GATE: Gaussian-Attentive Transformer for Uncertainty-Aware Age Estimation

– Supplementary Material –

A. Experimental Details

A.1. Datasets

This section provides additional descriptions for each facial age estimation dataset.

- MORPH II [28]: It contains about 55,000 facial images of 13,617 subjects, with ages ranging from 16 to 77 years. Based on these labels, multiple evaluation settings have been established [17, 30], as outlined below. Following the prevailing protocol in recent literature, the experiments are conducted on setting A in the main paper. Additional evaluations on settings B–D are included in this supplementary material to verify that GATE maintains consistently strong performance across all established protocols.
  - Setting A: 5,492 images of Caucasians are sampled and then randomly split into training and testing sets with a ratio of 8:2.
  - Setting B: About 21,000 images of Caucasians and Africans are randomly chosen such that the ratio between Caucasians and Africans is 1:1 and that between females and males is 1:3. The dataset is divided into three subsets (S1, S2, S3). Training and testing are repeated twice: (1) training on S1 and testing on S2+S3, and (2) training on S2 and testing on S1+S3.
  - Setting C: The whole dataset is randomly divided into five folds, ensuring that images of the same person belong to only one fold. Then, 5-fold cross-validation is performed.
  - Setting D: The whole dataset is randomly divided into five folds without any restriction, and 5-fold cross-validation is performed.
- CLAP2015 [7, 35]: It is designed for apparent age estimation. The apparent age of each image was rated by at least ten annotators, and the mean rating was used as the ground truth. This dataset also provides the standard deviation of the ratings for each image. It contains 4,691 facial images, which are divided into 2,476 for training, 1,136 for validation, and 1,079 for testing. The age range spans from 3 to 85 years.
- AgeDB [22]: It contains about 12,200 images for training, with a maximum bin density of 353 images, and a minimum bin density of 1. The validation and test sets contain about 2,100 images, respectively. The ages range from 0 to 101 years.
- UTK [38]: It consists of approximately 20,000 facial images covering a wide age range from 0 to 116 years. We adopt the same evaluation protocol as in [2, 8], using 13,147 images for training and 3,287 for testing.
- CACD [3]: It contains about 160K images of 2,000 celebrities, which are divided by identity into three subsets: 1,800 for training, 80 for validation, and 120 for testing. The age range spans from 14 to 62 years.
- Adience [13]: This dataset is used for age group estimation. It contains 26,580 facial images of 2,284 subjects, which are categorized into eight ordinal age groups: 0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, and 60+. We employ the 5-fold subject-exclusive (SE) cross-validation protocol following [5, 15, 19, 20].
- FG-NET [10]: It provides 1,002 color or grayscale facial images of 82 individuals, with ages ranging from 0 to 69 years. It is used exclusively for testing to assess generalization performance on an unseen dataset.

A.2. Implementation Details

All models are implemented in PyTorch and trained on NVIDIA RTX 4090 GPUs. We initialize the encoder with the ViT-B backbone from the CLIP algorithm [27], and the rank tokens in all variants are initialized from the CLIP class-token weights. We use the AdamW optimizer [21] with a weight decay of 5×10^{-2}, and apply a cosine annealing scheduler [9] to adjust the learning rate during training. For data augmentation, we use color jitter, random grayscale conversion, and random cropping to a resolution of 224×224, following [36].

A.3. Training and Hyperparameters

Dataset-specific training: For the single-domain configuration (GATE-s), each dataset is trained independently using the hyperparameters summarized in Table 1.

Table 1. Training configurations for dataset-specific GATE-s.

Dataset	MORPH II	CLAP2015	AgeDB	UTK	CACD	Adience
Learning rate	5×10^{-6}	5×10^{-6}	5×10^{-6}	5×10^{-7}	5×10^{-7}	5×10^{-7}
Batch size	64	64	64	16	64	64
Epochs	100	100	100	30	30	10
λ in (2)	0.1	0.1	0.1	0.1	0.1	0.1
τ in (6)	2.0	2.0	2.0	2.0	2.0	2.0
k in (9)	2	42	18	58	58	58

Mixed training: For the mixed-domain configurations (GATE-m and GATE-u), all datasets are trained jointly within a single model while maintaining a balanced composition of samples from different domains in each mini-batch. This strategy mitigates domain bias and promotes stable optimization across heterogeneous datasets. All experiments are trained for 30 epochs with a total batch size of 96 for both GATE-m and GATE-u. Training employs the AdamW optimizer with a weight decay of 0.05 and a cosine-annealing learning rate schedule starting from 5×10^{-6}. Because the dataset collection includes MORPH II and Adience, both organized with 5-fold splits, mixed-domain training is repeated five times using a common fold index applied to both datasets. The final performance is reported as the average over the five folds to ensure fair and robust evaluation. The key hyperparameters are fixed to $\lambda = 0.1$ in Eq. (2) and $\tau = 2.0$ in Eq. (6), and the per-dataset parameter k follows the same values as in Table 1.

A.4. Evaluation Metrics

In the main paper, mean absolute error (MAE) is used as the primary evaluation metric for all datasets. In this supplementary material, we additionally report cumulative score (CS) results for the MORPH II dataset in Table 2 and ϵ-error results for the CLAP2015 dataset in Table 3 to provide a more comprehensive evaluation. CS denotes the percentage of samples whose absolute errors are within a tolerance level l, which is set to 5 following [11, 17, 34]. In CLAP2015, each image includes the standard deviation of its apparent age ratings from multiple annotators, where a larger deviation indicates higher uncertainty. To incorporate this uncertainty, the ϵ-error is defined as

$$\epsilon\text{-error} = 1 - \exp\left(-\frac{(\hat{y}-y)^2}{2\sigma^2}\right), \tag{1}$$

where $\hat{y}$ and y denote the predicted and ground-truth ages of an image, and σ is the standard deviation of its annotator ratings. This metric penalizes errors more heavily when the annotation confidence is high and less when it is low. We report the average ϵ-error over all test images.

B. More Experiments

B.1. Comparative Assessment

Table 2 reports the comparison results on the four evaluation settings (A–D) of the MORPH II dataset. GATE-s achieves state-of-the-art performance on settings A, B, and D, and performs on par with the best existing method on setting C. These results confirm that GATE-s maintains strong accuracy and robustness under all established MORPH II protocols.

Table 2. Comparison of age estimation results on the four evaluation settings (A, B, C, and D) of the MORPH II dataset.

Algorithm	Backbone	Setting A		Setting B		Setting C		Setting D	
		MAE	CS(%)	MAE	CS(%)	MAE	CS(%)	MAE	CS(%)
AGEn [32]	VGG-16	2.52	85.0	2.70	-	-	-	-	-
BridgeNet [14]	VGG-16	2.38	91.0	2.63	86.0	-	-	-	-
OL [17]	VGG-16	2.41	91.7	2.75	88.2	2.68	88.8	2.22	93.3
C3AE [37]	custom CNN	-	-	-	-	-	-	2.75	-
DRC-ORID [11]	VGG-16	2.26	93.8	2.51	90.4	2.53	**90.5**	2.00	95.0
POE [15]	VGG-16	2.56	88.8	2.81	87.4	2.67	88.8	2.22	92.9
PML [4]	ResNet-34	2.31	-	-	-	-	-	2.15	88.0
AVDL [34]	ResNet-17	2.37	-	-	-	-	-	1.94	88.9
MWR [30]	VGG-16	2.13	94.2	2.53	90.4	2.53	**90.5**	2.00	95.0
GOL [12]	VGG-16	2.17	93.8	2.60	89.3	**2.51**	90.0	2.09	94.2
OrdinalCLIP [16]	VGG-16	2.31	90.2	2.68	85.9	-	-	-	-
RnC [36]	ResNet-18	2.44	88.3	2.88	83.2	2.80	84.0	2.15	91.7
CLOC [24]	ResNet-50	3.45	79.3	3.72	76.7	3.02	84.7	2.49	89.6
L2RCLIP [33]	ViT-B	2.13	-	-	-	-	-	-	-
NumCLIP [6]	ViT-B	2.11	91.6	2.51	86.9	2.74	84.8	1.93	92.8
MCGRL [31]	CNN+GCN	-	-	-	-	-	-	1.89	90.1
GATE-s	ViT-B	**1.98**	**94.5**	**2.45**	**90.7**	2.52	90.3	**1.64**	**96.0**

Table 3 shows the comparison results on the validation and test splits of the CLAP2015 dataset. GATE-s achieves the best performance in terms of both MAE and ϵ-error on both splits, outperforming recent CLIP-based approaches such as NumCLIP and L2RCLIP as well as traditional CNN methods. These results indicate that GATE-s generalizes effectively to apparent age estimation, maintaining strong accuracy under annotation uncertainty.

Table 3. Comparison on the validation and test splits of CLAP2015.

Algorithm	Backbone	Validation		Test	
		MAE	ϵ-error	MAE	ϵ-error
AgeNet [18]	GoogLeNet	3.33	0.29	-	0.26
Zhu et al. [39]	GoogLeNet	-	0.31	-	0.29
DEX [29]	VGG-16	3.25	0.28	-	0.26
AGEn [32]	VGG-16	3.21	0.28	2.94	0.26
BridgeNet [14]	VGG-16	2.98	0.26	2.87	0.26
MWR [30]	VGG-16	2.95	0.26	2.77	0.25
GOL [12]	VGG-16	3.88	0.34	3.38	0.31
RnC [36]	ResNet-18	-	-	4.72	
CLOC [24]	ResNet-50	5.09	0.43	4.45	0.39
L2RCLIP [33]	ViT-B	-	-	2.62	-
NumCLIP [6]	ViT-B	2.75	0.24	2.55	0.22
GATE-s	ViT-B	**2.57**	**0.22**	**2.42**	**0.21**

B.2. Ablation on Query-Update Layers

Figure 1 illustrates how patch tokens from intermediate ViT layers are used in the transformer decoder for age-query updates. Table 4 reports the MAE performance of GATE-s when different sets of ViT layers are used for this update step. The best performance is obtained with tokens from the 7th-8th-9th layers, which provide an effective balance between local detail and global semantics. Therefore, these layers are adopted as the default configuration in all experiments.

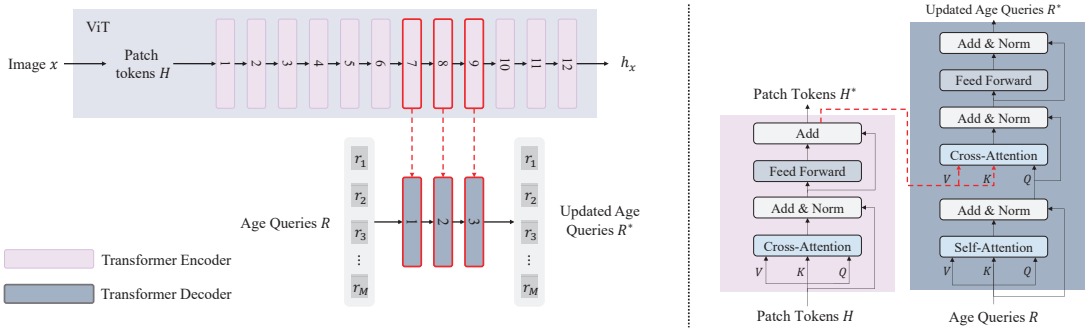

Figure 1. Architecture of the proposed transformer decoder. Patch tokens from intermediate ViT layers are fed into the transformer decoder during the age-query update. As shown in Table 4, the optimal performance is achieved when using tokens from the 7th-8th-9th layers.

Table 4. Performance of GATE-s with different query-update layers on the CLAP2015 dataset.

Query-update layers	1st-2nd-3rd	6th-9th-12th	7th-8th-9th	10th-11th-12th
MAE	2.418	2.422	**2.415**	2.437

B.3. Attention Visualization Across ViT Layers

Figure 2 visualizes how the age query attends to patch tokens extracted from different layers of the ViT encoder across four age groups in the CLAP2015 dataset. Each column corresponds to an age group — Baby (0–7), Teenager (8–19), Adult (20–49), and Elder (50+) — and each row, (a), (b), or (c), visualizes attention to patch tokens extracted from the 7th, 8th, or 9th layer of the ViT encoder, respectively. As the layer depth increases, attention expands from localized regions, such as the eyes, to broader facial areas, including the cheeks and mouth, indicating that deeper layers capture more global and semantically informative cues relevant to age estimation.

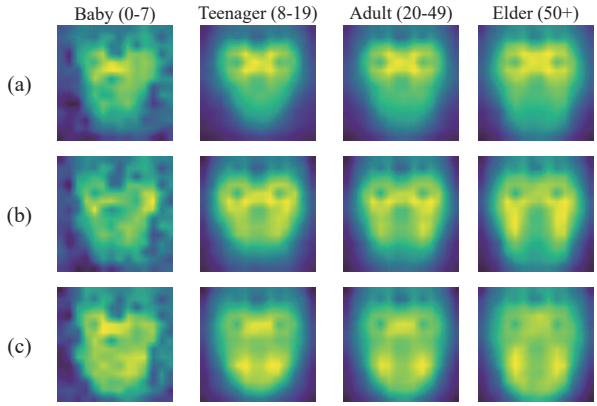

Figure 2. Layer-wise visualization of attention heatmaps showing how the age query attends to patch tokens in the transformer decoder. Rows (a)–(c) correspond to patch tokens extracted from the 7th, 8th, and 9th layers of the ViT encoder, respectively.

B.4. Hyperparameter Analysis

Performance according to λ in (2): Table 5 reports the MAE of GATE-s on the CLAP2015 dataset when varying the regularization weight λ in Eq. (2). Extremely small values ($\lambda = 10^{-3}$ or 10^{-2}) provide insufficient regularization, while a large weight ($\lambda = 1$) also degrades performance. The best result (2.415 MAE) is obtained at $\lambda = 10^{-1}$, indicating that a moderate balance between data fitting and regularization yields the most stable convergence. This setting is therefore adopted in all experiments.

Table 5. MAE performance of GATE-s according to λ on the CLAP2015 dataset. Lower values indicate better performance.

λ	10^{-3}	10^{-2}	10^{-1}	1	3	5
MAE	2.452	2.450	**2.415**	2.427	2.446	2.425

Performance according to τ in (6): Following [36], we use $\tau = 2.0$ as the default temperature, but we additionally analyze the effect of varying τ. Table 6 reports the MAE performance on CLAP2015 for different temperature values. Too small temperatures (*e.g.*, $\tau = 0.1$) produce overly sharp ranking probabilities and significantly degrade accuracy, whereas increasing τ to moderate levels stabilizes optimization and improves performance. The best result is achieved at $\tau = 2.0$, while excessively large values (*e.g.*, $\tau = 5.0$) overly flatten the ranking distribution and weaken the supervision signal.

Table 6. MAE performance of GATE-s according to τ on the CLAP2015 dataset. Lower values indicate better performance.

τ	0.1	0.5	1.0	2.0	5.0
MAE	3.523	2.466	2.418	**2.415**	2.432

Performance according to k in (9): Table 7 reports the MAE of GATE-s on CLAP2015 when varying the number of nearest neighbors k used in the k-NN inference stage. Small values (*e.g.*, $k = 2$ or 10) make the predictions sensitive to noise from individual neighbors, whereas large values (*e.g.*, $k = 50$ or 58) lead to over-smoothing and reduced discriminability. The best performance (2.415 MAE) is obtained at $k = 42$, suggesting that an intermediate neighborhood size provides the most stable aggregation of local embeddings. Accordingly, $k = 42$ is used as the default inference setting for CLAP2015, while the remaining datasets follow the values listed in Table 1 (*i.e.*, $k = 2$ for MORPH II, 18 for AgeDB, and 58 for larger datasets such as UTK, CACD, and Adience), reflecting differences in sample scale and intra-domain variability.

Table 7. MAE performance of GATE-s according to k on the CLAP2015 dataset. Lower values indicate better performance.

k	2	10	18	26	34	42	50	58
MAE	2.535	2.444	2.423	2.426	2.424	**2.415**	2.426	2.424

B.5. Analysis on GWA

Effect on adjacent-age discrimination: A potential concern of Gaussian Window Attention (GWA) is that a broadened attention scope under high uncertainty may impair fine-grained discrimination between adjacent ages. However, as shown in Table 8, GWA consistently improves cumulative score (CS) across all tolerance levels. Notably, the gain at small tolerances such as CS@1 indicates that GWA enhances adjacent-age discrimination rather than degrading it. This is because GWA adaptively modulates its attention scope based on uncertainty: samples requiring fine-grained discrimination yield a small σ, resulting in sharp and localized attention.

Table 8. Comparison of cumulative score (CS) between standard cross-attention and the proposed GWA on AgeDB.

Method	CS@0	CS@1	CS@2	CS@3	CS@4	CS@5
Standard cross-attention	6.45	20.61	32.43	44.07	54.21	63.22
GWA	**6.87**	**21.31**	**34.16**	**45.47**	**56.21**	**64.25**

B.6. Comparison with Uncertainty-Aware Baselines

To position GATE with respect to prior uncertainty-aware age estimation approaches, we compare it with previously reported Bayesian baselines on the APPA-REAL[1] benchmark. As shown in Table 9, GATE-s significantly outperforms Dropout, Dropconnect, and Deep Ensembles in terms of RMSE.

Table 9. Comparison with prior uncertainty-aware methods on APPA-REAL.

Method	Dropout [23]	Dropconnect [23]	Ensembles [23]	GATE-s
RMSE (↓)	12.272	12.968	11.101	**3.767**

Since prior results on APPA-REAL use different backbones, we further implement MC Dropout and ensemble baselines with the same ViT-B backbone on CLAP2015. As shown in Table 10, GATE achieves lower MAE with single-pass inference.

Table 10. Comparison with bayesian baselines on CLAP2015 under the same ViT-B backbone. Inference time is reported in parentheses.

Method	MC Dropout	Ensembles	Two-stage	GATE-s
MAE (time)	2.452 (172.19s)	2.422 (44.30s)	2.419	**2.415** (12.41s)

Training stability. Beyond predictive performance, we analyze the training behavior of the proposed uncertainty-guided mechanism. The uncertainty parameter σ is initialized with a relatively large value and optimized with a higher learning rate than the other parameters, after which it gradually decreases as prediction errors are reduced. This results in an implicit curriculum, where attention is initially broad and progressively becomes more focused over the course of training, preventing unreliable early-stage uncertainty estimates from dominating the attention mechanism. We further compare this end-to-end optimization strategy with a two-stage approach that first learns μ and σ, and then enables GWA. As shown in Table 10, the two-stage strategy does not outperform the proposed end-to-end training, further supporting the stability of the overall optimization.

B.7. Subgroup Analysis

Demographic subgroup analysis: We analyze the behavior of GATE-m across demographic subgroups. Figure 3 presents the performance and the corresponding predicted uncertainty across gender and race groups. The most sample-rich race group exhibits higher σ, indicating that the predicted uncertainty is not driven by demographic under-representation; it does not simply reflect demographic bias

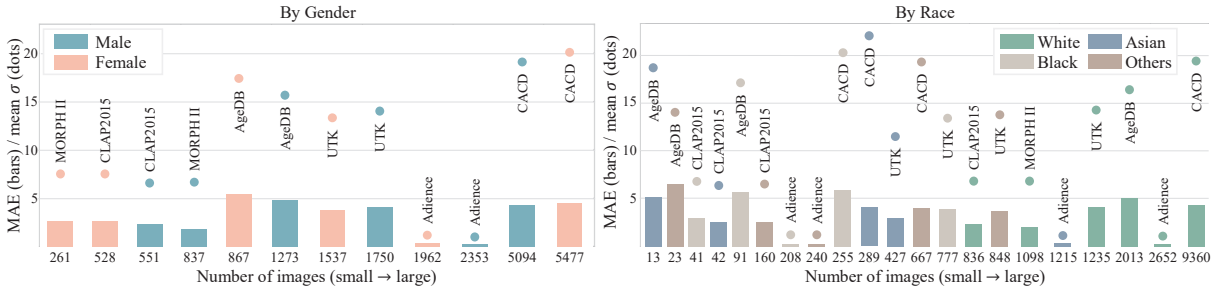

Figure 3. Subgroup analysis of GATE-m across gender and race. The predicted uncertainty does not monotonically follow subgroup frequency, suggesting that it captures visual ambiguity rather than demographic sample imbalance.

Robustness under class imbalance: To further analyze whether the predicted uncertainty is merely correlated with sample frequency, we report the MAE and average predicted σ across age groups on CLAP2015. As shown in Table 11, incorporating uncertainty yields lower or comparable MAE in data-scarce age groups than a standard cross-attention baseline, indicating improved robustness in tail classes. Also, the predicted σ is not strictly proportional to age frequency, suggesting that uncertainty reflects visual ambiguity rather than sample imbalance.

Table 11. Per-age-group analysis on CLAP2015. We report MAE and the average predicted uncertainty σ for each age group.

Method \ Age Group (# images)	0–9 (3)		10–19 (133)		20–29 (457)		30–39 (255)		40–49 (114)		50–59 (83)		60+ (34)	
	MAE	σ	MAE	σ	MAE	σ	MAE	σ	MAE	σ	MAE	σ	MAE	σ
Standard cross-attention	2.00	–	1.95	–	1.93	–	2.62	–	3.75	–	3.23	–	3.62	–
GATE-s	1.67	8.42	2.06	4.33	2.00	5.25	2.50	6.24	3.54	6.51	2.98	6.52	3.74	7.43
GATE-m	2.00	4.64	1.90	5.12	1.89	6.20	2.67	7.49	3.80	8.37	2.84	8.55	3.58	9.32

We also compare GATE-s and GATE-m in Table 11. Note that GATE-m achieves comparable or improved performance across age groups while maintaining consistent uncertainty estimates. This indicates that unified training does not amplify demographic bias and can instead help mitigate it by leveraging complementary distributions across datasets.

B.8. Complexity

Training time: All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. Training each dataset independently with GATE-s requires a total of 587 minutes summed over all domains, whereas unified multi-domain training with GATE-m or GATE-u completes in 80 minutes. This corresponds to nearly a 7× improvement in overall training efficiency, demonstrating the computational advantage of the unified architecture.

Table 12. Comparison of per-epoch training times for domain-specific (GATE-s) and mixed-domain (GATE-m and GATE-u) variants on a single RTX 4090 GPU.

Training configuration	Model	MORPHII	CLAP2015	AgeDB	UTK	CACD	Adience	All
Domain-specific training	GATE-s	52 min	43 min	137 min	70 min	450 min	14 min	587 min
Mixed training	GATE-m	–	–	–	–	–	–	80 min
	GATE-u	–	–	–	–	–	–	80 min

Model size: Table 13 compares the parameter counts of transformer-based age estimation models. Prior approaches such as SwinFace [25], NumCLIP [6], and Faceptor [26] require 109–179M parameters, whereas GATE is more compact with approximately 100M parameters. Notably, the multi-domain (GATE-m) and unified (GATE-u) variants introduce almost no additional parameters compared to the single-domain model (GATE-s), demonstrating that the proposed architecture maintains scalability across domains without increasing model size. This compact design highlights the efficiency of GATE relative to existing transformer-based methods.

Table 13. Model efficiency comparison in terms of parameter count.

Model	Backbone	# parameters (M)
SwinFace [25]	Swin-T	109.3
NumCLIP [6]	ViT-B	124
Faceptor [26]	ViT-B	178.9
GATE-s	ViT-B	100.17
GATE-m	ViT-B	100.18
GATE-u	ViT-B	100.17

C. Limitations

Figures 4–9 present qualitative examples from all six datasets, showing samples with similar ground-truth ages but different uncertainty levels. The hard samples are intentionally selected to illustrate situations where GATE produces large estimation errors, revealing factors that make age prediction inherently ambiguous:

- Atypical facial appearance – Unusually youthful or prematurely aged facial structure weakens the correlation between texture cues and chronological age.
- Pose and partial visibility – Large head rotations, self-occlusion, cropped facial regions, or blocked landmarks (e.g., eyes, cheeks, nasolabial folds) suppress key semantic evidence required for age reasoning.
- Lighting, shadows, and contrast – Harsh illumination, directional shadows, overexposure, or underexposure distort skin texture and conceal fine-scale age cues.
- Expression-induced deformation – Broad smiles, stretched expressions, pursed lips, or furrowed brows modify local geometry and introduce misleading age signals.
- Cosmetics, facial hair, and accessories – Heavy makeup, beards, hats, and glasses obscure or alter structural features that strongly correlate with age.
- Resolution limits and imaging artifacts – Blurry inputs, strong compression, motion blur, or low-pixel-density crops reduce the encoder's ability to extract high-frequency or localized aging cues.

These conditions reduce the reliability of visual evidence, leading GATE to predict larger variances σ and produce less accurate age estimates in general. While the proposed uncertainty mechanism adapts to such ambiguity, these hard samples highlight inherent limitations of appearance-based age estimation.

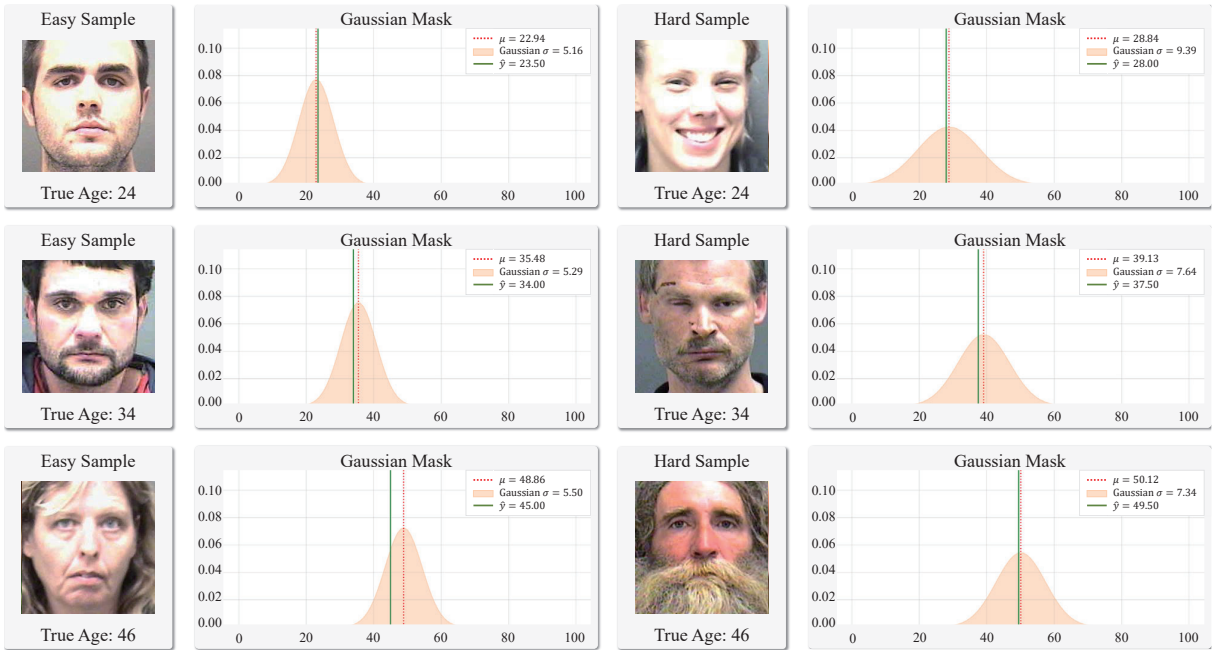

Figure 4. Qualitative examples from MORPH II. Examples with similar ground-truth ages but different prediction uncertainties (σ). Each shows the input face, ground-truth age y, predicted age $\hat{y}$, and predicted mean μ and standard deviation σ. Smaller σ values indicate confident, easy cases, whereas larger σ reflect greater uncertainty and prediction difficulty.

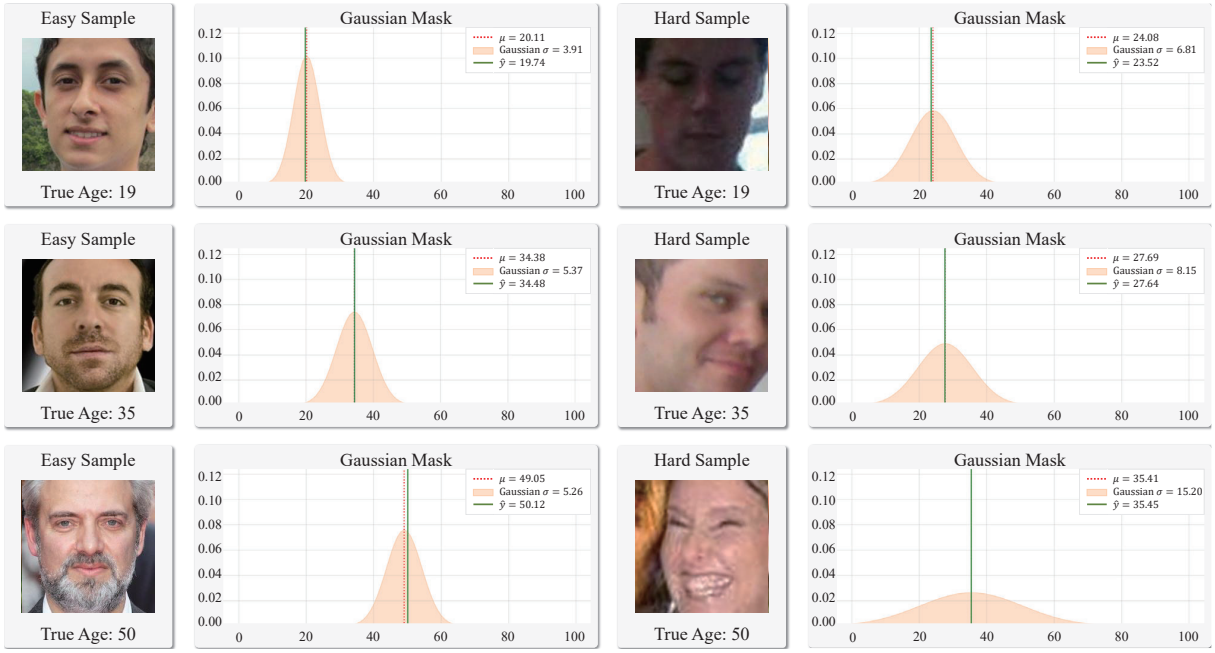

Figure 5. Qualitative examples from CLAP2015. Examples with similar ground-truth ages but different prediction uncertainties (σ). Each shows the input face, ground-truth age y, predicted age $\hat{y}$, and predicted mean μ and standard deviation σ. Smaller σ values indicate confident, easy cases, whereas larger σ reflect greater uncertainty and prediction difficulty.

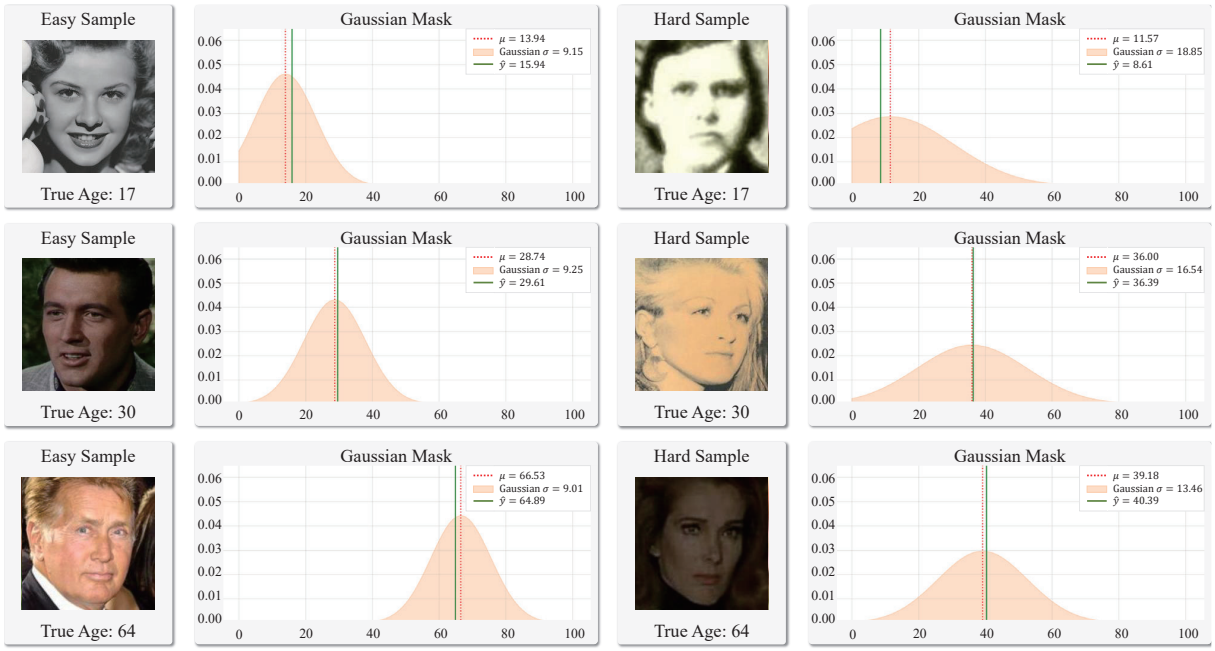

Figure 6. Qualitative examples from AgeDB. Examples with similar ground-truth ages but different prediction uncertainties (σ). Each shows the input face, ground-truth age y, predicted age $\hat{y}$, and predicted mean μ and standard deviation σ. Smaller σ values indicate confident, easy cases, whereas larger σ reflect greater uncertainty and prediction difficulty.

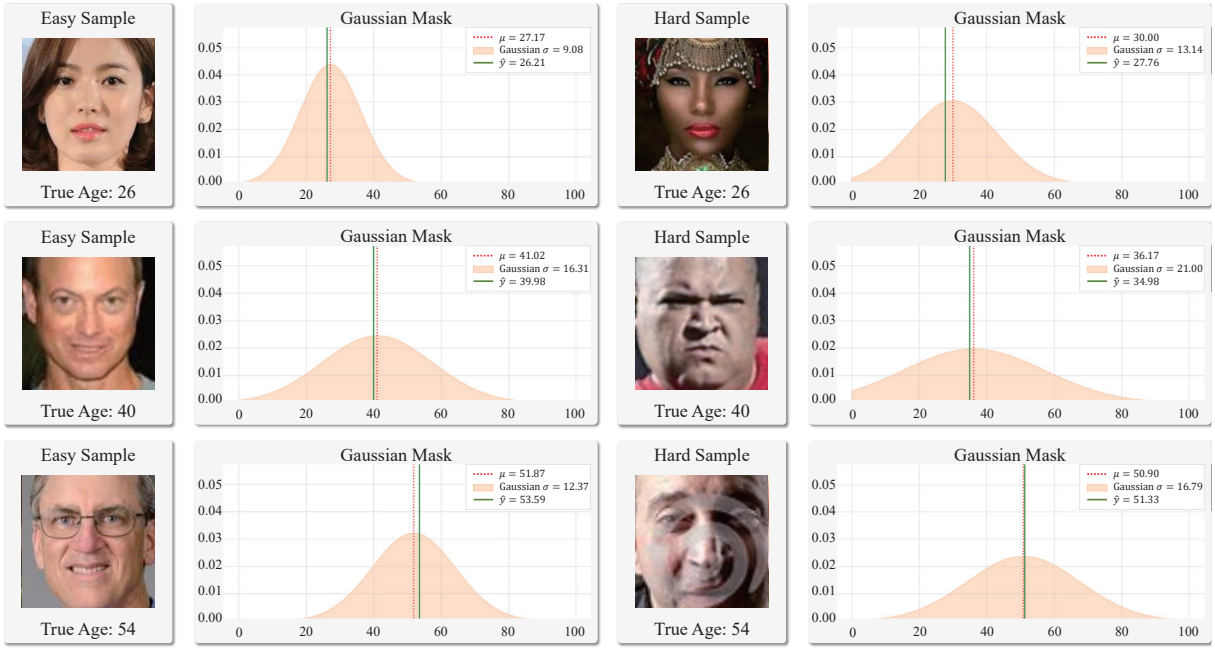

Figure 7. Qualitative examples from UTK. Examples with similar ground-truth ages but different prediction uncertainties (σ). Each shows the input face, ground-truth age y, predicted age $\hat{y}$, and predicted mean μ and standard deviation σ. Smaller σ values indicate confident, easy cases, whereas larger σ reflect greater uncertainty and prediction difficulty.

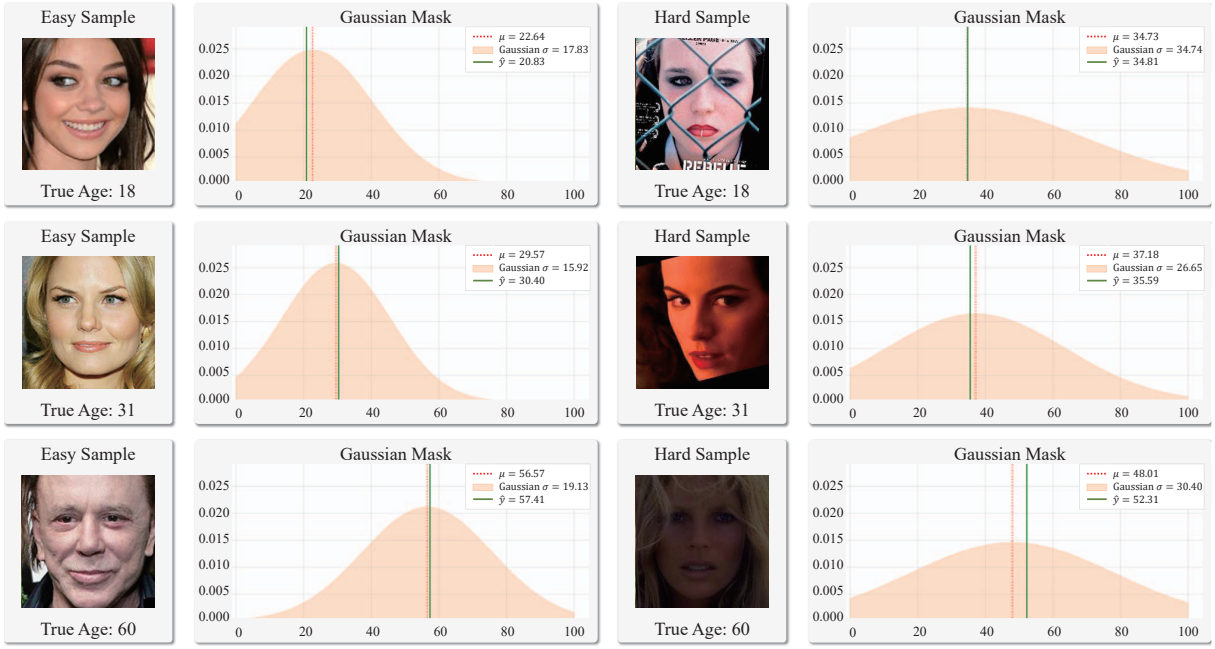

Figure 8. Qualitative examples from CACD. Examples with similar ground-truth ages but different prediction uncertainties (σ). Each shows the input face, ground-truth age y, predicted age $\hat{y}$, and predicted mean μ and standard deviation σ. Smaller σ values indicate confident, easy cases, whereas larger σ reflect greater uncertainty and prediction difficulty.

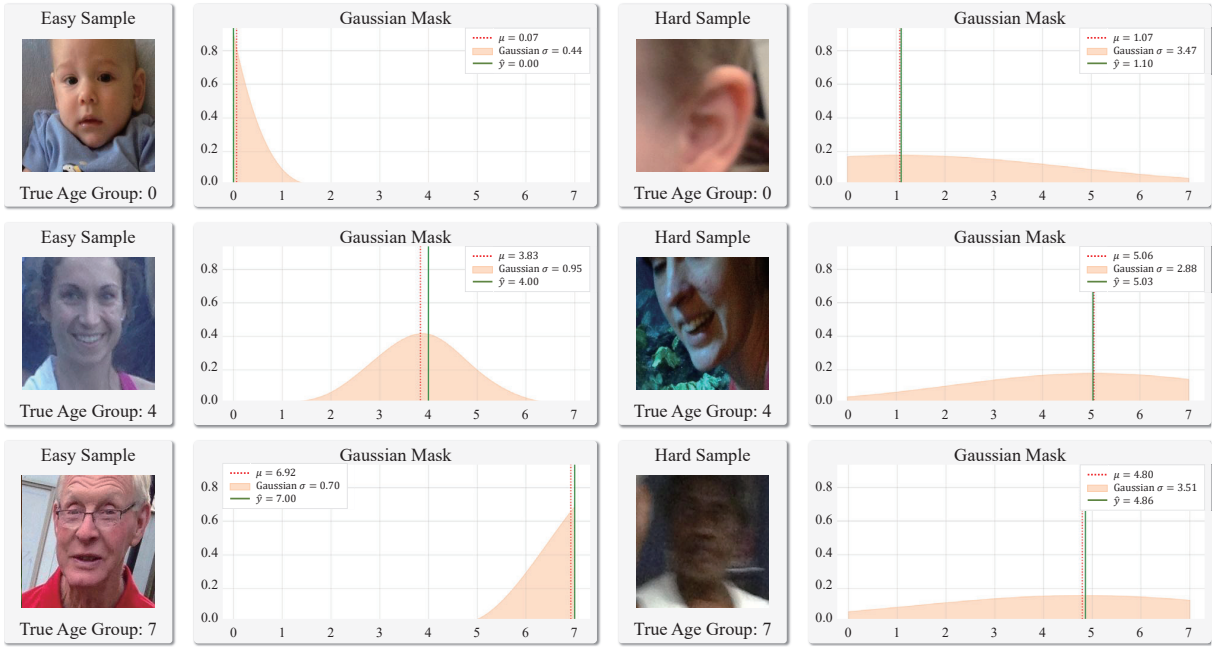

Figure 9. Qualitative examples from Adience. Examples with similar ground-truth ages but different prediction uncertainties (σ). Each shows the input face, ground-truth age y, predicted age $\hat{y}$, and predicted mean μ and standard deviation σ. Smaller σ values indicate confident, easy cases, whereas larger σ reflect greater uncertainty and prediction difficulty.

D. Broader Impacts

This work develops a method for learning age-related representations from facial images. Although it shows strong performance on age estimation tasks, its use should be approached with caution. Predicting personal attributes such as age is inherently sensitive and may raise ethical concerns, particularly when models exhibit demographic bias or are deployed without appropriate oversight. The proposed system should not be used to make decisions that directly affect individuals and should function only as a decision-support component within a controlled workflow. When applied in practice, fairness evaluations, transparent usage guidelines, and continuous monitoring are essential to avoid misuse or unintended harm. Given these considerations, we recommend using this model strictly for research purposes.

References

[1] Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baro, Isabelle Guyon, and Rasmus Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *Int. Conf. Autom. Face Gesture Recognit.*, 2017. 6

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

[3] 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 TMM*, 2015. 1

[4] 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. 3

[5] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, 2019. 1

[6] Yao Du, Qiang Zhai, Weihang Dai, and Xiaomeng Li. Teach CLIP to develop a number sense for ordinal regression. In *ECCV*, 2024. 3, 8

[7] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, 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 *ICCVW*, 2015. 1

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

[9] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. In *ICLR*, 2017. 1

[10] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE TPAMI*, 2002. 1

[11] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *ICLR*, 2020. 2, 3

[12] Seon-Ho Lee, Nyeong Ho Shin, and Chang-Su Kim. Geometric order learning for rank estimation. In *NeurIPS*, 2022. 3

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

[14] 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. 3

[15] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *CVPR*, 2021. 1, 3

[16] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. OrdinalCLIP: Learning rank prompts for language-guided ordinal regression. In *NeurIPS*, 2022. 3

[17] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2019. 1, 2, 3

[18] Xin Liu, Shaoxin Li, Meina Kan, Jie Zhang, Shuzhe Wu, Wenxian Liu, Hu Han, Shiguang Shan, and Xilin Chen. AgeNet: Deeply learned regressor and classifier for robust apparent age estimation. In *ICCVW*, 2015. 3

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

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

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[22] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: the first manually collected, in-the-wild age database. In *CVPRW*, 2017. 1

[23] Jakub Ondrejka. *Uncertainty Disentanglement in Face Age Estimation*. PhD thesis, 2024. 6

[24] Dileepa Pitawela, Gustavo Carneiro, and Hsiang-Ting Chen. CLOC: Contrastive learning for ordinal classification with multi-margin n-pair loss. In *CVPR*, 2025. 3

[25] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE TCSVT*, 2024. 8

[26] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *ECCV*, 2024. 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[28] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Int. Conf. Autom. Face Gesture Recognit.*, 2006. 1

[29] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2018. 3

[30] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *CVPR*, 2022. 1, 3

[31] Yuntao Shou, Xiangyong Cao, Huan Liu, and Deyu Meng. Masked contrastive graph representation learning for age estimation. *PR*, 2025. 3

[32] 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 TPAMI*, 2017. 3

[33] Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. In *NeurIPS*, 2023. 3

[34] 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. 2, 3

[35] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *ICML*, 2021. 1

[36] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: learning continuous representations for regression. In *NeurIPS*, 2023. 1, 3, 5

[37] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. C3AE: Exploring the limits of compact model for age estimation. In *CVPR*, 2019. 3

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

[39] Yu Zhu, Yan Li, Guowang Mu, and Guodong Guo. A study on apparent age estimation. In *ICCVW*, 2015. 3