Geometric Order Learning for Rank Estimation

Seon-Ho Lee **Nyeong-Ho Shin** **Chang-Su Kim**
School of Electrical Engineering
Korea University
`seonholee@mcl.korea.ac.kr`, `nhshin@mcl.korea.ac.kr`, `changsukim@korea.ac.kr`

A Broader Impacts

Recently, ethical concerns about the fairness of deep-learning-based systems have been raised (Castelvecchi, 2020; Roussi, 2020; Noorden, 2020). Especially, due to the intrinsic imbalance of facial datasets (Ricanek & Tesafaye, 2006; Zhang et al., 2017; Niu et al., 2016), most deep learning methods on facial analysis (Wen et al., 2020; Or-El et al., 2020) have unwanted gender or racial bias. The proposed algorithm is not free from this bias either when trained on such datasets. Hence, the bias should be resolved before any practical usage. Also, even though the proposed algorithm discovers some subclasses by ranking instances, these results should never be misinterpreted in such a way as to encourage any kind of discrimination. We recommend using the proposed algorithm for research only.

B Derivation of $L_{\mathrm{x}\prec\mathrm{y}}$ in (10)

From the metric constraint in (7), if $x \prec y$, it should be that $d_{\mathrm{e}}(h_x, h_y) > \gamma$, or equivalently

$$-d_{\mathrm{e}}(h_x, h_y) < -\gamma \tag{S.1}$$

Note that reference points and instances are roughly sorted according to their corresponding ranks in the embedding space by the order constraint in (5). Therefore, for $x \prec y$, h_x tends to be closer to each reference point r_i, $0 \le i \le \theta(x)$, than h_y is. In other words, for $x \prec y$, it is likely that

$$d_{\mathrm{e}}(r_i, h_x) - d_{\mathrm{e}}(r_i, h_y) < 0. \tag{S.2}$$

However, to satisfy the constraint in (S.1), we impose a stricter upper bound on the difference $d_{\mathrm{e}}(r_i, h_x) - d_{\mathrm{e}}(r_i, h_y)$ as follows. From the triangle inequality, we have

$$d_{\mathrm{e}}(r_i, h_y) \le d_{\mathrm{e}}(r_i, h_x) + d_{\mathrm{e}}(h_x, h_y), \tag{S.3}$$

which is equivalent to

$$-d_{\mathrm{e}}(h_x, h_y) \le d_{\mathrm{e}}(r_i, h_x) - d_{\mathrm{e}}(r_i, h_y). \tag{S.4}$$

Therefore, if

$$d_{\mathrm{e}}(r_i, h_x) - d_{\mathrm{e}}(r_i, h_y) < -\gamma, \tag{S.5}$$

the constraint in (S.1) is satisfied. This desirable condition in (S.5) is formulated into the first sum in $L_{\mathrm{x}\prec\mathrm{y}}$ in (10).

Similarly, for each reference point r_j, $\theta(y) \le j \le M-1$, it is likely that

$$d_{\mathrm{e}}(r_j, h_y) - d_{\mathrm{e}}(r_j, h_x) \le 0. \tag{S.6}$$

However, to satisfy the metric constraint, it is desirable that

$$d_{\mathrm{e}}(r_j, h_y) - d_{\mathrm{e}}(r_j, h_x) < -\gamma, \tag{S.7}$$

which is formulated into the second sum in $L_{\mathrm{x}\prec\mathrm{y}}$ in (10).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

C Implementation Details

C.1 Network Architecture

The structure of the encoder h is detailed in Table S-1, where '$k_h \times k_w$-s-c Conv' denotes the 2D convolution with kernel size $k_h \times k_w$, stride s, and c output channels. Similarly, '$k_h \times k_w$-s MaxPool' and '$k_h \times k_w$-s AvgPool' represent the 2D max pooling and 2D average pooling with a $k_h \times k_w$ kernel at stride s, respectively. Also, BN means batch normalization (Ioffe & Szegedy, 2015). The encoder is based on the VGG16 network and takes a 224 × 224 × 3 image as input.

Table S-1: The structure of the encoder h.

Layers	Output
3×3-1-64 Conv BN ReLU	224×224×64
3×3-1-64 Conv BN ReLU	224×224×64
3×3-2 MaxPool	112×112×64
3×3-1-128 Conv BN ReLU	112×112×128
3×3-1-128 Conv BN ReLU	112×112×128
3×3-2 MaxPool	56×56×128
3×3-1-256 Conv BN ReLU	56×56×256
3×3-1-256 Conv BN ReLU	56×56×256
3×3-1-256 Conv BN ReLU	56×56×256
3×3-2 MaxPool	28×28×256
3×3-1-512 Conv BN ReLU	28×28×512
3×3-1-512 Conv BN ReLU	28×28×512
3×3-1-512 Conv BN ReLU	28×28×512
3×3-2 MaxPool	14×14×512
3×3-1-512 Conv BN ReLU	14×14×512
3×3-1-512 Conv BN ReLU	14×14×512
3×3-1-512 Conv BN ReLU	14×14×512
14×14-1 AvgPool	1×1×512

C.2 Embedding Space Visualization

In Figure 1, ResNet18 is employed as an encoder, but the output dimension of its last fully connected layer is reduced to 3 for the visualization. The encoder takes a 28 × 28 × 3 image as input. We use the same training setting in Section 4.1.

Similarly, for the other 3D embedding space visualizations, including Figures 3 and 5, we modify the encoder in Table S-1 to append a fully connected layer with output dimension 3. We train the modified encoder using the same setting in Section 4.1.

C.3 Computing Environment

We do all experiments using PyTorch (Paszke et al., 2019) and an NVIDIA GeForce RTX 3090 GPU.

C.4 Training Time

We train the encoder until the loss converges. Table S-2 lists the training epochs and time for each dataset.

Table S-2: Training epochs and time.

	MORPH (setting A)	MORPH (setting C)	CACD (train split)	CACD (validation split)	UTK	Adience	HCI	Aesthetics
# epochs	250	150	10	30	50	80	150	50
Time (hrs)	5	30	7	1	3	4	1	3

D More Experimental Results

D.1 Hyper-Parameters

The proposed GOL algorithm has three hyper-parameters; τ in (1), γ in (7), and k in (14). Unless specified otherwise, we set $\tau = 0$ and $\gamma = 0.05$. Also, we fix $k = 50$ and $k = 16$ for facial age estimation tasks and the others, respectively. Let us describe how the hyper-parameters affect performances.

Hyper-parameter γ: Table S-3 compares the MAE scores at different γ's on the MORPH II, CACD, and UTK datasets. In this test, $\tau = 0$ and $k = 50$. Note that γ is a margin to impose the minimum distance between instances with a rank difference larger than τ in the embedding space. We see that $\gamma = 0.05$ performs well in general, so we set $\gamma = 0.05$ as the default option. Only for UTK, we set $\gamma = 0.25$, even though GOL outperforms the state-of-the-art MWR-G (4.49) at the other γ levels as well.

Table S-3: MAE performances according to γ on MORPH II, CACD, and UTK.

γ	MORPH (setting D)	CACD (validation split)	UTK
0.05	2.08	5.58	4.48
0.15	2.13	5.78	4.44
0.25	2.09	5.60	4.35

Hyper-parameter τ: Table S-4 compares the MAE scores at different τ's on MORPH II, CACD, and UTK. In this test, $\gamma = 0.05$ and $k = 50$. We see that $\tau = 0$ provides decent results. We hence set the default $\tau = 0$.

Table S-4: MAE performances according to τ on MORPH II, CACD, and UTK.

τ	MORPH II (setting D)	CACD (validation split)	UTK
0	2.08	5.58	4.48
1	2.12	5.71	4.45
2	2.18	5.99	4.46

Hyper-parameter k: Table S-5 compares the MAE scores according to k. In each age estimation task, there are more than 40 ranks, but there are no clear distinctive characteristics for each rank. Therefore, we set a relatively large $k = 50$ as the default option. However, note that the results are not very sensitive as long as $k \geq 18$.

Table S-5: MAE performances according to k on MORPH II, CACD, and UTK.

k	2	10	18	26	34	42	50	58
MORPH (setting D)	2.14	2.11	2.11	2.10	2.09	2.09	2.08	2.09
CACD (validation split)	5.73	5.62	5.59	5.59	5.58	5.58	5.58	5.57
UTK	4.63	4.52	4.51	4.50	4.49	4.48	4.48	4.49

Next, Table S-6 compares the MAE scores on different folds of the HCI dataset according to k. The HCI and aesthetics datasets contain only five ranks, respectively. Thus, we use a relatively small $k = 16$ for these tasks.

D.2 Embedding Spaces

Table S-7 compares the B2W and DRR scores on the test data of MORPH II, CACD, and Adience. GOL performs the best in all tests, as well as for the training data in Table 1. These results indicate that GOL can arrange the rank sets effectively in the embedding space to reflect their ordinal relationships even for unseen test instances.

Table S-6: MAE performances according to k on the HCI dataset. The performances for 5 folds out of 10 are listed.

k	4	8	12	16	20	24	28	32
Fold 1	0.528	0.532	0.512	0.508	0.500	0.504	0.504	0.504
Fold 2	0.584	0.596	0.588	0.584	0.584	0.584	0.588	0.588
Fold 3	0.548	0.544	0.540	0.532	0.532	0.528	0.528	0.528
Fold 4	0.516	0.512	0.512	0.508	0.508	0.508	0.512	0.512
Fold 5	0.532	0.532	0.528	0.524	0.532	0.536	0.540	0.540
Average	0.542	0.543	0.536	0.531	0.531	0.532	0.534	0.534

Table S-7: Comparison of B2W and DRR scores on the MORPH II, CACD and Adience test data.

	MORPH II (setting A)			CACD (validation split)			Adience		
Algorithm	B2W	$\text{DRR}_{1.0}$	$\text{DRR}_{0.5}$	B2W	$\text{DRR}_{1.0}$	$\text{DRR}_{0.5}$	B2W	$\text{DRR}_{1.0}$	$\text{DRR}_{0.5}$
ML (Schroff et al., 2015)	6.84	4.14	3.64	0.83	1.39	1.20	3.15	2.45	2.24
MV (Pan et al., 2018)	3.89	2.90	2.63	0.75	1.30	1.15	1.45	1.69	1.60
OL (Lim et al., 2020)	5.16	3.80	3.38	1.31	1.99	1.72	5.58	4.16	3.71
MWR-G (Shin et al., 2022)	7.23	4.85	4.20	1.09	1.80	1.55	4.71	3.63	3.25
Proposed GOL	9.47	5.21	4.54	1.71	2.29	1.97	11.49	7.24	6.39

D.3 Rank Estimation – Facial Age Estimation

Datasets: We provide facial age estimation results of GOL on four datasets: MORPH II (Ricanek & Tesafaye, 2006), CACD (Chen et al., 2015), UTK (Zhang et al., 2017), and Adience (Levi & Hassner, 2015). MORPH II has the Institutional Review Board approval. The other datasets were made for academic research purposes only. Any images getting deletion requests from the original owners will be discarded from the datasets. There are no name labels, except for CACD containing celebrity names. We exploit the datasets only for the performance assessment of GOL.

MORPH II (Ricanek & Tesafaye, 2006) is a popular dataset for facial age estimation, containing about 55,000 facial images of 13,617 subjects in the age range [16, 77]. In each image, the gender and race labels are annotated. Based on these labels, various evaluation protocols have been proposed. We employ the four evaluation settings A, B, C, and D (Lim et al., 2020; Lee & Kim, 2021; Shin et al., 2022).

- Setting A: 5,492 Caucasian images are randomly sampled and divided into training and testing sets with a ratio of 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 entire dataset is randomly split into five folds, subject to the constraint that the same person should belong to only one fold, and the 5-fold cross-validation is performed.
- Setting D: The whole dataset is randomly divided into five folds without any constraint, and the 5-fold cross-validation is performed.

CACD (Chen et al., 2015) provides 160k images from 2,000 celebrities. It is split into three subsets by celebrities: 1,800 for training, 80 for validation, and 120 for testing. We provide two results by training GOL on the train set and on the validation set, respectively, as done in (Rothe et al., 2018; Shen et al., 2018; Shin et al., 2022). The age range is [14, 62].

UTK (Zhang et al., 2017) consists of 20,000 facial images in a wide age range [0, 116]. We adopt the evaluation protocol in (Gustafsson et al., 2020; Berg et al., 2021).

Adience (Levi & Hassner, 2015) is used for age group estimation. There are 26,580 facial images from 2,284 subjects, which are grouped into 8 ordinal classes: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and over 60-year-olds. We employ the 5-fold subject-exclusive (SE) cross-validation evaluation setting in (Liu et al., 2018, 2019; Diaz & Marathe, 2019; Li et al., 2021).

More comparison: Table S-8 provides comparison results with more conventional algorithms on MORPH II. It is an extended version of Table 2.

Table S-8: Comparison of facial age estimation results in the four evaluation settings (A, B, C, and D) of MORPH II. Here, * means that IMDB-WIKI pre-training is performed.

	Setting A		Setting B		Setting C		Setting D	
	MAE	CS(%)	MAE	CS(%)	MAE	CS(%)	MAE	CS(%)
RED-SVM (Chang et al., 2010)	-	-	-	-	-	-	6.49	49.0
OHRank (Chang et al., 2011)	-	-	-	-	-	-	6.07	56.3
KPLS (Guo & Mu, 2011)	-	-	4.18	-	-	-	-	-
CPLF (Yi et al., 2014)	-	-	3.63	-	-	-	-	-
Huerta *et al.* (Huerta et al., 2015)	-	-	-	-	3.88	-	-	-
OR-CNN (Niu et al., 2016)	-	-	-	-	-	-	3.27	73.0
Tan *et al.* (Zichang et al., 2016)	-	-	3.03	-	-	-	-	-
Ranking-CNN (Chen et al., 2017)	-	-	-	-	-	-	2.96	85.0
DEX (Rothe et al., 2018)*	2.68	-	-	-	-	-	-	-
DMTL (Hu et al., 2017)	-	-	-	-	3.00	85.3	-	-
CMT (Yoo et al., 2018)	-	-	-	-	2.91	-	-	-
DRFs (Shen et al., 2018)	2.91	82.9	2.98	-	-	-	2.17	91.3
AGEn (Tan et al., 2017)*	2.52	85.0	2.70	83.0	-	-	-	-
MV (Pan et al., 2018)*	-	-	-	-	2.79	-	2.16	-
C3AE (Chao et al., 2019)*	-	-	-	-	-	-	2.75	-
BridgeNet (Li et al., 2019)*	2.38	91.0	2.63	86.0	-	-	-	-
AVDL (Wen et al., 2020)*	2.37	-	2.53	-	-	-	**1.94**	-
OL (Lim et al., 2020)*	2.41	91.7	2.75	88.2	2.68	88.8	2.22	93.3
DRC-ORID (Lee & Kim, 2021)*	2.26	**93.8**	**2.51**	89.7	2.58	89.5	2.16	93.5
MWR-G (Shin et al., 2022)*	2.24	93.5	2.55	**90.1**	2.61	89.5	2.16	93.0
Proposed GOL	**2.17**	**93.8**	2.60	89.3	**2.51**	**90.0**	2.09	**94.2**

D.4 Rank Estimation – HCI Classification

Table S-9 is an extended version of the HCI results in Table 4.

Table S-9: Accuracy (%) and MAE comparison on the HCI dataset.

	HCI	
Algorithm	Accuracy (%)	MAE
Frank & Hall (Frank & Hall, 2001)	41.4	0.99
Cardoso *et al.* (Cardoso & da Costa, 2007)	41.3	0.95
Palermo *et al.* (Palermo et al., 2012)	44.9	0.93
RED-SVM (Lin & Li, 2012)	35.9	0.96
Martin *et al.* (Martin et al., 2014)	42.8	0.87
OR-CNN (Niu et al., 2016)	38.7	0.95
CNNPOR (Liu et al., 2018)	50.1	0.82
GP-DNNOR (Liu et al., 2019)	46.6	0.76
DRC-ORID (Lee & Kim, 2021)	44.7	0.80
POE (Li et al., 2021)	54.7	0.66
MWR-global (Shin et al., 2022)	52.2	0.60
Proposed GOL	**56.2**	**0.55**

D.5 Rank Estimation – Aesthetic Score Regression

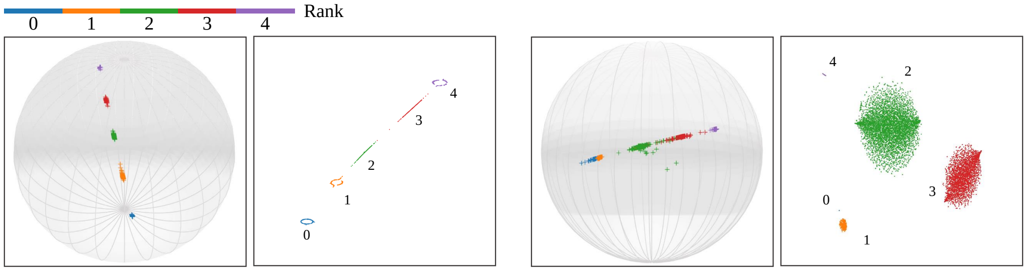
The aesthetics dataset (Schifanella et al., 2015) provides 15,687 image URLs on Flickr, where 13,929 images are available but the others are lost. Each image is annotated with a 5-scale aesthetic score. We adopt the 5-fold cross-validation. For training, we set the learning rate of all parameters, including reference points, to 10^{-6}. The other settings are the same as in Section 4.1. Table S-10 is an extended version of Table 5.

Table S-10: Accuracy (%) and MAE comparison on the aesthetics dataset.

	Nature		Animal		Urban		People		Overall	
Algorithm	Accuracy (%)	MAE	Accuracy (%)	MAE	Accuracy (%)	MAE	Accuracy (%)	MAE	Accuracy (%)	MAE
RED-SVM (Lin & Li, 2012)	70.7	0.31	61.1	0.41	65.4	0.37	61.2	0.41	64.6	0.38
CNNm (Liu et al., 2018)	71.0	0.31	68.0	0.34	68.2	0.36	71.6	0.32	69.5	0.33
OR-CNN (Niu et al., 2016)	69.8	0.31	69.1	0.33	66.5	0.35	70.4	0.31	69.0	0.33
CNNPOR (Liu et al., 2018)	71.9	0.29	69.3	0.32	69.1	0.33	69.9	0.32	70.1	0.32
SORD (Diaz & Marathe, 2019)	73.6	**0.27**	70.3	0.31	73.3	0.28	70.6	0.31	72.0	0.29
Li *et al.* (Li et al., 2021)	73.6	**0.27**	71.1	0.30	72.8	0.28	**72.2**	**0.29**	72.4	0.29
Proposed GOL	**73.8**	**0.27**	**72.4**	**0.28**	**74.2**	**0.26**	69.6	0.31	**72.7**	**0.28**

D.6 Analysis

More visualizations: Figure S-1 visualizes the embedding spaces of the proposed algorithm on the HCI and aesthetics datasets. For each dataset, the left subfigure shows the reduced 3D embedding space and the right subfigure is the t-SNE visualization (Maaten & Hinton, 2008) of original 512D embedding space.



(a) HCI

(b) Aesthetics

Figure S-1: Visualization of the embedding spaces for the HCI and aesthetics datasets.

Young, median, and old instances: Figure S-2 shows sampled instances from each rank set $\mathcal{X}_i$ of the Adience dataset. Specifically, Figures S-2 (a) and (c) list the instances in each $\mathcal{X}_i$ that are the farthest from the reference point r_i in the backward and forward directions, respectively. Also, Figures S-2 (b) shows the closest instance to r_i among all instances in $\mathcal{X}_i$. In Adience, each rank set $\mathcal{X}_i$ consists of instances in an age range, marked at the top of the figure. Thus, from each rank set, young, median, and old instances tend to be selected in (a), (b), and (c), respectively, which means that the instances are well sorted in the embedding space although their exact ages are not annotated.

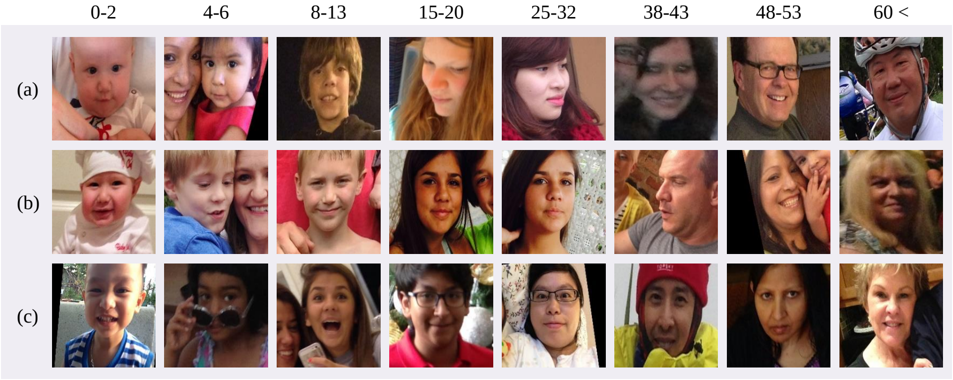


Figure S-2: Sampling (a) young, (b) median, and (c) old instances from each rank set of Adience.

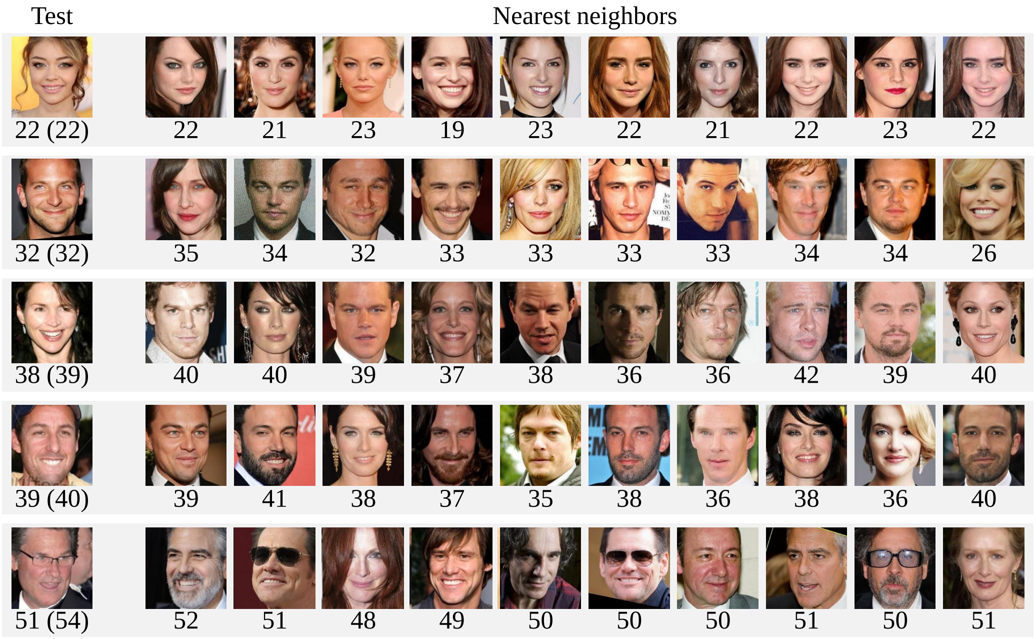
Testing time: Table S-11 lists the testing time of GOL according to the number of samples for the k-NN search on the MORPH II and CACD datasets. Note that the distances to all samples can be computed efficiently in a parallel manner. Therefore, the proposed algorithm performs fast even with 44,000 samples.

Table S-11: Testing time according to the number of samples for the k-NN search.

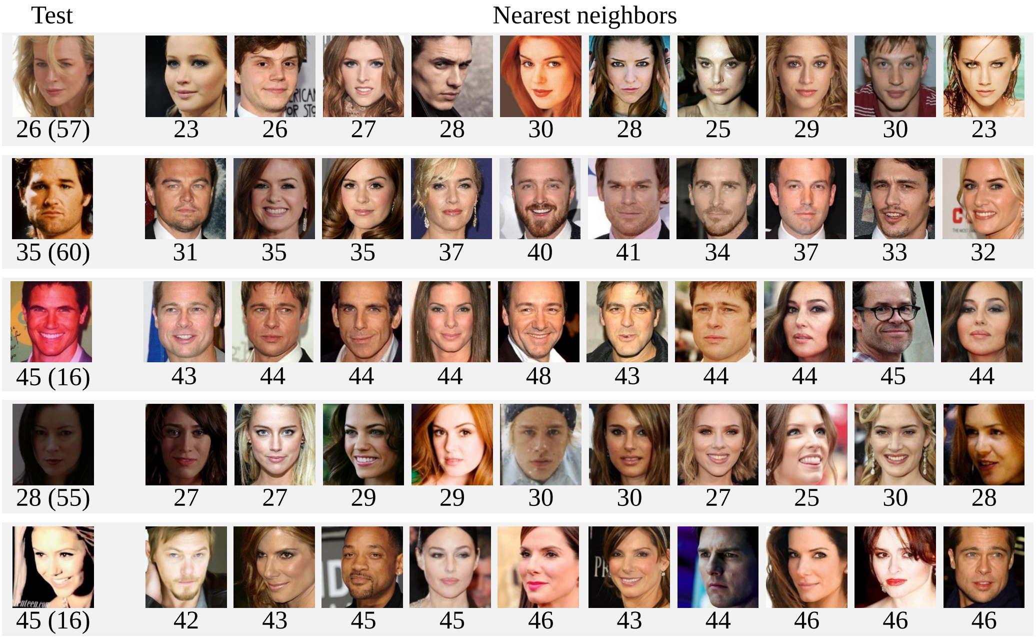
	MORPH (setting A)	MORPH (setting B)	MORPH (setting C)	CACD (validation split)
# samples	4,394	7,000	44,000	7,600
Time (ms)	0.05	0.06	0.08	0.06

Limitations: Figure S-3 shows some success and failure cases of GOL on the CACD dataset. In Figure S-3 (a), the ages of test instances are predicted accurately with absolute errors less than 4. However, in Figure S-3 (b), GOL performs poorly on some hard cases, in which various factors, such as poor illumination, overexposure, and low-quality photographs, hinder accurate estimation.

Also, since GOL predicts a rank based on the k-NN search, it suffers when there are insufficient training instances. For example, GOL yields relatively poor results on MORPH II setting B, which consists of 7,000 training samples and 14,000 test samples. Similarly, GOL tends to yield less accurate estimates on minority classes, such as toddlers and elders, since most age estimation datasets contain fewer instances in such classes. These limitations of GOL might be alleviated by generating pseudo-references in the embedding space, which we leave as future work.



(a)



(b)

Figure S-3: (a) Success and (b) failure cases of the proposed algorithm in facial age estimation. For each test image x, the estimate $\hat{\theta}(x)$ is reported with the ground-truth $(\theta(x))$ within the parentheses. Also, the ten nearest neighbors of x are shown with their ages.

References

Axel Berg, Magnus Oskarsson, and Mark O'Connor. Deep ordinal regression with label diversity. In *ICPR*, 2021. 4

Jaime S. Cardoso and Joaquim F. Pinto da Costa. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8:1393–1429, 2007. 5

Davide Castelvecchi. Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–349, 2020. 1

Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. A ranking approach for human age estimation based on face images. In *ICPR*, 2010. 5

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

Zhang Chao, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3AE: Exploring the limits of compact model for age estimation. In *CVPR*, 2019. 5

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. 4

Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-CNN for age estimation. In *CVPR*, 2017. 5

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, 2019. 4, 6

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *ECML-PKDD*, 2001. 5

Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR*, 2011. 5

Fredrik K. Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B. Schon. Energy-based models for deep probabilistic regression. In *ECCV*, 2020. 4

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. 5

Ivan Huerta, Carles Fernández, Carlos Segura, Javier Hernando, and Andrea Prati. A deep analysis on age estimation. *Pattern Recog. Lett.*, 68:239–249, 2015. 5

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2

Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identitiy decomposition. In *ICLR*, 2021. 4, 5

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *CVPR Workshops*, 2015. 4

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. 5

Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *CVPR*, 2021. 4, 5, 6

Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2020. 4, 5

Hsuan-Tien Lin and Ling Li. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 24(5):1329–1367, 2012. 5, 6

Yanzhu Liu, Adams W. K. Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *CVPR*, 2018. 4, 5, 6

Yanzhu Liu, Fan Wang, and Adams W. K. Kong. Probabilistic deep ordinal regression based on Gaussian processes. In *CVPR*, 2019. 4, 5

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 6

Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. In *Proc. ACM ICMR*, 2014. 5

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output CNN for age estimation. In *CVPR*, 2016. 1, 5, 6

Richard V. Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834): 354–358, 2020. 1

Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, 2020. 1

Frank Palermo, James Hays, and Alexei A. Efros. Dating historical color images. In *ECCV*, 2012. 5

Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, 2018. 4, 5

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaisonet, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NIPS*, 2019. 2

Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *FGR*, 2006. 1, 4

Rasmus Rothe, Radu Timofte, and Luc V. Gul. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis.*, 126(2-4):144–157, 2018. 4, 5

Antoaneta Roussi. Resisting the rise of facial recognition. *Nature*, 587(7834):350–353, 2020. 1

Rossano Schifanella, Miriam Redi, and Luca M. Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of Flickr pictures. In *ICWSM*, 2015. 5

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 4

Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Yuille. Deep regression forests for age estimation. In *CVPR*, 2018. 4, 5

Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *CVPR*, 2022. 4, 5

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(11): 2610–2623, 2017. 5

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, 5

Dong Yi, Zhen Lei, and Stan Z. Li. Age estimation by multi-scale convolutional network. In *ACCV*, 2014. 5

Byungin Yoo, 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. 5

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 1, 4

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. 5