	4.49	0.46/62.2
Local ρ-regressors	**2.13/94.2**	**2.53/90.4**	**2.53/90.5**	**2.00/95.0**	**2.95/0.26**	**2.77/0.25**	**2.23/91.1**	**4.41**	**5.68**	**4.37**	**0.45/62.6**

2.3. Performance According to K

Figure S-1 plots MAE according to the number K of NNs for predicting an initial estimate $\hat{\theta}^0(x)$. The impacts of K are negligible, as long as the initial estimation is conducted, *i.e.* $K \geq 1$. In the case of $K = 0$, we set $\hat{\theta}^0(x)$ to the midpoint of the entire range. In such a case, more than 10 iterations are often required for the convergence, and the MAE performance is degraded by about 0.007 as compared with the default $K = 5$.

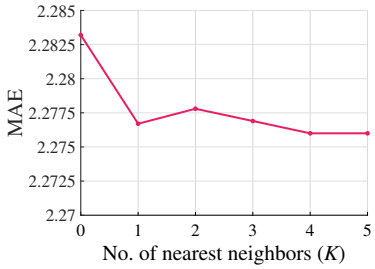


Figure S-1. Plot of MAE according to the number K of NNs on setting A of MORPH II.

2.4. Storage Costs

Table S-3 lists memory requirements for reference features. This additional memory is negligible in most practical applications. Especially, on MORPH II, only 469KB of additional memory is required for MWR using five local ρ-regressors.

Table S-3. Memory requirements for reference features for the test split of CLAP2015 and setting D of MORPH II.

	Global ρ-regressor	Local ρ-regressors
CLAP2015 (Test)	255KB	601KB
MORPH II (D)	225KB	469KB

2.5. Model Complexity

Table S-4 compares the complexity of the proposed ρ-regressors with those of conventional algorithms. We see that the ρ-regressors have relatively lightweight architecture.

Table S-4. Comparison of model complexities.

	DEX [33]	MV [30]	OL [26]	Global ρ-regressor	Local ρ-regressors
Parameters (M)	138	138	15.51	15.77	78.85

2.6. More Comparison with State-of-the-Arts

Table S-5 compares the proposed MWR for facial age estimation with DLDL-v2 [13] and DHAA [37] in terms of MAE. Except for settings B and D of MORPH II, MWR clearly outperforms these existing techniques. Note that DLDL-v2 uses MS-Celeb-1M [15], which is about 20 times bigger than IMDB-WIKI [33], for pre-training. Also, DHAA employs facial keypoints additionally for training. Nevertheless, without using a bigger dataset or extra information, the proposed algorithm provides competitive results. Especially, on FG-NET, the proposed algorithm outperforms DHAA by a significant MAE margin of 0.37.

Table S-5. MAE comparison on settings A, B, and D of MORPH II, the validation split of CLAP2015, and FG-NET.

Algorithm	MORPH II (A)	MORPH II (B)	MORPH II (D)	CLAP2015 (Val)	FG-NET
DLDL-v2 [13]	-	-	1.97	3.14	-
DHAA [37]	2.49	**2.49**	**1.91**	3.05	2.60
Proposed	**2.13**	2.53	2.00	**2.95**	**2.23**

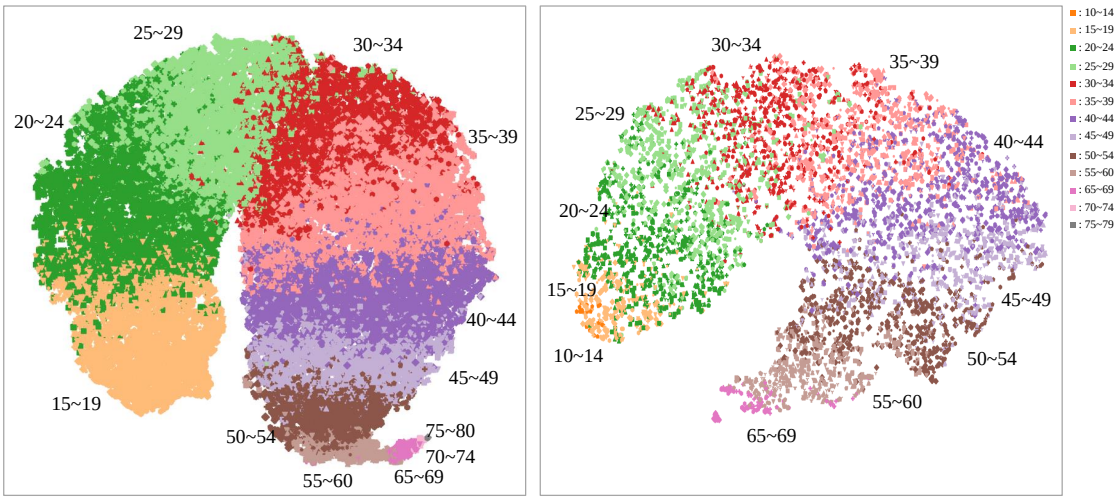
Next, Table S-6 compares the proposed MWR with PML [10]. For a fair comparison, as done in [10], we adopt ResNet34 [17] as the encoder. The proposed algorithm provides much better results than PML in both cases.

Table S-6. MAE comparison on settings A and D of MORPH II using the same backbone of ResNet34. IMDB-WIKI pre-training is not performed.

Algorithm	MORPH II (A)	MORPH II (D)
PML [10]	2.31	2.15
Proposed	**2.20**	**1.97**

2.7. Visualization of Feature Spaces

Figure S-2 visualizes the feature spaces of MORPH II and CACD. In both datasets, feature vectors are roughly aligned according to the ranks (ages).



MORPH II CACD

Figure S-2. t-SNE visualization of the feature spaces of MORPH II and CACD.

2.8. Examples of Selected References in Facial Age Estimation

Figure S-3 shows examples of reference pairs selected by the min γ and max γ schemes in facial age estimation. In Figure S-3(a), most faces look straight ahead without occlusion. Hence, the min γ scheme provides more accurate estimates. In contrast, in Figure S-3(b), images have overexposure or poor illumination, degrading age estimation performances.

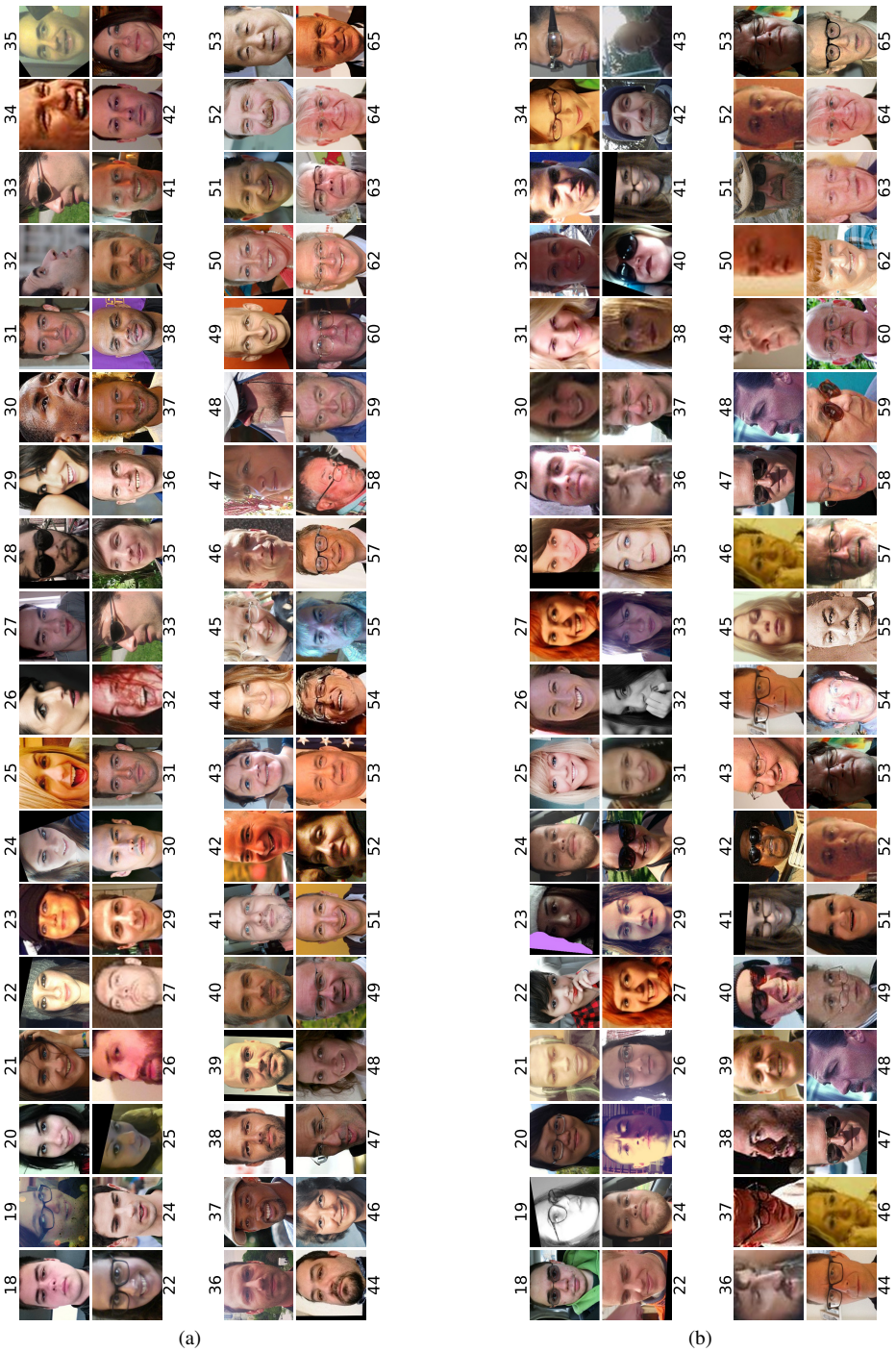


Figure S-3. Example of reference pairs selected by the (a) min γ and (b) max γ schemes on the CLAP2015 test split during the global regression. The numbers above or below the images indicate the corresponding ages.

3. Impacts of Applications

The proposed MWR is applicable to general ordinal regression tasks. In this paper, we apply the MWR to facial age estimation, historical color image classification, and aesthetic score regression. Especially, the proposed facial age estimator has diverse potential uses. For example, it can be used for forensic search [1] and social media [33]. Also, it can facilitate age-based customization of advertisements. However, as well as positive impacts, it has negative ones. For instance, age estimation errors lead to undesirable results, such as recommending unsuitable content. Moreover, although age information itself is not enough for identifying an individual, intermediate features of the proposed algorithm can be utilized as ancillary information in facial recognition systems [20], inducing serious problems such as unwanted surveillance and invasion of privacy [31]. Hence, ethical considerations should be made before the use of the proposed algorithm. We recommend using the proposed age estimator for research only until standard rules on the usage of facial analysis research are established by the governments [9] and international research bodies.

References

[1] Kanar Alkass, Bruce A. Buchholz, Susumu Ohtani, Toshiharu Yamamoto, Henrik Druid, and Kirsty L. Spalding. Age estimation in forensic sciences: Application of combined aspartic acid racemization and radiocarbon analysis. *Molecular & Cellular Proteomics*, 9:1022–1030, 2007. 6

[2] Axel Berg, Magnus Oskarsson, and Mark O'Connor. Deep ordinal regression with label diversity. In *Proc. IEEE ICPR*, 2021. 1

[3] Yoo ByungIn, Youngjun Kwak, Youngsung Kim, Changkyu Choi, and Junmo Kim. Deep facial age estimation using conditional multitask learning with weak label expansion. *IEEE Signal Process. Lett.*, 25:808–812, 2018. 2

[4] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. A ranking approach for human age estimation based on face images. In *Proc. IEEE ICPR*, 2010. 2

[5] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*, 2011. 2

[6] Zhang Chao, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3AE: Exploring the limits of compact model for age estimation. In *CVPR*, 2019. 2

[7] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 17:804–815, 2015. 1

[8] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age estimation. In *CVPR*, 2017. 2

[9] Yvette D. Clarke. H.R.2231 - Algorithmic Accountability Act of 2019. 2019. 6

[10] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classification. In *CVPR*, 2021. 4

[11] Raúl Díaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, 2019. 1

[12] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo J. Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *ICCV Workshops*, 2015. 1

[13] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *IJCAI*, 2018. 4

[14] Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR*, 2011. 2

[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 *ECCV*, 2016. 4

[16] Fredrik K. Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B. Schön. DCTD: Deep conditional target densities for accurate regression. In *arXiv preprint arXiv:1909.12297*, 2019. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[18] Han Hu, Anil K. Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40:2597–2609, 2017. 2

[19] Ivan Huerta, Carles Fernández, Carlos Segura, Javier Hernando, and Andrea Prati. A deep analysis on age estimation. *Pattern Recog. Lett.*, 68:239–249, 2015. 2

[20] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *ICBA*, 2004. 6

[21] Andreas Lanitis, Chris J. Taylor, and Timothy F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24:442–455, 2002. 1

[22] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *ICLR*, 2021. 1, 2

[23] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshops*, 2015. 1

[24] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *CVPR*, 2021. 1

[25] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. BridgeNet: A continuity-aware probabilistic network for age estimation. In *CVPR*, 2019. 1, 2

[26] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2020. 1, 2, 3

[27] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *CVPR*, 2018. 1

[28] Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *ICCV*, 2019. 1

[29] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *CVPR*, 2016. 2

[30] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, 2018. 1, 2, 3

[31] I.D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020. 6

[32] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *FGR*, 2006. 1, 2

[33] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126:144–157, 2018. 1, 2, 3, 4, 6

[34] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Yuille. Deep regression forests for age estimation. In *CVPR*, 2018. 1, 2

[35] Wei Shen, Kai Zhao, Yilu Guo, and Alan Yuille. Label distribution learning forests. In *NIPS*, 2017. 1

[36] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z. Li. Efficient group-n encoding and decoding for facial age estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40:2610–2623, 2018. 1, 2

[37] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li. Deeply-learned hybrid representations for facial age estimation. In *IJCAI*, 2019. 4

[38] 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. In *ECCV*, 2020. 1, 2

[39] Dong Yi, Zhen Lei, and Stan Z. Li. Age estimation by multi-scale convolutional network. In *ACCV*, 2014. 2

[40] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017b. 1

[41] Tan Zichang, Shuai Zhou, Jun Wan, Zhen Lei, and Stan Z. Li. Age estimation based on a single network with soft softmax of aging modeling. In *ACCV*, 2016. 2