Distribution Highlighted Reference-based Label Distribution Learning for Facial Age Estimation

Supplementary Material

A. Facial Age Estimation Datasets

We used three challenging datasets for evaluation: MORPH II [36], UTKFace [50], and CACD [6]. These datasets each contain over 20,000 facial images, which is considered large-scale in facial age estimation. We also used IMDB-WIKI [37, 38] for pre-training. This appendix provides detailed explanations of these datasets.

A.1. Details of Datasets

MORPH II [36] is the most common dataset for age estimation. It contains 55,134 facial images from 13,617 individuals whose ages range from 16 to 77. On this dataset, we used four evaluation settings, A, B, C, and D, which are explained in the next subsection.

UTKFace [50] contains over 20,000 facial images with a wide age span, ranging from 0 to 116. These images cover large variations in pose, facial expression, illumination, occlusion, resolution, etc. We used the evaluation protocol implemented in previous studies [1, 18, 24, 40].

CACD [6] contains 163,446 facial images from 2,000 celebrities. The dataset is split into three subsets based on the celebrities: 1,800 for training, 80 for validation, and 120 for testing. The age range is from 14 to 62. We provided two separate results by training the DNN on the training and validation sets, respectively, as done in several previous studies [24, 39, 40].

IMDB-WIKI [37, 38] consists of 523,051 facial images. Although this is the largest facial dataset with age labels, it is known to have too much label noise. Therefore, in general, IMDB-WIKI is not suitable for evaluations [44]. Instead, it is commonly used to pre-train the DNN for many state-of-the-art methods [23, 25, 26, 28, 30, 48]. Following this practice, we pre-trained the DNN using IMDB-WIKI.

A.2. Evaluation Settings of MORPH II

For MORPH II, we evaluated DHRL in four widely used settings. In each setting, to ensure fair comparisons, we used the same data splitting provided by Shin *et al.* [40].

- **Setting A.** 5,492 images of the Caucasian race were selected and then randomly divided into two non-overlapping parts: 80% for training and 20% for testing.
- **Setting B.** About 21,000 images of Africans and Caucasians were selected to satisfy two constraints: the ratio between Africans and Caucasians should be 1 : 1, and that between females and males should be 1 : 3. They were split into three disjoint subsets S1, S2, and S3. The training and testing were repeated twice: 1) training on S1, testing on S2+S3, and 2) training on S2, testing on S1+S3. The average result of the two experiments is reported.
- **Setting C.** This setting is the 5-fold cross-validation on the entire dataset. Images were randomly split into five folds, but the same individual's images should belong to only one fold. The average result of the five experiments is reported.
- **Setting D.** The entire dataset was randomly divided into five folds without any restrictions. Thus, this setting is similar to Setting C, but the same individual's images may belong to both training and test sets. The average result of the five experiments is reported.

B. Details of IMDB-WIKI Pre-training

Following previous studies [23, 25, 26, 28, 30, 48], we pre-trained the DNN on IMDB-WIKI for better initialization. Among all IMDB-WIKI images, we used 311,085 facial images selected to mitigate label noise, which were provided by Paplhám and Franc [32].

We used the DNN trained on ImageNet for initializing parameters before pre-training. In the pre-training with IMDB-WIKI, the SGD optimizer with a momentum of 0.9 and a batch size of 32 was used. The optimization was conducted for 10 training epochs with a consistent learning rate of 0.001. We used the MAE loss defined in Eq. (3) as the loss function.

C. Details of Optimization for Existing LDL Methods

Many LDL methods [14, 15, 25, 30] were evaluated in only one or two settings in MORPH II. For a fair and comprehensive comparison, we re-evaluated these methods across all settings. Note that the official source codes for these methods are not publicly available, so we used reproduction codes provided online by Paplhám and Franc [32]. For optimization, we used the SGD optimizer with a momentum of 0.9. The remaining hyper-parameters are listed in Table 10 and were consistent with the original papers, except for hyper-parameters not explicitly described. We experimentally adjusted the hyper-parameters not explicitly described to achieve high estimation performance. While the original paper of DLDL-v2 [15] used a different DNN architecture than VGG-16, we used VGG-16 for a fair comparison. Since proper hyper-parameters may vary depending on the DNN architecture, we compared the estimation perfor-

Method	Learning rate	Weight decay	Batch size	Epochs	L.r. decay epochs	γ
DLDL [14]	$0.001^{\dagger}$	0.0005	32	10	None	—
Mean-var. [30]	0.001	0	64	30	15	0.1
DLDL-v2 [15]	0.001	0	64	120	60	0.1
Uni.-con. [25]	0.01	0	128	174	29, 58, $\cdots$	0.5

Table 10. Hyper-parameters used in our re-evaluations. γ represents multiplicative factor for learning rate (L.r.) decay. † indicates that learning rate for last layer of DNN was multiplied by 10.

Method	Setting A	Setting B	Setting C	Setting D
DLDL [14]	2.48	2.85	2.75	2.38 (2.42)
Mean-var. [30]	2.28	2.64	2.62 (2.79)	2.13 (2.16)
DLDL-v2 [15]	2.18	2.58	2.62	1.94
Uni.-con. [25]	2.34	3.42	2.80	2.31 (1.86)

Table 11. Our re-evaluation results in four evaluation settings (A, B, C, and D) of MORPH II. Values within parentheses represent original results reported in their respective papers.

Method	Setting B	Setting C	Setting D
DLDL	2.85 → 2.55	2.75 → 2.53	2.38 → 2.04
DLDL-v2	2.58 → 2.57	2.62 → 2.59	1.94 → **1.91**
Mean-var.	2.64 → 2.63	2.62 → 2.56	2.13 → 2.07
Uni.-con.	3.42 → 2.90	2.80 → 2.65	2.31 → 2.16
Expectation	2.68 → **2.52**	2.62 → **2.52**	2.15 → 1.94

Table 12. MAE results of various reference DNNs on DHRL in Settings B to D. Values to left and right of arrow indicate results of reference and target DNNs, respectively.

Method	Setting A	Setting B	Setting C	Setting D
DLDL	$2.52_{\pm 0.05}$	$2.85_{\pm 0.01}$	$2.75_{\pm 0.02}$	$2.38_{\pm 0.03}$
Mean-var.	$2.32_{\pm 0.04}$	$2.64_{\pm 0.05}$	$2.62_{\pm 0.03}$	$2.13_{\pm 0.02}$
DLDL-v2	$2.18_{\pm 0.03}$	$2.58_{\pm 0.01}$	$2.62_{\pm 0.02}$	$1.94_{\pm 0.01}$
Uni.-con.	$2.64_{\pm 0.45}$	$3.42_{\pm 0.72}$	$2.80_{\pm 0.22}$	$2.31_{\pm 0.30}$
DHRL	$2.10_{\pm 0.04}$	$2.52_{\pm 0.01}$	$2.52_{\pm 0.02}$	$1.94_{\pm 0.01}$

Table 13. Facial age estimation results on four evaluation settings of MORPH II with standard deviations.

mance between the original hyper-parameters and those in DHRL. We found that the hyper-parameters in DHRL gave a better performance and therefore used them for DLDL-v2, as shown in Table 10. For hyper-parameters other than those related to optimization, we closely followed the original paper settings. An exception was the unimodal-concentrated loss [25], where the original paper set the hyper-parameter λ (distinct from λ in DHRL) to 1000. However, this setting failed to converge and resulted in poor estimation performance in our re-evaluation. We observed that setting $\lambda = 10$ allowed the optimization to converge, although the results remained unstable (see Fig. 4). We therefore adopted this value instead. For all methods, we utilized the same pre-trained VGG-16 on ImageNet and IMDB-WIKI as DHRL for initialization.

Table 11 lists the re-evaluated estimation results for each LDL method. Values within parentheses represent the original results reported in their respective papers. As shown in the second (DLDL) and third (the mean-variance loss) rows, our re-evaluation reproduced the original results or achieved a slightly higher estimation performance. We believe these results indicate the reliability of our re-evaluation. We also found that the unimodal-concentrated loss [25] did not match the original results. This discrepancy may have resulted from differences in training details, such as the deep learning framework, image pre-processing, or optimization details in the pre-training. Unfortunately, the exact reason remains unclear, as the official source code is not available. We emphasize that our re-evaluations are based on the publicly available reproduction codes by Paplhám and Franc [32]. Thus, we present our re-evaluation results in Table 2 as the officially reproducible performance.

D. Results of Various Reference DNNs on DHRL in Settings B to D

Table 12 shows the MAE results using the reference DNNs pre-trained with various LDL methods in Settings B to D. We observed that DHRL consistently improves the performance of the reference DNN, which is similar to the trend in Table 1. Furthermore, except in Setting D, the expectation regression loss for the pre-training achieved the best performance in DHRL.

E. Statistical Analysis of Performance Gaps

We statistically analyzed the performance gap between DHRL and other LDL methods using the results of MORPH

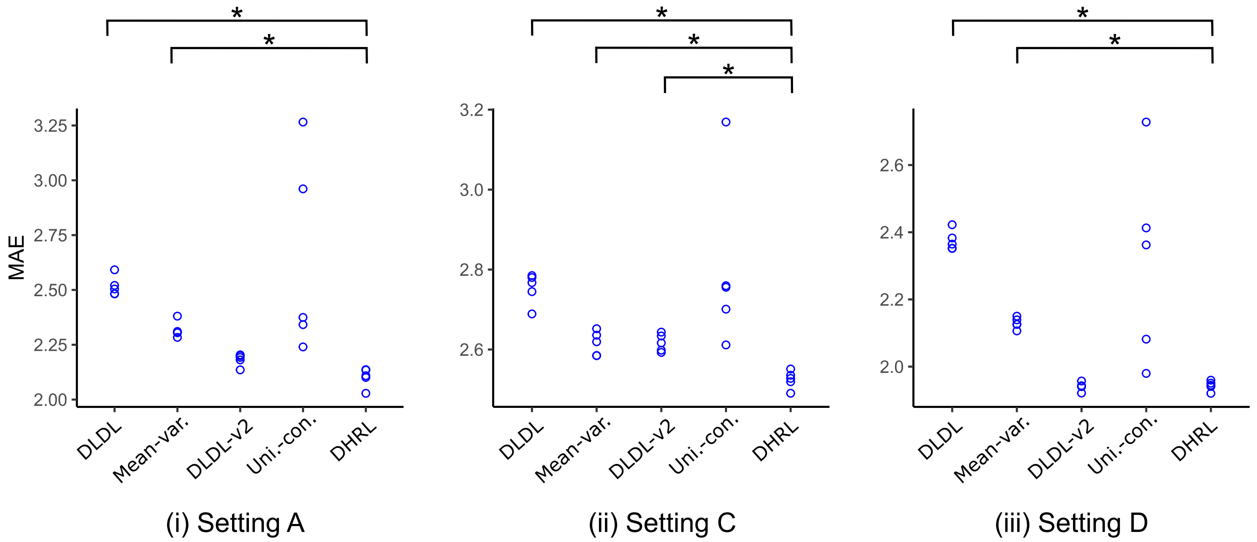


Figure 4. Scatter plots of MAE values obtained by five facial age estimation methods for Settings A, C, and D of MORPH II. Lines marked with * denote statistically significant results of DHRL when compared with each method, as determined by paired t-test at significance level of $p < 1.25 \times 10^{-2}$. Here, significance level was corrected by dividing original level 5.0×10^{-2} by 4 (*i.e*., Bonferroni correction).

II. Table 13 summarizes performance on four settings in MORPH II, including standard deviations. Since Setting A does not adopt multiple data splits (as described in Appendix A.2), we extended this setting to 5-fold cross-validation, where one fold corresponds to the original Setting A split. As shown in the table, DHRL exhibits relatively small standard deviations, indicating that its results are stable.

To provide a more detailed statistical analysis, we present scatter plots of the MAE values for Settings A, C, and D with 5-fold cross-validation in Fig.4[5]. The figure also reports the results of paired t-tests between DHRL and each method. Since four t-tests were conducted for each setting, we applied the Bonferroni correction by dividing the significance level 5.0×10^{-2} by 4. The corrected significance level is $p < 1.25 \times 10^{-2}$. In Fig. 4, the lines marked with * denote statistically significant results of DHRL when compared with each method. From the figure, DHRL shows consistent statistical significance over DLDL and the mean-variance loss. In contrast, no significant difference is observed against DLDL-v2 in Setting D, which is also evident from the scatter plots. Furthermore, in several cases, there are no statistical significances, even though differences can be visible in the scatter plots and in Table 13, *i.e*., the unimodal-concentrated loss across all settings, and DLDL-v2 in Setting A. We consider this mainly due to the small sample size in the t-test (only five folds). In particular, as described in Appendix C, the unimodal-concentrated loss shows unstable training with large variance, which reduces the statistical power in limited-sample settings. This likely explains why significance was not detected, even though DHRL qualitatively outperforms the unimodal-concentrated loss. As a future direction, we plan to conduct larger-scale experiments by increasing the number of folds to analyze the statistical support for DHRL.

Method	MORPH II	UTKFace	CACD
Cross-entropy	2.81	4.38	3.96
Regression	2.83	4.72	4.06
OR-CNN [29]	2.83	4.40	4.01
DLDL [14]	2.81	4.39	3.96
DLDL-v2 [15]	2.82	4.42	3.96
SORD [10]	2.81	4.36	3.96
Mean-var. [30]	2.83	4.42	4.07
Uni.-con. [25]	2.78	4.47	4.10
DHRL	**2.69**	**4.29**	**3.92**

Table 14. Comparison of facial age estimation results on MORPH II, UTKFace, and CACD using same evaluation setting as Paplhám and Franc [32].

F. Evaluation Results under a Different Setting

In Tables 2 and 3, we followed the evaluation setting of the prior works [23, 40]. Recently, Paplhám and Franc [32] provided a new evaluation setting of facial age estimation using different data splits and DNN architecture (ResNet-50). In this appendix, we evaluated DHRL using this evaluation setting to confirm DHRL's versatility for diverse conditions.

In this experiment, we used the same hyper-parameters

[5] Since Setting B contains only two samples (results), it is not suitable for statistical analysis. Therefore, we excluded it.

Method	MAE
Reference DNN	2.26
DHRL w/o NAE & DSM ($\tau = 1/2$)	2.19
DHRL w/o NAE & DSM ($\tau = 1$)	2.18
DHRL w/o NAE & DSM ($\tau = 2$)	2.19
DHRL w/o NAE & DSM ($\tau = 1/2$ and 1)	2.18
DHRL w/o NAE & DSM ($\tau = 1/2$ and 2)	2.15
DHRL w/o NAE & DSM ($\tau = 1$ and 2)	2.17
DHRL w/o DSM ($\tau = 1/2$)	2.14
DHRL w/o DSM ($\tau = 1$)	2.16
DHRL w/o DSM ($\tau = 2$)	2.17
DHRL w/o DSM ($\tau = 1/2$ and 1)	2.15
DHRL w/o DSM ($\tau = 1/2$ and 2)	2.15
DHRL w/o DSM ($\tau = 1$ and 2)	2.13
DHRL w/o NAE	2.14
DHRL	**2.10**

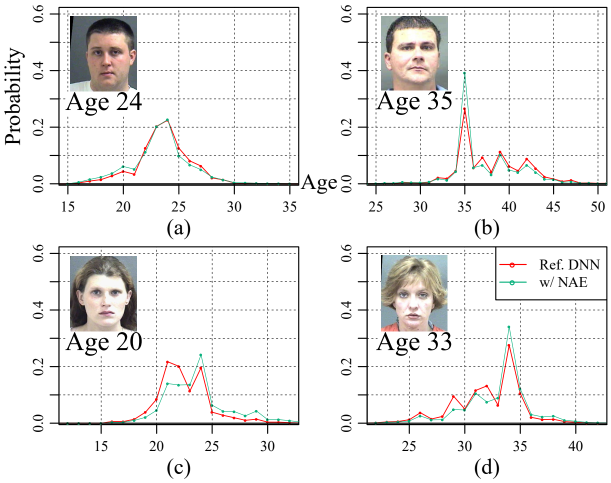
Table 15. Full results of our ablation study, including variations with other temperatures and combinations of two temperatures.

	(a) Age 24	(b) Age 35	(c) Age 20	(d) Age 33
DHRL	23.73 (**-0.27**)	36.61 (**+1.61**)	22.15 (**+2.15**)	33.31 (**+0.31**)
DLDL-v2	22.99 (-1.01)	37.92 (+2.92)	22.70 (+2.70)	33.47 (+0.47)
Mean-var.	24.64 (+0.64)	38.96 (+3.96)	23.20 (+3.20)	31.51 (-1.49)

Table 16. Estimated ages for each facial image in Fig. 3. Values within parentheses represent age estimation errors. Bold indicates smallest value of absolute estimation error in each column.

of DHRL defined in Subsection 4.1, *i.e.*, $\varepsilon = 0.005$, $K = 3$, $\tau_1 = 2$, $\tau_2 = 1/2$, $\tau_3 = 1$, and $\lambda = 4$. For the training of the reference and target DNNs in DHRL, we mainly follow the settings in Subsection 4.1, but found that using the exact same optimization settings resulted in poor estimation performance. This is because the suitable learning rate varies between ResNet and VGG. In fact, the initial learning rates used in their original papers differ for training with ImageNet [19, 41]. Therefore, we set the initial learning rate to 0.01 while keeping other hyper-parameters unchanged. Furthermore, we used the IMDB-WIKI pre-trained DNN provided by Paplhám and Franc [32] for initialization.

Table 14 lists the estimation results of DHRL and existing methods on MORPH II, UTKFace, and CACD. From the table, we found that DHRL consistently outperformed the existing facial age estimation methods even in this different evaluation setting. From these results, together with those in Tables 2 and 3, we confirm that the proposed method demonstrates effectiveness across a wide range of evaluation settings.



(a) (b) (c) (d)

Figure 5. Predicted label distributions of reference DNN (*i.e.*, input-dependent reference label distributions) and input-dependent reference label distributions generated using NAE for four facial images. X-axis represents discrete age values.

G. Full Results of Ablation Study

Table 4 in the main paper shows the selected results of the ablation study, and here, we provide the full results in Table 15, including the variations with other temperature (τ) values and the combinations of two temperatures. DHRL achieved the best MAE score, as expected. We also observed that (1) all variations of DHRL without both NAE and DSM outperformed the reference DNN, and (2) adding NAE generally improved the estimation performance, excluding the $\tau = 1/2$ and 2 variations. These observations complement those in Table 4, further supporting the effectiveness of DHRL, including NAE and DSM.

H. Additional Results from Analysis of DHRL

Table 16 lists the estimated ages for each facial image in the analysis of label distribution (Fig. 3). Values within parentheses represent age estimation errors. As shown, DHRL achieved the smallest absolute errors.

Figure 5 shows the input-dependent reference label distributions, which were computed using the reference DNN, as red lines. Compared with the label distributions of the mean-variance loss, the input-dependent reference label distributions exhibited fewer constraint misses (*e.g.*, shown in (b)). From this observation, the expectation regression loss using only Eq. (3) may help reduce the negative effects of heuristic constraints, as we hypothesized in Subsection 4.2. In Fig. 5, we also show the input-dependent reference label distributions generated using NAE as green lines. We observed that NAE modified the input-dependent reference label distributions in (b) to (d). These modifications appear to produce more proper label distributions, such as increas-

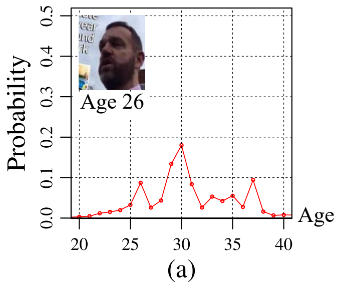


Figure 6. Failure cases of label distribution predicted by DHRL. X-axis represents discrete age values.

ing the probability for the correct age label (b), creating a more unimodal distribution (c), and smoothing fluctuated areas (d). These results qualitatively confirm the effect of NAE that highlights the label ambiguity hidden in the label distributions.

From the above observations, NAE appears to generate the label distributions that accurately model the label ambiguity. Interestingly, NAE did not lead to improved estimation performance. Specifically, the MAE score with NAE was 2.34, which is worse than that of the reference DNN (2.26). Nonetheless, the target DNN achieved a better estimation performance when using NAE, as shown in Table 4. This indicates that the usefulness of the input-dependent reference label distribution in DHRL does not always correlate directly with its estimation performance.

In Fig. 6, we present failure cases of DHRL, where overly jagged label distributions were predicted for two test facial images in UTKFace. One possible reason is that the inputs are challenging test images. Specifically, in case (a), the right half of the face is barely visible, while in case (b), the upper part of the head is cropped. In such cases, DHRL may fail to predict appropriate distributions due to limited cues for age estimation. Consequently, the predicted ages (30.7 and 59.3) largely deviate from the ground truth. These analyses suggest that incorporating data augmentation techniques, such as Cutout [9], could further improve the performance of DHRL. This direction is interesting and promising for future work.

I. Optimization Details and Additional Results for Analysis using EfficientNets

This appendix first describes the optimization details for DHRL with EfficientNet-B0 and B2. For pre-training with IMDB-WIKI, we used the same optimization settings described in Appendix B. For the training of the reference and target DNNs in DHRL, we observed that using the same optimization settings as VGG led to poor estimation performance. This is because the suitable learning rate varies between EfficientNet and VGG. In fact, the initial learning rates used in their original papers differ for training with ImageNet [41, 43]. Therefore, we set the initial learning rate to 0.01 while keeping other hyper-parameters unchanged. As a result, DHRL using EfficientNets achieved a better estimation performance than when using VGG-16, as shown in Table 8.

	MAE
Reference DNN (EfficientNet-B0)	2.24
Target DNN (VGG-16)	2.11

Table 17. Performance of DHRL using EfficientNet-B0 and VGG-16 as reference and target DNNs.

Method	Memory [MiB]	Training time [sec/epoch]
Reference DNN	19,962	38.9
DLDL-v2	19,962	43.8
Mean-var.	19,962	41.4
DHRL	25,980	81.2

Table 18. Comparison of computational costs in Setting A of MORPH II.

We also evaluated the performance of DHRL using EfficientNet-B0 and VGG-16 as the reference and target DNNs. Namely, these two DNNs have different architectures from each other. Table 17 shows the performance of the reference and target DNNs in Setting A of MORPH II. As shown, even when the reference DNN had a significantly different architecture, the target DNN obtained an MAE of 2.11, which is almost the same as the original setting, using VGG-16 as the reference DNN (MAE of 2.10). Since DHRL uses the reference DNN as a generator of input-dependent reference label distributions, it is likely not sensitive to the reference DNN's architecture. Therefore, DHRL can be considered robust to differences in architecture between the reference and target DNNs.

J. Details of Additional Computational Costs of DHRL in Training

Table 18 lists the computational costs of DHRL and existing methods during training in Setting A of MORPH II. These experiments were conducted using a single RTX A6000 GPU. As shown, DHRL increases the memory usage by about 1.3 times and the training time by about 2 times. This is because DHRL uses the reference DNN, which is not utilized in the existing methods. However, we argue that DHRL is worth the increased cost. When increasing the number of training epochs to be the same GPU-hours, the reference DNN, DLDL-v2, the mean-variance loss, and DHRL resulted in 2.21, 2.17, 2.33, and **2.10** in MAE. These results emphasize the effectiveness of DHRL even on the fixed computational cost.

	10,000	50,000	100,000	Full
Reference DNN	5.71	4.93	4.55	4.52
Target DNN	5.35	4.49	4.33	4.30

Table 19. MAE scores of reference and target DNNs using different training data sizes on CACD. “Full” setting means using all 145,275 training images, which corresponds to MAE result in Table 3.

K. Effect of Training Data Size on DHRL

Finally, we evaluated the effect of training data size on DHRL. To this end, we used the large-scale CACD dataset and created subsets of 10,000, 50,000, and 100,000 training images by removing training images of the same individual. For each subset, both the reference and target DNNs were trained, and the results are shown in Table 19. Here, the “Full” setting means using all 145,275 training images, which corresponds to the result in Table 3. From the table, we found that increasing the number of training images improves the performance of both reference and target DNNs, as expected. Furthermore, DHRL consistently improves the target DNN regardless of the training data size. This result indicates that the effectiveness of DHRL is robust to variations in training data size.