Appendix

This appendix is organized as follows:

- Section A.1 provides the method implementation details for partition learning with continuous partition matrix (Section 4.1) and different forms of classification loss $\mathcal{L}_{cls}$ (Section 4.2).
- Section A.2 presents the experiment details of training settings and the estimation of demographic attribute labels with pretrained models (Section 5.1).
- Section A.3 shows additional analysis and results. Specifically, Section A.3.1 presents experiments with different $\mathcal{L}_{cls}$ and demographic attributes. Section A.3.2 discusses ablation study on the choice of $\mathcal{L}_{inv}$ and settings of partition set $\mathcal{P}$ during feature learning, as well as the effect of loss weight λ during partition learning (Section 5.4). Section A.3.3 conveys analysis of causal effect learning and provides additional visualization results (Section 5.5).

A.1. Method Implementation in Section 4

This section begins by presenting the implementation details of continuous partition matrix optimization in partition learning (Section 4.1), followed by the discussion on various forms of classification loss $\mathcal{L}_{cls}$ in feature learning (Section 4.2).

A.1.1. Partition Learning in Section 4.1

During partition learning, we maximize the objective $\mathcal{L}_{inv}$ to obtain the partition matrix $\mathbf{P}$ and the confounding demographic attribute. In practice, to enable back-propagation in optimization, we adopt a continuous partition matrix, *i.e.*, soft partition denoted as $\tilde{\mathbf{P}} \in \mathbb{R}^{C\times K}$. Specifically, $\tilde{\mathbf{P}}$ is initialized with random values and then updated using Eq. (4) in Section 4.1. After that, we threshold the learned soft partition matrix $\tilde{\mathbf{P}}$ to obtain the hard partition matrix $\mathbf{P} \in \{0,1\}^{C\times K}$. Denote the conventional Softmax function as $\mathcal{N}$, we apply Softmax normalization on $\tilde{\mathbf{P}}$ such that:

$$\mathcal{N}\left(\tilde{\mathbf{P}}\right)_{i,k} \in [0,1], \text{ and } \sum_{k=1}^{K} \mathcal{N}\left(\tilde{\mathbf{P}}\right)_{i,k} = 1, \forall i \in \{1...C\}. \tag{A1}$$

Here $\mathcal{N}\left(\tilde{\mathbf{P}}\right)_{i,k}$ represents the confidence value of the i-th identity belonging to the k-th subset. Therefore, based on the confidence values of each subset, we re-weigh the computation of supervised contrastive loss [5] $\mathcal{L}_{cls}$ in $\mathcal{L}_{inv}$ of Eq. (4), as formulated below:

$$\mathcal{L}_{cls}(\Phi, \cdot, \tilde{\mathbf{P}}, k) = \sum_{\mathbf{x}\in\mathcal{X}} \mathcal{N}(\tilde{\mathbf{P}})_{y,k} \cdot \sum_{\mathbf{x}^+\in\mathcal{X}} -\log \frac{e^{\Phi(\mathbf{x})^\intercal \Phi(\mathbf{x}^+)}}{\sum_{\mathbf{x}^*\neq\mathbf{x}}^{\mathbf{x}^*\in\mathcal{X}} e^{\Phi(\mathbf{x})^\intercal \Phi(\mathbf{x}^*)}}, \tag{A2}$$

where $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ represents the set of all the N training images, $\mathbf{x}^+$ is the corresponding positive image sharing the same identity with $\mathbf{x}$ in $\mathcal{X}$. Thus we can enable back-propagation to optimize the partition matrix. After learning soft partition $\tilde{\mathbf{P}}$, the hard partition matrix $\mathbf{P}$ can be obtained as below:

$$\mathbf{P}_{i,k} = \begin{cases} 0, \text{ if } \mathcal{N}\left(\tilde{\mathbf{P}}\right)_{i,k} \leqslant 0.5; \\ 1, \text{ otherwise.} \end{cases} \tag{A3}$$

A.1.2. Classification Loss $\mathcal{L}_{cls}$ in Section 4.2

For invariant feature learning, we optimize the objective $\mathcal{L}_{inv}$ in Eq. (6) where the definition of classification loss $\mathcal{L}_{cls}$ has a variety of options. As our INV-REG is orthogonal to the existing face technologies, it can be plugged into the latest classification losses without further modification. In this section, we first illustrate the mathematical forms of $\mathcal{L}_{cls}$ for Arcface and CIFP baselines adopted in Section 4. Furthermore, we provide additional examples of other marginal losses, including Cosface and Cosface-based CIFP variation. Experimental results using these two losses are presented in Section A.3.1.

Arcface. In margin-based losses, the objective is to optimize the decision boundary in angular space based on the L2 normalized weights and features. Denote the angle between the extracted feature $\Phi(\mathbf{x})$ and a certain weight vector of class y as θ, the corresponding angle is computed as $\theta_y = \arccos(\Phi(\mathbf{x}), f_y)$. The $\mathcal{L}_{cls}$ of Arcface loss [2] is then formulated as:

$$\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k) = -\sum_{\mathbf{x}\in\mathcal{X}_k} \log \frac{e^{s\cdot cos(\theta_{y_i}+m)}}{e^{s\cdot cos(\theta_{y_i}+m)} + \sum_{j=1, j\neq y_i}^{C} e^{s\cdot cos\theta_j}}, \tag{A4}$$

where s is the scale factor, and m is the addictive margin in angular space. The decision boundary between the current class y_i and its negative classes $j (j \neq y_i)$ is decided by the margin m, where $m = 0$ degenerates to the conventional Softmax cross-entropy loss, and $m > 0$ guarantees the intra-class compactness and inter-class discriminability with minimum m margin in feature space. In practice, m is a predefined fixed value.

CIFP. Unlike Arcface which applies fixed margin to all the samples, CIFP [9] assigns an adaptive margin to each instance based on the false positive rate (FPR) indicator, where hard samples (*i.e.*, misclassified in training) have larger margin values. CIFP can be regarded as margin-based sample re-weighting to promote recognition consistency across different groups. Depending on the utilized marginal losses, CIFP can have different variations. In Section 4, we adopted CIFP based on Arcface margin, and the

corresponding $\mathcal{L}_{cls}$ is expressed as follows:

$$\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k) = -\sum_{\mathbf{x}\in\mathcal{X}_k} \log \frac{e^{s\cdot cos(\theta_{y_i}+m)}}{e^{s\cdot cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{C} e^{s\cdot cos\left(\theta_j + \frac{\gamma_i^+}{\gamma_u^+}\right)}}, \tag{A5}$$

where γ_i^+ and γ_u^+ are the instance and overall FPR indicators respectively. Please refer to [9] for more details.

Cosface. Both Arcface and Cosface [8] target at maximizing inter-class variance and minimizing intra-class variance. In comparison, Cosface removes the addictive margin to the outside of $cos(\theta)$ and applies it on logit level as derived below.

$$\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k) = -\sum_{\mathbf{x}\in\mathcal{X}_k} \log \frac{e^{s\cdot(cos\theta_{y_i}+m)}}{e^{s\cdot(cos\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{C} e^{s\cdot cos\theta_j}}. \tag{A6}$$

CIFP-Cosface. For the Cosface variation of CIFP, denoted as CIFP-Cosface, both the fixed margin and the adaptive FPR-based margin are applied on logit level. The formula is given below.

$$\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k) = -\sum_{\mathbf{x}\in\mathcal{X}_k} \log \frac{e^{s\cdot(cos\theta_{y_i}+m)}}{e^{s\cdot(cos\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{C} e^{s\cdot\left(cos\theta_j + \frac{\gamma_i^+}{\gamma_u^+}\right)}}. \tag{A7}$$

A.2. Experiment Details

This section discusses the details of training settings and the estimation of demographic attribute labels (Section 5.1).

A.2.1. Training Settings

In the experiments, we employed SGD optimizer with momentum of 0.9 and weight decay of 5e-4 to train the models. The learning rate started at 0.1 and was reduced by a factor of 10 after 110K, 190K, and 220K iterations. The total number of training epochs was 21 for all the experiments. We adopted the modified ResNet architecture [4, 2] as the backbone, where Batch Normalisation (BN), Dropout, fully connected layer, and BN were appended to the last convolutional layer as the output. The embedding size was 512. For partition learning using Eq. A2, we referred to [5] and appended a multi-layer perceptron to the backbone, which consists of a single hidden layer of size 512 and output vector of size 128. For the training losses in feature learning, we configured the scale as 64 and the margin as 0.4 for Arcface and CIFP, and 0.35 for Cosface and CIFP-Cosface. We adhered to the original work for the other settings in CIFP.

Attributes		#Samples	#Identities
Race	Caucasian	3933K	58K
	African	578K	9K
	South Asian	537K	8K
	East Asian	771K	11K
Gender	Female	2034K	25K
	Male	3788K	61K

Table A1: Demographic distribution statistics of train set.

A.2.2. Demographic Attributes Estimation

To analyze the demographic statistics of the training data in Section 5.1, as well as investigate the effect of ground-truth partitions in Section 5.4, we employed pretrained attribute classification models to estimate the demographic attributes. In particular, the annotations of race (*i.e.*, Caucasian, African, South Asian, and East Asian) and gender (*i.e.*, female and male) for the MS-Celeb-1M dataset [2, 3] were predicted. A majority voting strategy was employed in post-cleaning to ensure that images of the same identity have consistent demographic attributes. Statistics of the cleaned attributes are shown in Table A1. We observe that the dataset is heavily skewed towards Caucasian race and male gender with more samples and identities than the counterparts, leading to a degradation of recognition fairness in conventional training.

A.3. Additional Analysis and Results

This section first presents experiments involving different $\mathcal{L}_{cls}$ and demographic attributes. It then discusses the ablation study on the choice of $\mathcal{L}_{inv}$, the setting of partition set $\mathcal{P}$, and the loss weight λ during partition learning. Finally, the analysis of causal effect learning and additional visualization results are provided.

A.3.1. Additional Experiments

Results on Different $\mathcal{L}_{cls}$. Referring to the definitions in A.1.2, we plugged our INV-REG into Cosface and CIFP-Cosface baselines. The models were trained with ResNet-50 backbone, and evaluated on MFR dataset. We observe in Table A2 that, our method demonstrates superiority over both baselines with improved average accuracy and reduced standard deviation. Moreover, the performance of underrepresented minority racial groups is greatly enhanced without sacrificing the majority accuracy. This validates that our method is orthogonal to diverse loss functions and brings further performance benefits.

Results on Children and Masked Faces. In addition to the multi-race attributes, we conducted evaluation on MFR dataset for children and masked face attributes. The children category contains 14K identities and 157K images,

Method	African (AF)	Caucasian (CA)	South Asian (SA)	East Asian (EA)	Avg	Std	All
Cosface* [8]	74.50	84.81	82.55	53.84	73.92	12.21	78.67
Ours-Cosface	76.11	84.38	82.69	55.17	74.59	**11.63**	79.29
CIFP-Cosface* [9]	77.15	85.53	83.83	54.50	75.26	12.38	79.68
Ours-CIFP-Cosface	**78.24**	**86.70**	**84.78**	**55.66**	**76.35**	12.34	**80.25**

Table A2: Verification accuracy (%) on MFR dataset. ("*": self-implemented results based on the officially released code. "Ours-": our results achieved by plugging our INV-REG into other baselines. "Avg"/"Std": average/standard deviation of the accuracy on four races. "All": accuracy on all the samples.)

Method	Children	Masked
Arcface [2]	55.11	59.91
Ours-Arcface	55.65	63.10
CIFP [9]	59.21	65.03
Ours-CIFP	**59.24**	**65.87**
Cosface [8]	54.50	64.14
Ours-Cosface	55.19	65.02
CIFP-Cosface [9]	58.28	64.33
Ours-CIFP-Cosface	**61.76**	**65.09**

Table A3: Verification accuracy (%) on MFR dataset.

Method	AF	CA	SA	EA	Avg	Std
CIFP	77.26	85.52	83.76	55.74	75.57	11.86
Ours†	79.40	**86.73**	84.82	57.11	77.01	11.80
Ours	**79.41**	86.53	**84.99**	**57.82**	**77.19**	**11.49**

Table A4: Verification accuracy (%) on MFR dataset with different $\mathcal{L}_{inv}$ losses. (Ours†: REx in Eq (3). Ours: IRMv1 in Eq. (2).)

while the masked faces category contains 6.9K identities with 6.9K masked images and 13K non-masked images. For the evaluation metric, true accept rate (TAR) with false accept rate (FAR) 1e-4 is obtained as in Table A3. Our INV-REG method outperforms the baselines on both attributes regardless of the marginal losses, demonstrating its generalization ability under diverse demographic attributes.

A.3.2. Additional Ablation Study

Choice of $\mathcal{L}_{inv}$ in Eq. (6). Section 3.2 discussed different implementations of $\mathcal{L}_{inv}$, including IRMv1 [1] in Eq. (2) and REx [6] in Eq. (3). To investigate the choice of $\mathcal{L}_{inv}$, we performed feature learning with both implementations. Table A4 lists the evaluation results on MFR dataset with CIFP baseline and ResNet-50 backbone. We observe that, both implementations of $\mathcal{L}_{inv}$ greatly improve the multi-race performance with higher average accuracy and lower standard deviation compared with CIFP. This validates the effectiveness of invariant learning with IRM theory [1]. In addition, $\mathcal{L}_{inv}$ with IRMv1 achieves the top performance

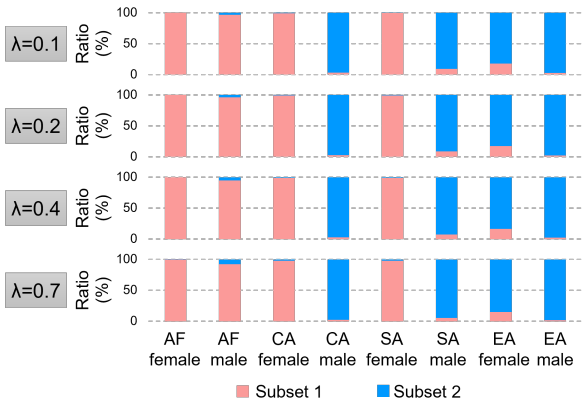

Figure A1: Proportion of the demographic groups in the partition subsets with different λ.

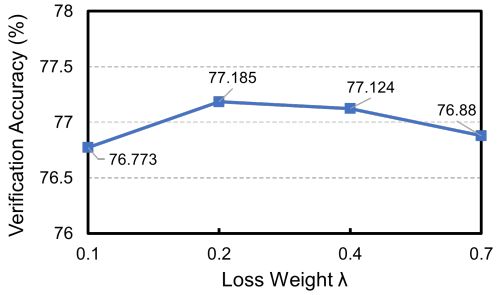

Figure A2: Accuracy (%) on MFR dataset with different λ.

on most attributes, especially on minority races (*e.g.*, improving East Asian from 55.74 to 57.82). Thus we adopted IRMv1 in our experiments.

Loss Weight λ in Eq. (4). Section 5.4 presented the ablation of loss weight λ in $\mathcal{L}_{inv}$ for feature learning. Here we further study the effect of λ for partition learning. We increased λ from 0.1 to 0.7 and analyzed the resulted partition subsets and verification accuracy on MFR. Figure A1 shows that the distribution of demographic attributes among partition subsets remains relatively consistent across various λ. For example, subset 1) is dominated by African people and Caucasian and South Asian females in all the settings. Furthermore, Figure A2 depicts the average multi-race ac-

Method	AF	CA	SA	EA	Avg	Std
Arcface	74.54	84.43	81.47	53.27	73.43	12.18
Ours†	76.24	84.71	82.19	54.42	74.39	**11.93**
Ours	**77.00**	**85.30**	**82.88**	**54.93**	**75.03**	11.99

Table A5: Verification accuracy (%) on MFR dataset with different settings of partition set $\mathcal{P}$. (Ours†: keeping only the most recent partition. Ours: maintaining all previously learned partitions.)

Method	Avg	Std	All
Arcface	73.19 ± 0.18	12.49 ± 0.27	77.68 ± 0.22
Ours-Arcface	$\mathbf{75.07 \pm 0.09}$	$\mathbf{11.93 \pm 0.08}$	$\mathbf{79.28 \pm 0.26}$

Table A6: Mean and variation of main results on MFR dataset computed from 5 independent runs.

Method	Avg (21)	Std (21)	Avg (23)	Std (23)
CIFP baseline	75.57	11.86	75.67	11.88
Ours (1 partition)	76.75	12.20	76.74	12.32
Ours (2 partitions)	76.89	11.85	76.90	11.87
Ours (3 partitions)	**77.19**	**11.49**	77.30	11.48
Ours (4 partitions)	76.82	11.90	**77.36**	**11.17**

Table A7: Performance with longer training (# of epochs) on MFR dataset.

curacy using different λ, where our method demonstrates robustness by maintaining stable accuracy levels. The best performing $\lambda = 0.2$ is adopted.

Partition set $\mathcal{P}$. Table A5 lists the performance comparison using different settings in partition set $\mathcal{P}$, *i.e.*, retaining all previously discovered partitions as described in Section 5, or keeping only the most recent partition. We observe that both settings result in performance improvement over the Arcface baseline with ResNet-50 backbone, yet maintaining all previously discovered partitions yields superior accuracy. We postulate that the model trained solely on the most recent partition may again exhibit bias towards previously deconfounded demographic attributes in earlier partitions. This is due to the model's tendency to overfit to the confounding demographic-specific features. On the other hand, by maintaining all the learned partitions in $\mathcal{P}$, diverse demographic biases will be mitigated progressively in the trained models, improving the robustness across all demographic groups.

Sensitivity to randomness. In the experiments, we utilized Std to evaluate model fairness on different demographic groups. To address the variability of the Std metric due to randomness, we performed sensitivity study on MFR dataset with 5 random seeds in Table A6, and the low variability validates the significance of our results.

Model Convergence. In Table A7, we extended Table 9

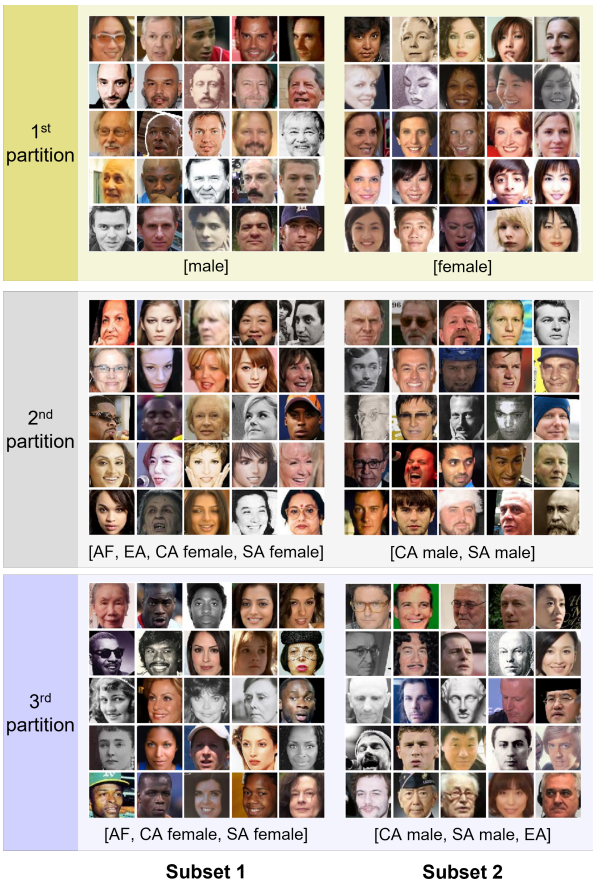

Figure A3: Visualization results of face images in the subset of different partitions. Demographic attribute statistics of each subset (same as Figure 7) are given.

results by training CIFP and our method with more epochs. CIFP hardly benefits from more training, validating that it converges at 21 epoch. In contrast, our method with more partitions brings further improvement given more epochs, validating that using more partitions requires larger number of training epochs to fully converge.

A.3.3. Additional Analysis

Learning Causal Effect. The causal effect from the image X to prediction Y is given by $P(Y|do(X))$ [7]. In Figure 3, $P(Y|do(X))$ can be expanded using the backdoor adjustment formula:

$$P(Y|do(X)) = \sum_{d} P(Y|X,d)P(D=d), \qquad \text{(A8)}$$

which corresponds to learning $P(Y|X,d)$ within each confounder stratum (*i.e.*, demographic group), and combining them with the fixed population-level prior $P(D)$. In contrast, conventional learning using the full training data cor-

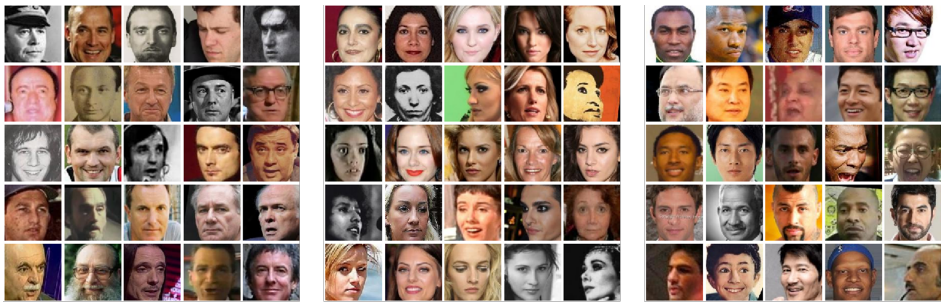

Figure A4: Visualization of face images in partition of three subsets.

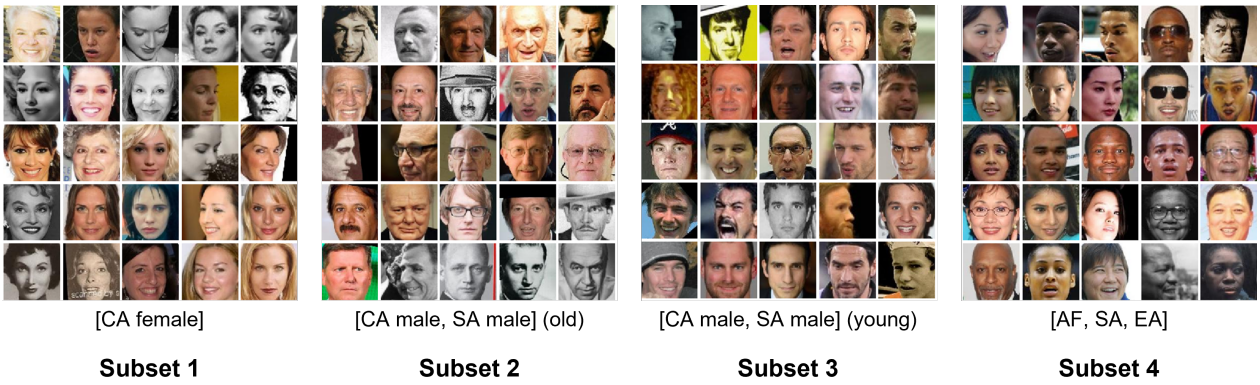

Figure A5: Visualization of face images in partition of four subsets.

responds to maximizing the likelihood $P(Y|X)$, given by:

$$P(Y|X) = \sum_{d} P(Y|X,d)P(D=d|X), \tag{A9}$$

where the key difference is that the weight of each $P(Y|X,d)$ changes from the population-specific $P(D=d)$ to the sample-specific $P(D=d|X)$. However, the appearance of X is affected by D, $P(D=d|X)$ tends to have an extremely large value on the demographic group $D=d$ that X belongs to.

Hence when group d contains spurious demographic-specific context, the training can be dominated by $P(Y|X,d)$, such that the model recklessly captures the context for prediction. Note that the key to capture the causal effect $P(Y|do(X))$ is *not* to learn a representation independent from the demographic attributes, but to fairly adjust the contribution of each $P(Y|X,d)$ by $P(D=d)$, where d indeed participates in the prediction by $P(Y|X,d)$, *i.e.*, the demographic attribute contains valid cue to differentiate identities.

Visualization of Partition Subsets. In Figure 7, we presented the demographic distribution of different partition subsets. Here we further visualized the face images in each subset as shown in Figure A3. There exhibits a distinct contrast between the two subsets in each partition. For instance, in the first partition, subset 1) is primarily composed of male faces whereas subset 2) contains mostly female images. This further verifies our analysis in Section 5.5 that each partition encodes a certain demographic attribute. In addition, we extended our analysis to visualize the partition results when more than two subsets are utilized (Table 8), *i.e.*, number of subsets $K=3$ and $K=4$ as shown in Figure A4 and Figure A5, respectively. We can see those subsets of $K=3$ behave with different preferences in the demographic attributes, where subset 1) and 2) are dominated by Caucasian males and females respectively, while all other attributes are categorized into subset 3). As for $K=4$, in addition to the attributes of race and gender, age is also considered during partition learning. As an illustration, both subset 2) and 3) consist mostly males of Caucasian and South Asian, yet the faces in subset 2) are older in age compared to those in subset 3).

Failure Cases Analysis. We supplement more visualization results on failure cases using RFW dataset. Due to factors such as hairstyle, wearing hat, and head poses, Arcface baseline presents low similarity scores for faces from the same identity as shown in Figure A6a, yet high simi-

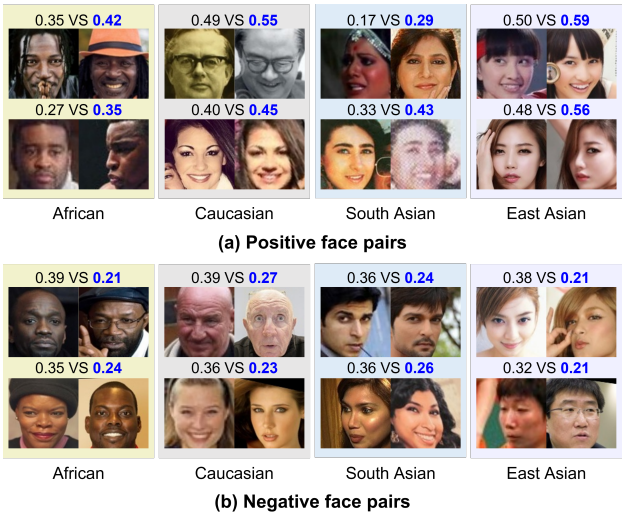

Figure A6: Failure cases analysis on different races. (a) Positive face pairs from the same identity. (b) Negative face pairs from different identities. Similarity score of each pair is given to compare Arcface versus our method (in bold blue).

larity values for the negative face pairs of different identities in Figure A6b. In comparison, our INV-REG effectively captures the causal feature that is invariant of the spurious demographic-specific attributes, resulting in improved positive similarity and decreased negative similarity scores.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 2, 3

[3] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 1, 2

[6] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 3

[7] Judea Pearl. *Causality*. Cambridge university press, 2009. 4

[8] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2, 3

[9] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 578–586, 2021. 1, 2, 3