Diverse Visual Experience Promotes Integrated Representations and Mitigates Bias in Deep Neural Networks for Face Perception

Elaheh Akbari[1,2]
Katharina Dobs[2,3,4]

1. Department of Psychology, Justus Liebig University Giessen
2. Center for Mind, Brain and Behavior, Universities of Marburg, Giessen, and Darmstadt
3. Department of Computer Science, Justus Liebig University Giessen
4. Center for Applied Computer Science and Data Science, Justus Liebig University Giessen

Abstract

Humans are experts at recognizing faces, yet this expertise is not uniform: people perceive faces from familiar facial groups more accurately than those from unfamiliar ones—a phenomenon known as the Other-Race Effect (ORE). Although the ORE is well established, it remains unclear how exposure diversity reshapes the underlying representations that give rise to it. Here, we combine deep convolutional neural networks (CNNs) with human behavioral data to investigate how training diversity influences the emergence and mitigation of recognition biases. CNNs trained exclusively on either Asian or White faces reproduced ORE-like biases, showing reduced performance for faces from unfamiliar groups. In contrast, a CNN trained on both Asian and White faces showed minimal bias, balanced performance across groups, and an integrated representational geometry spanning both groups. Critically, this model best captured human face-matching behavior across both Asian and White participants, outperforming single-trained networks on cross-group trials. These findings show how diverse visual experience reorganizes representational geometry to support cross-group generalization in artificial systems and provide computational insight into how diverse experience may shape more integrated face representations in humans.

Keywords: other-race effect, face perception, convolutional neural networks, visual experience, representational geometry, perceptual bias

Introduction

Humans are remarkably proficient at visual recognition, yet our perception consistently reflects systematic biases. These biases shape how we interpret sensory information, ranging from low-level visual illusions, such as perceiving motion in static images or misjudging size and color based on context, to higher-level biases like better recognition of one's own social group. One prominent example is the Other-Race Effect (ORE), the difficulty humans experience when recognizing faces from populations less familiar to them. The ORE is robust and extensively documented across cultures and contexts (Meissner & Brigham, 2001; Platz & Hosch, 1988; Rossion & Michel, 2012), in experimental paradigms such as memory and perceptual discrimination (Turk et al., 2005; Walker & Hewstone, 2006; Walker & Tanaka, 2003; Wong et al., 2021; Zhao et al., 2014), as well as in neuroimaging studies (Ficco et al., 2023; Golby et al., 2001).

Developmental research suggests that the ORE emerges from asymmetrical visual experience: although absent at birth (Kelly et al., 2005), own-group preferences arise by 3 months and persist into adulthood (Bar-Haim et al., 2006; Kelly, Liu, et al., 2007; Kelly, Quinn, et al., 2007). Early exposure to diverse faces can significantly mitigate the ORE (Anzures et al., 2012, 2013; Bar-Haim et al., 2006; McKone et al., 2019; Spangler et al., 2013; Suhrke et al., 2014; Wan et al., 2015), a pattern supported by behavioral (Brigham & Malpass, 1985; Hancock & Rhodes, 2008), computational (O'Toole et al., 1991), and neuroimaging evidence (Walker et al., 2008). Importantly, recent findings show that not only which faces one encounters, but also the richness of face experience shape recognition abilities: individuals from smaller or less socially diverse communities show reduced face recognition performance (Balas & Saville, 2017). Together, these results suggest that the diversity of one's visual environment broadly constrains the development of robust face representations.

In parallel, alternative accounts emphasize socio-cognitive influences on the ORE, such as attentional biases or reduced motivation to individuate unfamiliar faces (Bernstein et al., 2007; Shriver et al., 2008). Thus, while experience shapes behavioral outcomes in face recognition, the underlying mechanisms remain unclear. In particular, it is unclear whether limited experience restricts the richness or separability of the representational space needed to individuate unfamiliar faces, and whether diverse exposure leads to segregated group-specific representations or a more integrated representational space that supports cross-group generalization? Addressing these questions in humans is challenging due to limited control over individual perceptual histories and the difficulty of disentangling perceptual from social factors.

Artificial neural networks—lacking social cognition, and motivation—provide a unique opportunity to isolate the role of visual experience in shaping face perception. Deep convolutional neural networks (CNNs), in particular, have proven valuable models of biological vision: when trained on object recognition, they exhibit strong similarities to the primate visual system in both performance and

representational geometry (Jacob et al., 2021; Kriegeskorte, 2015; Yamins & DiCarlo, 2016). Critically, CNNs also reproduce hallmark signatures of human face perception, including the inversion effect (Dobs et al., 2023; Yovel et al., 2022), Mooney face perception (Zeman et al., 2022), and ORE itself (Dobs et al., 2023; Tian et al., 2021). Manipulating visual experience in CNNs further reveals that exposure to faces is necessary for view-invariant representations (Abudarham et al., 2021), and that the amount and kind of facial experience shape view invariance and sensitivity to critical face features, resembling human perceptual sensitivity (Rosemblaum et al., 2025). Taken together, CNNs offer a unique methodological advantage: they allow full control over the training 'diet', enabling systematic tests of how visual experience shapes representational spaces and their properties (Doerig et al., 2023; Kanwisher et al., 2023; O'Toole & Castillo, 2021; Van Dyck & Gruber, 2023).

Here, we leverage CNNs to investigate the origins of the ORE and the role of exposure diversity in shaping face representations. Specifically, we asked (1) whether diverse visual exposure mitigates ORE-like biases in CNNs, (2) which representational mechanisms support or prevent such mitigation, and (3) whether the resulting latent representations predict human behavioral choices across groups. To address these questions, we trained three CNNs: one only on Asian faces, one only on White faces, and one on both Asian and White faces. We first tested whether single-trained CNNs show an ORE, and whether training diversity reduces this bias. We then examined how these differences are reflected in the networks' internal representational geometry and how well each model accounts for human behavioral data. We hypothesized that single-trained CNNs would show group-specific feature spaces, mirroring behavioral biases observed in humans with limited exposure to other-group faces. In contrast, diverse training would yield more integrated representations that reduce such biases. By combining computational modeling with human behaviour, this work provides a computational account of how visual experience shapes face representations and contributes to perceptual bias, and offers insights into how diverse experience may support more generalized face perception in humans.

Methods

Training Deep Convolutional Neural Networks

To assess the impact of training diversity on face representations and recognition performance, we trained three CNNs, all based on the VGG16 architecture (Simonyan & Zisserman, 2015); Figure 1A). VGG16 was chosen because it is optimized for face recognition (Parkhi et al., 2015), achieves human-level accuracy on diverse face recognition tasks (Gwyn et al., 2021; M. Wang & Deng, 2021) and

reproduces behavioral and neural signatures of human face perception (Dobs et al., 2023; Grossman et al., 2019).

Each network was randomly initialized and trained for identity recognition using datasets that differed in facial composition (Figure 1B): First, a Single Asian CNN was trained exclusively on 1,654 Asian identities (891 female) from a subset of the Asian Face Dataset (Xiong et al., 2018). Second, we trained a Single White CNN exclusively on 1,654 White identities (827 female) drawn from subsets of the VGGFace2 (Cao et al., 2018) and CASIA-WebFace (Yi et al., 2014) datasets. For both the single-trained CNNs, 94 training images and 5 validation images per identity were used. The number of identities and images per identity was matched across groups and was constrained by the available size of the Asian dataset, ensuring that differences between models reflect training diversity rather than differences in dataset size. Third, we trained a Dual CNN on the union of both groups (3,308 identities total). To equate the total number of training samples across all models (155,476 total training images), we used 47 training images per identity (validation images remained 5 per identity).

All three CNNs were trained using identical hyperparameters: Training was conducted with a batch size of 128, using a cross-entropy loss function and the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a learning rate of 0.001. To prevent overfitting, a weight decay of 0.0001 was utilized. A learning rate scheduler (ReduceLROnPlateau, Pytorch) reduced the learning rate by a factor of 0.1 if the validation loss failed to decrease five consecutive epochs ('patience' = 5). This helped fine-tune learning when training reached a plateau.

During training, all images underwent a series of data augmentations. Each image was resized to 256 x 256 pixels and then randomly cropped to 224 x 224 pixels. A random grayscale transformation was applied with a probability of 0.2, ensuring that the networks learned features beyond color cues. Additionally, 50% of the images were randomly flipped horizontally to increase variability. Finally, all images were normalized to a mean of 0.5 and a standard deviation of 0.5 for each RGB channel. For validation and testing, images were resized to 256 x 256 pixels, center-cropped to 224 x 224 pixels, and normalized using the same parameters.

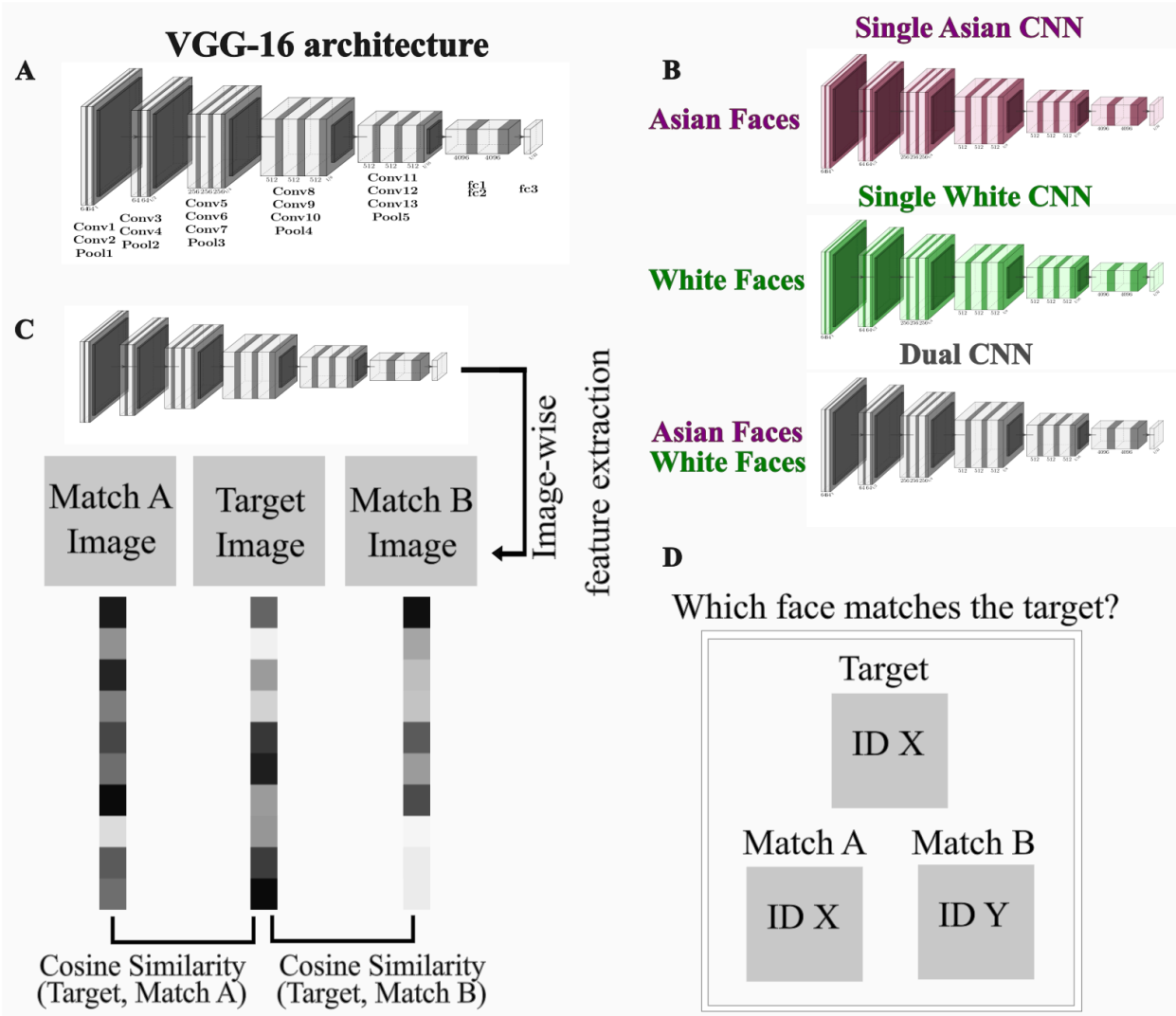

Figure 1. ***Overview of the methods.*** *A) VGG16 architecture consisting of 13 convolutional, 5 pooling, and 3 fully connected layers. B) Overview of trained CNNs: a Single Asian CNN trained only on Asian identities, a Single White CNN trained only on White identities, and a Dual CNN trained on both Asian and White identities. C) Target-matching task used in CNNs. Image-wise feature vectors were extracted from the penultimate fully-connected layer of each CNN, and cosine similarity was computed between the target image and each match (Match A and Match B). This cosine similarity is subsequently used to compute accuracy and choice probability (see Methods for details). D) Human target-matching task, in which participants chose whether Match A or Match B depicted the same identity as the target face. Images displayed here are not part of the stimulus set. Images are publicly available at https://commons.wikimedia.org.*

Other-race Effect in CNNs

To assess whether the ORE was present in the CNNs, we evaluated each CNN on a target-matching task using novel faces not included during training (Figure 1C). The test set consisted of young (20-35 years), female White and Asian identities that were not included in test or training sets. Specifically, we used a

dataset of 40 Asian and 40 White identities, each with 5 images (200 images per group; 400 images total) (Dobs et al., 2023). The images were collected from colleagues or public Instagram accounts (<2,000 followers). For each group, we generated 1,560 triplets (40 × 39), consisting of two images of a target identity and one distractor image from a different identity. Performance was quantified based on the cosine similarity between image feature vectors from the penultimate fully connected layer (fc7) of each CNN. We chose this layer because it captures the model's learned identity-relevant features without being tied to the specific output units (O'Toole et al., 2018). For each triplet, the image with the higher cosine similarity to the target was simulated as the model's choice (Figure 1C).

ORE Index: To quantify recognition bias across facial groups for all CNNs, we defined a difficulty-weighted ORE index based on normalized accuracies. First, we estimated a group-specific performance ceiling using the trained group accuracy of the corresponding single-trained model:

$$Ceiling_{group_A} = Acc_{Singlegroup_A}(group_A)$$

For each model *m*, we then compute normalized accuracies:

$$NormAcc_m(group_A) = \frac{Acc_m(group_A)}{Ceiling_{group_A}}$$

Finally, we defined the difficulty-weighted ORE index as:

$$ORE_m = \frac{\left|NormAcc_m(group_A) - NormAcc_m(group_B)\right|}{NormAcc_m(group_A)}$$

For the Dual CNN, the ORE index was computed in both directions, allowing us to assess the symmetry of bias across groups. This normalization procedure corrects for inherent differences in group-specific task difficulty, ensuring that any observed ORE differences across models reflect true representational bias rather than differences in baseline recognizability arising from the distinct stimulus sets.

Obtaining Properties of the Representational Feature Space

To investigate the properties of the learned representational feature space across CNNs, we used three complementary approaches. First, we applied a lesioning analysis in the Dual CNN to test whether diverse training experience led to an integrated feature space or to group-specific subspaces within the network. Second, we performed Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) to compare the internal representational geometries across all CNNs. Finally, we visualized filter-preferred stimuli by iteratively optimizing random noise images to maximize the activation of filters most selective for faces. This provided qualitative insight into how training affected feature preferences in each CNN.

Lesioning Analysis in the Dual CNN

To test whether the processing of different facial groups (Asian and White faces) was segregated or integrated within the last convolutional layer (conv13) of the Dual CNN, we applied a lesioning analysis following the procedure described in Dobs et al. (2023). First, we obtained group-specific importance rankings by ablating each filter in the conv13 individually and measuring the resulting increase in loss separately for each group, using 50 batches of training images per group. We focused on conv13 because it is the final layer containing spatially localized filters, thus making it the last point at which feature selectivity can be meaningfully attributed to individual filters. Filters were then ranked independently for the Asian and White face-recognition tasks based on their respective contributions to the loss.

From this initial ranking, we performed progressive, conditional lesioning in 1.6% increments: 1) Lesion the top 1.6% of filters (for a given group) simultaneously and keep them lesioned. 2) Recompute the individual loss ranking on the remaining filters (with previously selected filters still lesioned). 3) Add the top 1.6% from this updated ranking to the lesioned set and keep them lesioned in the next iteration. We repeated steps (2)-(3) until 20% of filters were selected for each group. This procedure conditions each selection on prior lesions, reducing redundancy and capturing units that remain most informative given previously removed filters.

Finally, for each group-specific set, we lesioned the full 20% of filters and evaluated performance on the entire validation set for both groups, allowing us to test whether the Dual CNN relies on group-specific subsets (segregated subspaces) or a shared, integrated feature space. We chose a 20% lesion proportion because prior work (Dobs et al., 2023) showed that this level of ablation produces strong and interpretable task-specific effects, providing a balance between too little lesioning (insufficient to reveal functional specialization) and too much (at which point the sets of “critical” filters begin to overlap across tasks, diminishing task specificity).

Representational Similarity Analysis

To characterize the representational geometry of each CNN, we conducted Representational Similarity Analysis (RSA) using the same stimulus set of 400 images (40 identities × 5 images each; Asian and White faces) used to evaluate the ORE. Activations from the fc7 layer were extracted for all images, and pairwise dissimilarities were computed using 1 – cosine similarity, resulting in a 400 x 400 representational dissimilarity matrix (RDM) for each model.

To evaluate how training diversity shapes representational geometries, we quantified dissimilarities within facial groups (Asian–Asian, White–White) and between groups (Asian–White) for

each CNN. We additionally visualized the RDMs using Multidimensional Scaling (MDS) to obtain lower-dimensional embeddings that qualitatively reveal clustering and separation patterns.

Finally, we performed cross-model RSA by correlating RDMs across networks using the Pearson correlation coefficient to assess representational similarity between CNNs trained on different datasets. Together, these analyses allowed us to directly compare how differently trained CNNs encode facial identity and group structure in their representational spaces.

Comparing Human Participants and CNNs

To directly assess the correspondence between CNN performance and human behavior, we compared the decision patterns of the CNNs with those of Asian (n = 102) and White (n = 269) participants from a previously published study (Dobs et al., 2023). Human behavioral responses were obtained using the same target-matching task illustrated in Figure 1D, in which participants selected which of two faces matched the target identity. This allowed a direct comparison between CNNs and human decision patterns across facial groups and stimulus conditions. For detailed information on participant recruitment, experimental setup, and data collection procedures, please refer to Dobs et al. (2023).

Stimuli and Target-matching Task in Humans

The stimuli used for comparing the human and CNNs choices were identical to those used for evaluating the ORE and behavioral data were drawn from Dobs et al. (2023). Each participant completed 40 target-matching trials (20 per facial group) presented in randomized order, in which they chose which of two faces matched a target identity (Figure 1D). For detailed information on the experimental setup, please refer to Dobs et al. (2023).

To assess the reliability of participants' choices, we conducted a split-half reliability analysis across triplets. In each of 1,000 iterations, we randomly divided participants into two independent groups. For each group, we computed average triplet-choice patterns and retained those triplets that were shared between both groups to ensure valid comparisons. We then calculated Pearson correlation coefficients between the two halves and applied the Spearman-Brown prophecy correction to correct for biases introduced by split-half calculations. We used the formula $\rho' = \frac{2\rho}{1+\rho}$ where ρ is the Pearson correlation between the two halves. To express the corrected reliability on the same scale as Pearson correlation coefficients, we report the square root of ρ'. Note that because each participant viewed only a small subset of all triplets, not all triplets were shared across participants, which likely led to a slight underestimation of the true noise ceiling.

CNN-Human Similarity Analysis

We then quantified the correspondence between CNN and human decision patterns on a triplet level. For each CNN, activations for the triplets (Target, Match A and Match B) used in the human experiment were extracted from the fc7 layer. We selected fc7 because it is the penultimate layer, proximal to the network's decision output yet not directly constrained by the classification layer, thus providing a meaningful approximation of the model's representational behavior. Cosine similarities between each match and the target were calculated and the probability of choosing Match A was obtained from:

$$probability\ of\ choice\ A\ =\ \frac{Cosine\ Similarity_{(Target,\ MatchA)}}{Cosine\ Similarity_{(Target,\ MatchA)} + Cosine\ Similarity_{(Target,\ MatchB)}}$$

Next, the probabilities of correct choices for triplets were computed as averages across participants. Finally, these behavioral choice probabilities were correlated with CNNs' choice probabilities obtained from the previous step using a Pearson correlation. This correlation quantifies the level of correspondence between model and human decision patterns.

Statistical Inference

Nonparametric statistical methods were employed to avoid assumptions about data distribution. For the ORE analysis, we bootstrapped the 1,560 triplets 10,000 times. For each resample, average identity-matching accuracy was computed, resulting in a distribution of ORE values for each CNN. We defined 95% confidence intervals as the 2.5th and 97.5th percentiles of the bootstrapped distributions. Pairwise comparisons between CNNs were performed separately for each facial group (Asian and White faces), with two-tailed p-values derived from the bootstrapped distribution of accuracy differences. All p-values were adjusted for multiple comparisons (two pairwise contrasts) using the Bonferroni correction, with a significance threshold of $p < .05$.

For the CNN-human consistency analysis, we calculated Pearson correlation coefficients between CNN and human responses across triplets within each bootstrap sample. Distributions of correlation values were compared between CNNs, stratified by both stimulus group (Asian vs. White faces) and participant group (Asian vs. White participants). Error bars indicate the variability of correlation coefficients, quantified as the standard deviation across bootstrap samples. To compare the consistency of two CNNs, as in the ORE analysis, two-tailed p-values were obtained from bootstrapped differences. All p-values were adjusted for multiple comparisons (three pairwise contrasts) using the Bonferroni correction, with a significance threshold of $p < .05$.

Data and Code Availability

All analyses were conducted using Python 3.9.12, and model training was performed using Python 3.8 and PyTorch 1.13.0. All analysis scripts are available at https://github.com/VCCN-lab/ORE_in_CNNs_and_humans. Code for training the networks is available at https://github.com/martinezjulio/sdnn (Dobs et al., 2022). The trained CNN models will be made available on the Hugging Face platform (https://huggingface.co) upon publication. All stimuli used for computing ORE, RSA, and for human-model behavioral comparisons are available at OSF (https://osf.io/dbks3/) (Dobs et al., 2023).

Results

Diverse training mitigates ORE-like biases in CNNs

To investigate how training diversity influences recognition biases, we evaluated three CNNs—Single White, Single Asian, and Dual CNN—on an identity-matching task using a novel set of 40 White and 40 Asian facial identities (five images per identity; Figure 1C) that were not included in training. First, we tested whether single-trained CNNs exhibited an ORE similar to humans. Both networks performed significantly better on faces from their trained group than on faces from the untrained group, as measured by face-matching accuracy in our target-matching task (Single Asian CNN: Asian-face accuracy = .94 vs. White-face accuracy = .90; Single White CNN: Asian-face accuracy = .90 vs. White-face accuracy = .97; all $p < .0001$, Bonferroni-corrected; Figure 2A), confirming a clear ORE. In line with previous findings (Dobs et al., 2023; Tian et al., 2021), these results demonstrate that asymmetric visual experience alone is sufficient to produce group-specific recognition biases.

Next, we asked whether training diversity mitigates these biases while preserving high performance across both facial groups. The Dual CNN, trained on both facial groups, maintained high accuracy for white and Asian faces (Asian = .93, White = .96, $p < .0001$; Figure 2A). Notably, its overall accuracy across groups (mean = .95) exceeded that of either single-trained model (Single Asian: mean = .92; Single White: mean = .94), indicating that diverse experience not only balances performance but also improves it. Its performance on White faces did not differ significantly from that of the Single White CNN ($p = .06$), whereas the Single Asian CNN slightly but significantly outperformed the Dual CNN on Asian faces ($p = .03$). Crucially, this balanced performance is particularly impressive given that the Dual CNN was trained on only half as many images per identity as the single-trained CNNs, indicating that training diversity, rather than data quantity, drives its broader generalization.

To further quantify these recognition biases, we computed an ORE index for each CNN (Figure 2B), measuring the relative drop in recognition accuracy for untrained versus trained groups (higher values = stronger bias). The results mirrored the accuracy pattern observed above: single-trained CNNs showed larger ORE indices, indicating greater difficulty recognizing untrained facial groups (Single White CNN = 4.7%; Single Asian CNN = 7.1%). In contrast, the Dual CNN demonstrated minimal bias, with a baseline-weighted ORE index of ~0.6% for both groups. These results underscore the effectiveness of diverse training in mitigating representational bias. Importantly, this pattern parallels developmental findings in humans, where early exposure to diverse faces has been shown to reduce the ORE (Anzures et al., 2012, 2013; Bar-Haim et al., 2006; Brigham & Malpass, 1985; Hancock & Rhodes, 2008; McKone et al., 2019; Spangler et al., 2013; Suhrke et al., 2014).

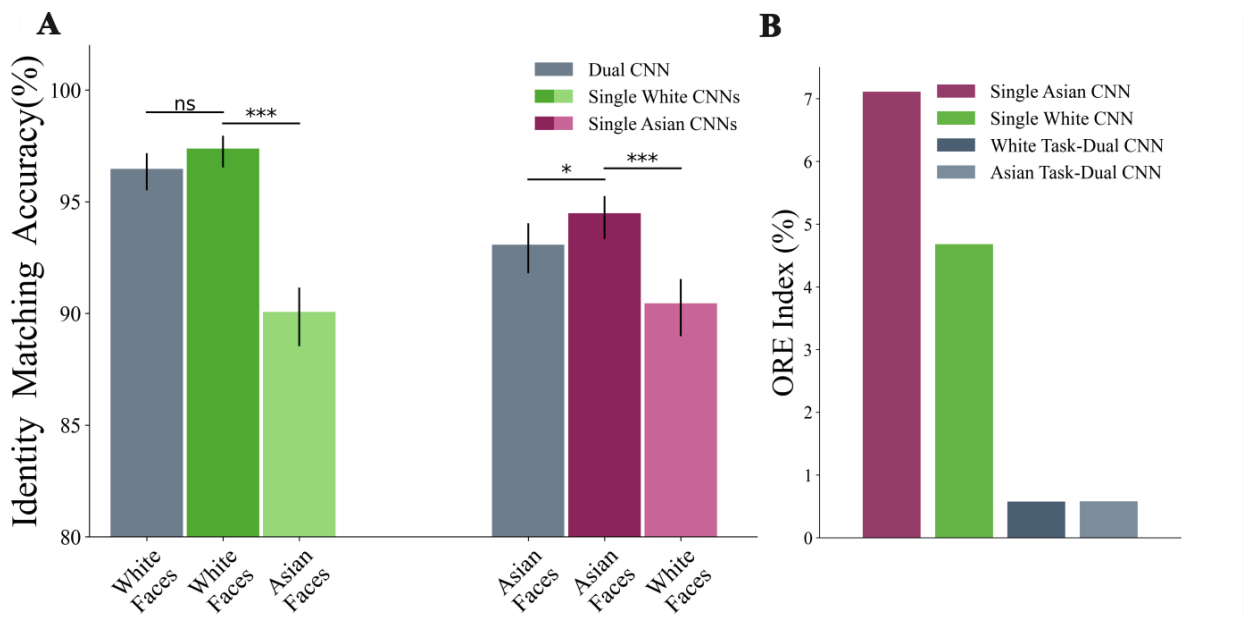

Figure 2. ***Recognition biases in CNNs.*** *A) Identity-matching accuracy for White and Asian faces across three CNNs: Single White (green), Single Asian (purple), and Dual (gray) CNN. All pairwise comparisons were statistically significant (p < .05, Bonferroni-corrected), except for the comparison between the Dual CNN and the Single White CNN on White faces (p = .06, Bonferroni-corrected). Error bars represent 95% bootstrapped confidence intervals. B) ORE index for each model. Single-trained CNNs exhibit substantial bias favoring their trained facial group, whereas the Dual CNN shows minimal ORE, indicating balanced recognition performance across facial groups. Asterisks indicate significance level as follows: ns, not significant; *, p < .05; ***, p < .0001.*

Lesioning reveals an integrated feature space in the Dual CNN

The performance results showed that training diversity substantially reduced recognition biases in the Dual CNN. We next asked how this bias reduction is reflected in the networks' internal representations. Specifically, we assessed whether the Dual CNN's reduced ORE arises from a shared, integrated

representational space that encodes faces across groups, or from the coexistence of separate group-specific subspaces supporting each facial group. To test this, we employed a lesioning analysis targeting the most critical group-specific filters in the Dual CNN's last convolutional layer (Figure 3A). Filters were ranked by the increase in loss caused by their ablation during identity recognition for either facial group. We then lesioned the top 20% of filters most critical for each group-specific task (e.g., recognition of all Asian identities) and measured the resulting change in validation performance.

Lesioning group-specific filters in the Dual CNN resulted in substantial declines in performance across both tasks (lesioning White-selected filters: White faces: -65%, Asian faces: -48%; Lesioning Asian-selected filters: White faces: -49%, Asian faces: -63%). Although impairments were somewhat larger for the facial group corresponding to the lesioned filters, both lesion types strongly disrupted performance for both groups. This pattern indicates that the Dual CNN relies on substantially shared representational resources for identity recognition across groups.

Training diversity leads to balanced representational geometry

To further probe the structure of this shared representational space and compare it with the single-trained CNNs, we examined the geometry of their feature spaces using Representational Similarity Analysis (RSA). For each model, we first constructed representational dissimilarity matrices (RDMs) by computing pairwise cosine similarities between activation patterns for all face identities within and between facial groups (Figure 3B; RDMs). In the Single White CNN, within-Asian similarity was significantly higher than within-White similarity (mean = .43 vs. .18; $p < .0001$; bootstrap test), indicating lower discriminability among Asian identities—consistent with the previously observed ORE. Similarly, in the Single Asian CNN, within-White similarity exceeded within-Asian similarity (mean = .37 vs. .26; $p < .0001$), aligning with reduced recognition of White faces (Figure 3B; mean-dissimilarity plots). This is supported by MDS visualizations, where the trained groups are dispersed but the untrained groups are compressed in each single-trained CNN, indicating poor discriminability for the untrained group compared to the trained group. On the other hand, in the Dual CNN, the similarities are evenly distributed, reflecting shared representational space with comparable discriminability for both groups (Figure 3B; MDS plots).

Interestingly, the Dual CNN displayed within-group similarity comparable to each of the corresponding single-trained models (within-Asian similarity: Dual = .29 vs. Single Asian = .26; $p < .001$; within-White similarity: Dual = .19, Single White CNN = .18, $p < .05$), reflecting a more uniform representational structure characterized by evenly distributed identity clusters. Analyses of between-group similarity revealed that the Dual CNN organized the two facial groups into more distinct but coexisting subspaces (between-group similarity = .13, compared to the Single Asian CNN = .19 and the Single White CNN (0.15); all pairwise *p*s < .0001; Bonferroni-corrected) (Figure 3B; mean-dissimilarity plots). These findings suggest a hierarchically integrated representational organization: the Dual CNN

encodes both groups within a common feature space that supports cross-group generalization, yet preserves sufficient separation to maintain high identity performance (see Figure 2A) for each facial group. Consistent with this interpretation, cross-model RSA showed that the Dual CNN's representational geometry was more similar to both single-trained networks (Single White CNN: r = .83; Single Asian CNN: .57; all pairwise *p*s < .0001; bootstrap test, Bonferroni-corrected) than the single-trained models were to each other (r = .31, $p < .0001$), with a somewhat stronger resemblance to the White-trained network (Figure 3C). This pattern further supports the view that the Dual CNN developed an integrated, but structured, representational space that bridges both facial domains.

Together, these results highlight a shared conclusion: training diversity promotes the emergence of integrated and generalizable feature representations. These representations are less constrained by group-specific biases, supporting more balanced identity encoding across facial groups and enabling improved cross-group face recognition performance in the Dual CNN.

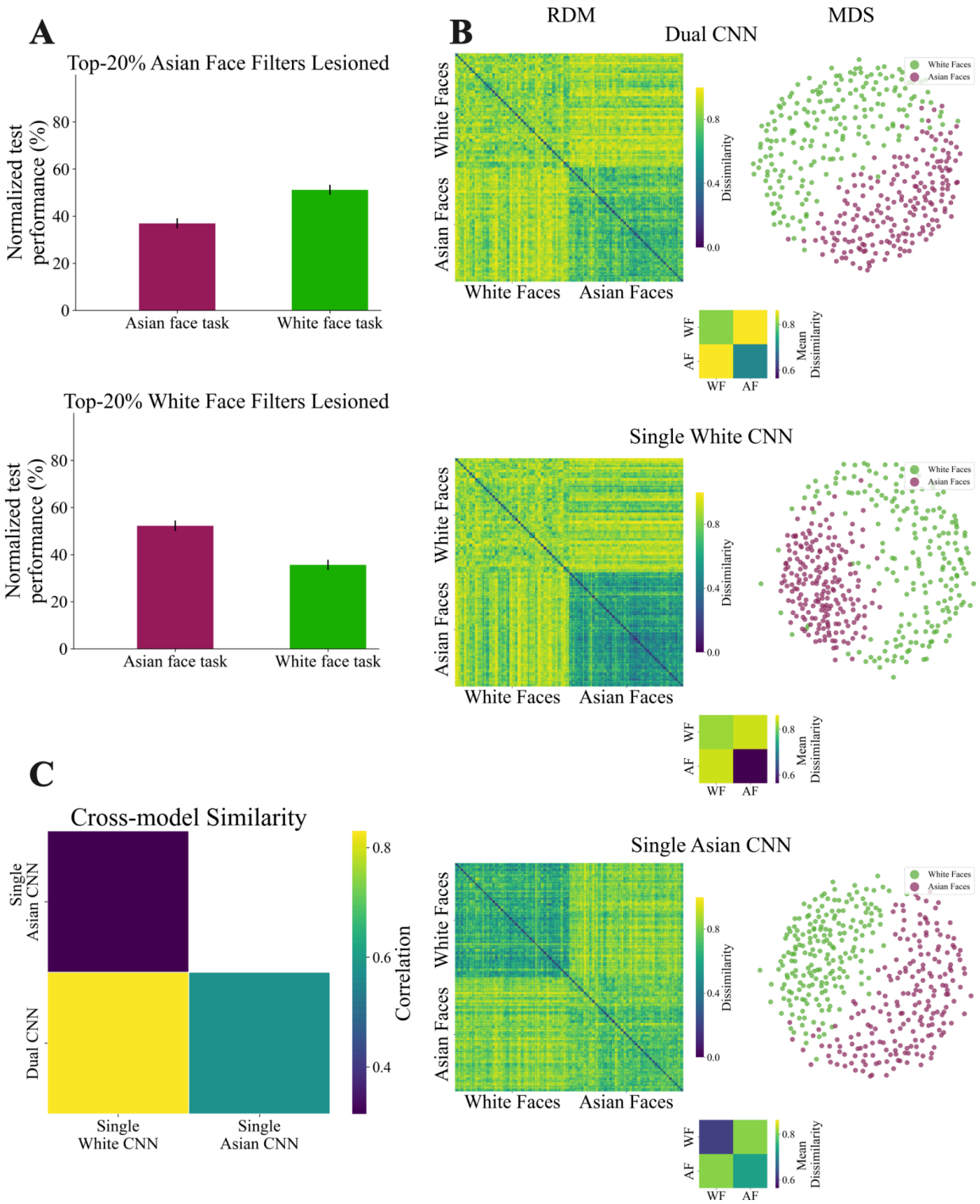

Figure 3. ***Properties of the representational feature spaces.*** *A) Lesioning analysis. The top 20% of* group-*specific filters were ablated, and performance was evaluated across both facial groups. Lesioning impaired not only the target* group *but also the non-target* group*, indicating shared representational resources within the Dual CNN. Performance is shown relative to each task's baseline accuracy. Error bars represent 95% bootstrapped confidence intervals. B) Representational Dissimilarity Matrices*

(RDMs). Pairwise dissimilarities (1 – cosine similarity) among White and Asian face identities for each CNN. To the right of each RDM, we show (i) MDS embedding of the dissimilarities (White faces: green; Asian faces: purple) and (ii) a 2×2 mean-dissimilarity plot quantifying within- and between-group similarity (AF: Asian faces; WF: White faces). Single-trained CNNs exhibit lower within-group dissimilarity for the untrained group*, indicating reduced identity discriminability. In contrast, the Dual CNN shows balanced dissimilarity patterns across both* groups*, consistent with a more integrated representational space. C) Cross-model RSA. Using Pearson correlation to compare representational geometries across CNNs.*

Dual CNN best captures human cross-group behavior

To evaluate the behavioral relevance of these representational differences, we next compared the CNNs' decision patterns to human face-matching behavior. Specifically, we computed Pearson's correlations between each model's predictions and the average human responses per triplet in the target-matching task (Figure 1D), previously collected from Asian (n = 102) and White (n = 269) participants (Dobs et al., 2023). This analysis quantified the extent to which each CNN captured human-like choice behavior at the triplet level (Figure 4). A noise ceiling was estimated via permutation of triplet labels to determine the upper bound of explainable variance.

For White participants, the Single White CNN best predicted behavioral responses on White-face trials (r = .80, p < .0001; bootstrap test), followed by the Dual CNN (r = .78, p < 0.0001) and the Single Asian CNN (r = .70, p < 0.0001). The Single White CNN correlated significantly more strongly with human behavior than either the Dual CNN or the Single Asian CNN (all p < .0001; Bonferroni-corrected). When White participants viewed Asian faces, behavior was best explained by both the Single Asian CNN (r = .75, p < .0001) and the Dual CNN (r = .75, p < .0001), with no significant difference between the two (p = 1). Both, however, predicted behavior significantly better than the Single White CNN (r = .71; p < .0001; p < .001 vs. Single Asian; p < .0001 vs. Dual).

For Asian participants, a similar group-specific pattern emerged. On Asian-face triplets, both the Single Asian CNN (r = .78, p < .0001; bootstrap test) and the Dual CNN (r = .77, p < .0001) showed strong correspondence with human behavior and did not differ significantly from each other (p = .056; Bonferroni-corrected). In contrast, when Asian participants viewed White faces, the Single White CNN (r = .75, p < .0001) and the Dual CNN (r = .74, p < .0001) provided a better fit than the Single Asian CNN (r = .67 ; p < .0001). The Single White CNN and Dual CNN performed comparably (p = .1) and both outperformed the Single Asian CNN (both p < .0001). Across both participant groups, this asymmetric generalization pattern likely reflects differences in real-world visual experience: unlike the single-trained CNNs, which never encountered the other facial group, human observers have at least some diverse exposure (Balas & Saville, 2017). These differences help explain why the Single White CNN aligns better with Asian participants viewing White faces than the Single Asian CNN, and why the Single Asian CNN aligns better with White participants viewing Asian faces: in each case, the model trained on a given facial

group provides a closer match to the observers' greater relative experience with that group than the model that had no exposure at all.

Taken together, the single-trained CNNs aligned best with participants who were most familiar with the facial group each model was trained on, showing the strongest correspondence on own-group faces and capturing a substantial portion of the explainable variance in human behavior. Importantly, the Dual CNN performed comparably to the single-trained CNNs across all conditions, even outperforming them on cross-group trials. This pattern indicates that diverse training fosters more flexible, generalizable representations, yielding a model whose decision patterns more closely reflect human face processing across groups.

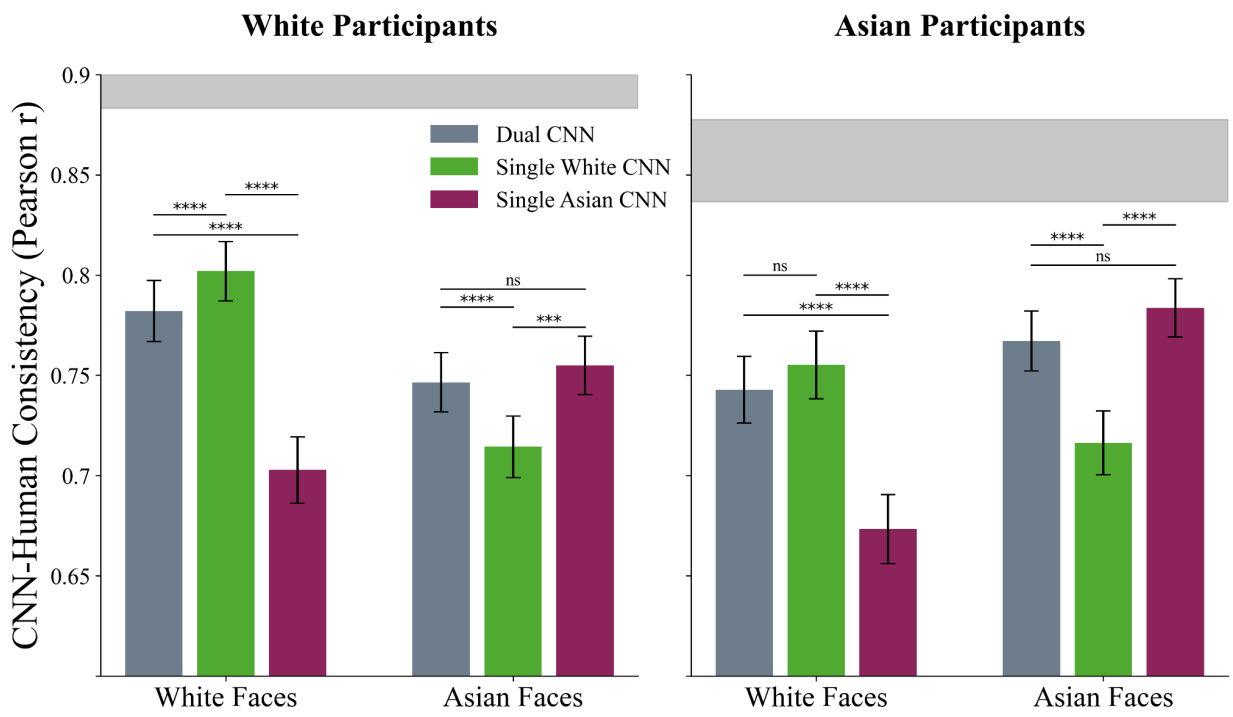

Figure 4. ***Behavioral consistency between human participants and CNNs.*** *CNNs' choice probabilities (see Methods) were correlated with the mean human responses across participants, computed separately for White (left panel; n = 269) and Asian (right panel; n = 102) participant groups. Error bars represent bootstrapped standard deviations of the correlation coefficients. Gray horizontal bars indicate Spearman-Brown-corrected noise ceilings. Asterisks denote significance levels: ns, not significant (p > 0.05); ** p < .001; *** p < .0001.*

Discussion

Our aim in this study was to investigate the computational mechanisms underlying the ORE by using CNNs as controlled models of visual experience. We addressed three questions: (1) whether diverse visual exposure reduces ORE-like biases in CNNs, and if so, (2) through which representational mechanisms, and (3) whether these learned representational geometries predict human behavioral patterns across groups. Consistent with our first question, we replicated that CNNs trained exclusively on a single face set exhibited a clear ORE, whereas the Dual CNN showed markedly reduced bias and robust identity recognition across both groups. Lesioning and RSA analyses revealed that this mitigation arises not from segregated group-specific subsystems but from an integrated representational space that uses partially overlapping feature sets to individuate faces across groups. Finally, the Dual CNN provided the closest match to human face-matching behavior for both participant groups, suggesting that diverse visual experience leads to representational geometries that align more closely with human perceptual strategies.

Our findings mirror developmental studies in humans, where early exposure to diverse faces is associated with reduced ORE and individuated processing of less familiar groups (Anzures et al., 2012, 2013; Bar-Haim et al., 2006; Brigham & Malpass, 1985; Hancock & Rhodes, 2008; McKone et al., 2019; Spangler et al., 2013; Suhrke et al., 2014). Much like infants exposed to visually diverse environments, the Dual CNN developed more generalizable face representations, suggesting that diversity during training, whether biological or artificial, plays a critical role in shaping the flexibility of face-identity processing systems. Importantly, in humans, the timing of exposure appears crucial: diverse experience early in development has a stronger mitigating effect than exposure later in life (McKone et al., 2019). Future work could investigate whether artificial systems exhibit analogous “developmental windows”.

These findings align with the “face space” theory (Valentine, 1991), which posits that other-race faces are represented in a more compact and thus less informative representational space. Recent EEG and neural decoding studies support this account, showing that other-race faces elicit less discriminable neural signals and homogeneous representations (Shoura et al., 2025; Vizioli et al., 2010). This mirrors the compressed and less separable representation as we observed in the single-trained CNNs. In contrast, the Dual CNN showed that training diversity expands the “face space” enabling finer-grained and more structured identity representations across groups. This broadened representational geometry is consistent with recent computational work showing that richer and more varied visual training improves the separability and organization of learned object representations (Cohen et al., 2020; Mehrer et al., 2021).

Previous computational studies have documented ORE-like biases in CNNs trained on homogeneous face datasets (Dobs et al., 2023; Tian et al., 2021; A. Wang et al., 2023). Our results

replicate these effects but extend them in several key ways. First, we systematically compared models trained on fully controlled and large-scale datasets, isolating the contribution of training diversity with fewer confounds than in earlier work. Second, by integrating lesioning with representational similarity analyses, we moved beyond correlational descriptions of representational geometry to causally test whether the features supporting recognition are shared or group-specific, allowing us to distinguish integrated from segregated representational geometries. Third, by comparing model predictions with human behavioral data across two participant groups, we show that the Dual CNN provides the strongest overall correspondence with human performance, underscoring the relevance of these computational mechanisms for understanding human perception.

Together, these results support an interpretation of the ORE as a form of representational overfitting when exposure is restricted to a narrow subset of faces: the system allocates representational resources to maximize discrimination within the familiar group, at the cost of compressed and less informative representations of unfamiliar groups. Importantly, such specialization may not be purely detrimental. In biological systems, this tuning may reflect an adaptive optimization to the statistics of one's social environment, enhancing discrimination for the identities observers most frequently need to individuate. In this sense, the ORE may constitute a "feature" of efficient resource allocation, albeit one that becomes a "bug" in modern multicultural environments where cross-group recognition is increasingly important. Future work should investigate the benefits and costs associated with maintaining a shared versus specialized feature space, and may draw parallels to other domains such as language, where bilingualism also shows complex trade-offs (Lehtonen et al., 2023).

Human-model comparisons also revealed a meaningful asymmetry between artificial training and modern human experience. Single-trained CNNs best captured participant choices for their own-group faces, but on other-group trials, human behavior aligned more closely with the Dual CNN or with the CNN trained on the corresponding group. Unlike CNNs exposed exclusively to one facial group, human observers, no matter how limited their reported contact, inevitably accumulate heterogeneous visual input from indirect sources such as media, advertising, or incidental encounters. Such incidental exposure contributes to recognition ability (Balas & Saville, 2017), whereas self-reported contact explains only modest variance in other-group performance (Wong et al., 2020). More broadly, our findings indicate that cross-group performance reflects not only the observer's perceptual history, but also the structure of the recognition task itself. Depending on which visual cues are diagnostic in a given trial, task demands can either accentuate or attenuate group differences in both humans and models.

Our work also provides insights into the potential origins of the ORE by showing that perceptual-expertise mechanisms (Tanaka et al., 2013) are sufficient to produce ORE-like biases in a system optimized solely for face recognition. Because CNNs lack any socio-motivational constructs, the emergence of ORE in these models must arise from perceptual learning dynamics alone, specifically from asymmetrical experience. Moreover, we show that this bias can be substantially mitigated when the

model receives balanced, diverse facial input. Together, these findings suggest that the human ORE can, in principle, be explained by perceptual factors alone, while socio-cognitive factors (Bernstein et al., 2007; Shriver et al., 2008) may modulate or amplify this bias rather than generate it de novo. This challenges pure socio-cognitive explanations and instead supports hybrid accounts in which perceptual-expertise mechanisms provide the foundational structure on which socio-cognitive influences act (Ficco et al., 2023; Young et al., 2012).

From a machine learning standpoint, our findings echo broad concerns in the AI fairness literature: models trained on imbalanced datasets not only underperform on unfamiliar inputs, but also exhibit amplified representational biases. This is consistent with prior work showing how skewed training distributions can exacerbate bias in computer vision systems (Raji & Fried, 2021; Singh et al., 2021; Tian et al., 2021). Such imbalances can lead to poorly generalizable models that behave unreliably or unfairly when deployed in real-world settings. By showing that a model trained on a more diverse data set (Dual CNN) improves performance across groups and develops more robust, generalizable representational geometry, our results reinforce calls in the AI community for proactive dataset diversification as a critical step toward building fairer and more reliable vision systems.

This study has several limitations. First, all analyses were performed using a feedword CNN architecture. Although well established for face recognition (Parkhi et al., 2015), it remains to be tested whether similar mechanisms emerge in architectures incorporating recurrence, attention, or feedback. Second, our training regime implemented an extreme imbalance (100% Asian vs. 100% White faces), future work should examine more graded imbalances (e.g., 80/20 distributions) and variations in exposure quality (e.g., many identities collapsed into a single category) to further differentiate theoretical accounts of the ORE (Qian et al., 2017). Finally, when relating models to human behavior, we focused on observers whose everyday experience is skewed toward their own facial group and did not include individuals with substantial, balanced exposure to both groups. Consequently, we could not directly test whether the Dual CNN best captures behavior in observers with genuinely diverse face experience.

In summary, our study shows that training diversity in CNNs enhances recognition performance across groups and promotes integrated, generalizable representational geometries that more closely resemble human perceptual organization. By combining behavioral alignment with representational similarity analysis and lesioning, we show that diverse experience mitigates group-specific overfitting, improves identity discrimination, and yields representational geometries that better mirror those in human observers, even when those observers have limited contact with other-group faces. More broadly, our results underscore the central role of exposure diversity in shaping both artificial and biological face-processing systems and highlight representational diversity as essential for building accurate, equitable, and cognitively plausible AI models. Together, these contributions move beyond documenting ORE in CNNs and toward a computational account of its origins, its representational consequences, and the conditions under which it can be reduced.

Acknowledgements

We thank the members of the VCCN lab for valuable feedback and discussion, and Ben de Haas for comments on an earlier version of the manuscript. This work was supported by the European Research Council (ERC Starting Grant DEEPFUNC, ERC-2023-STG-101117441), the Hessian Ministry of Higher Education, Research, Science and the Arts (LOEWE Start Professorship), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223) and SFB/TRR 135 (Project No. 222641018, TP C9).

References

Abudarham, N., Grosbard, I., & Yovel, G. (2021). Face Recognition Depends on Specialized Mechanisms Tuned to View-Invariant Facial Features: Insights from Deep Neural Networks Optimized for Face or Object Recognition. *Cognitive Science*, *45*(9), e13031. https://doi.org/10.1111/cogs.13031

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., Tanaka, J. W., & Lee, K. (2013). Developmental Origins of the Other-Race Effect. *Current Directions in Psychological Science*, *22*(3), 173–178. https://doi.org/10.1177/0963721412474459

Anzures, G., Wheeler, A., Quinn, P. C., Pascalis, O., Slater, A. M., Heron-Delaney, M., Tanaka, J. W., & Lee, K. (2012). Brief daily exposures to Asian females reverses perceptual narrowing for Asian faces in Caucasian infants. *Journal of Experimental Child Psychology*, *112*(4), 484–495. https://doi.org/10.1016/j.jecp.2012.04.005

Balas, B., & Saville, A. (2017). Hometown size affects the processing of naturalistic face variability. *Vision Research*, *141*, 228–236. https://doi.org/10.1016/j.visres.2016.12.005

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and Nurture in Own-Race Face Processing. *Psychological Science*, *17*(2), 159–163. https://doi.org/10.1111/j.1467-9280.2006.01679.x

Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect: Mere Social Categorization Is Sufficient to Elicit an Own-Group Bias in Face Recognition. *Psychological Science*, *18*(8), 706–712.

Brigham, J. C., & Malpass, R. S. (1985). The Role of Experience and Contact in the Recognition of Faces Of Own- and Other-Race Persons. *Journal of Social Issues*, *41*(3), 139–155. https://doi.org/10.1111/j.1540-4560.1985.tb01133.x

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). *VGGFace2: A dataset for recognising faces across pose and age* (No. arXiv:1710.08092). arXiv. https://doi.org/10.48550/arXiv.1710.08092

Cohen, U., Chung, S., Lee, D. D., & Sompolinsky, H. (2020). Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, *11*(1), 746. https://doi.org/10.1038/s41467-020-14578-5

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, *8*(11), eabl8913. https://doi.org/10.1126/sciadv.abl8913

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge

in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, *120*(32), e2220642120. https://doi.org/10.1073/pnas.2220642120

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., Van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Ficco, L., Müller, V. I., Kaufmann, J. M., & Schweinberger, S. R. (2023). Socio-cognitive, expertise-based and appearance-based accounts of the other-'race' effect in face perception: A label-based systematic review of neuroimaging results. *British Journal of Psychology*, *114*(S1), 45–69. https://doi.org/10.1111/bjop.12595

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*(8), 845–850. https://doi.org/10.1038/90565

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., Khuvis, S., Herrero, J. L., Irani, M., Mehta, A. D., & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, *10*(1), 4934. https://doi.org/10.1038/s41467-019-12623-6

Gwyn, T., Roy, K., & Atay, M. (2021). Face Recognition Using Popular Deep Net Architectures: A Brief Comparative Study. *Future Internet*, *13*(7), 164. https://doi.org/10.3390/fi13070164

Hancock, K. J., & Rhodes, G. (2008). Contact, configural coding and the other-race effect in face recognition. *British Journal of Psychology*, *99*(1), 45–56. https://doi.org/10.1348/000712607X199981

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*(1), 1872. https://doi.org/10.1038/s41467-021-22078-3

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254. https://doi.org/10.1016/j.tins.2022.12.008

Kelly, D. J., Liu, S., Ge, L., Quinn, P. C., Slater, A. M., Lee, K., Liu, Q., & Pascalis, O. (2007). Cross-Race Preferences for Same-Race Faces Extend Beyond the African Versus Caucasian Contrast in 3-Month-Old Infants. *Infancy*, *11*(1), 87–95. https://doi.org/10.1207/s15327078in1101_4

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The Other-Race Effect Develops During Infancy: Evidence of Perceptual Narrowing. *Psychological Science*, *18*(12), 1084–1089. https://doi.org/10.1111/j.1467-9280.2007.02029.x

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., Ge, L., & Pascalis, O. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*(6). https://doi.org/10.1111/j.1467-7687.2005.0434a.x

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/neuro.06.004.2008

Lehtonen, M., Fyndanis, V., & Jylkkä, J. (2023). The relationship between bilingual language use and executive functions. *Nature Reviews Psychology*, *2*(6), 360–373. https://doi.org/10.1038/s44159-023-00178-9

McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., Kidd, E., & Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports*, *9*(1), 12820. https://doi.org/10.1038/s41598-019-49202-0

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). *An ecologically motivated image dataset for deep learning yields better models of human vision*.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

O'Toole, A. J., & Castillo, C. D. (2021). Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning. *Annual Review of Vision Science*, *7*(1), 543–570. https://doi.org/10.1146/annurev-vision-093019-111701

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. *Trends in Cognitive Sciences*, *22*(9), 794–809. https://doi.org/10.1016/j.tics.2018.06.006

O'Toole, A. J., Deffenbacher, K., Abdi, H., & Bartlett, J. C. (1991). Simulating the 'Other-race Effect' as a Problem in Perceptual Learning. *Connection Science*, *3*(2), 163–178. https://doi.org/10.1080/09540099108946583

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *Procedings of the British Machine Vision Conference 2015*, 41.1-41.12. https://doi.org/10.5244/C.29.41

Platz, S. J., & Hosch, H. M. (1988). Cross-Racial/Ethnic Eyewitness Identification: A Field Study[1]. *Journal of Applied Social Psychology*, *18*(11), 972–984. https://doi.org/10.1111/j.1559-1816.1988.tb01187.x

Qian, M. K., Quinn, P. C., Heyman, G. D., Pascalis, O., Fu, G., & Lee, K. (2017). Perceptual individuation training (but not mere exposure) reduces implicit racial bias in preschool children. *Developmental Psychology*, *53*(5), 845–859. https://doi.org/10.1037/dev0000290

Raji, I. D., & Fried, G. (2021). *About Face: A Survey of Facial Recognition Evaluation*.

Rosemblaum, M., Guy, N., Grosbard, I., Kliger, L., Abudarham, N., & Yovel, G. (2025). Concurrent emergence of view invariance, sensitivity to critical features, and identity face classification through visual experience: Insights from deep learning algorithms. *Journal of Vision*, *25*(8), 2. https://doi.org/10.1167/jov.25.8.2

Rossion, B., & Michel, C. (2012). *An experience-based holistic account of the other-race face effect. The Oxford handbook of face perception*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199559053.013.0012

Shoura, M., Liang, Y. Z., Sama, M. A., De, A., & Nestor, A. (2025). Revealing the neural representations underlying other-race face perception. *Frontiers in Human Neuroscience*, *19*, 1543840. https://doi.org/10.3389/fnhum.2025.1543840

Shriver, E. R., Young, S. G., Hugenberg, K., Bernstein, M. J., & Lanter, J. R. (2008). Class, Race, and the Face: Social Context Modulates the Cross-Race Effect in Face Recognition. *Personality and Social Psychology Bulletin*, *34*(2), 260–274. https://doi.org/10.1177/0146167207310455

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (No. arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Singh, R., Majumdar, P., Mittal, S., & Vatsa, M. (2021). *Anatomizing Bias in Facial Analysis* (No. arXiv:2112.06522). arXiv. https://doi.org/10.48550/arXiv.2112.06522

Spangler, S. M., Schwarzer, G., Freitag, C., Vierhaus, M., Teubert, M., Fassbender, I., Lohaus, A., Kolling, T., Graf, F., Goertz, C., Knopf, M., Lamm, B., & Keller, H. (2013). The Other-Race Effect in a Longitudinal Sample of 3-, 6- and 9-Month-Old Infants: Evidence of a Training Effect. *Infancy*, *18*(4), 516–533. https://doi.org/10.1111/j.1532-7078.2012.00137.x

Suhrke, J., Freitag, C., Lamm, B., Teiser, J., Fassbender, I., Poloczek, S., Teubert, M., VÃ¶hringer, I. A., Keller, H., Knopf, M., Lohaus, A., & Schwarzer, G. (2014). The other-race effect in 3-year-old German and Cameroonian children. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00198

Tian, J., Xie, H., Hu, S., & Liu, J. (2021). Multidimensional Face Representation in a Deep Convolutional

Neural Network Reveals the Mechanism Underlying AI Racism. *Frontiers in Computational Neuroscience*, *15*, 620281. https://doi.org/10.3389/fncom.2021.620281

Turk, D. J., Handy, T. C., & Gazzaniga, M. S. (2005). Can perceptual expertise accountfor the own-race bias in face recognition? A split-brain study. *Cognitive Neuropsychology*, *22*(7), 877–883. https://doi.org/10.1080/02643290442000383

Valentine, T. (1991). A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

Van Dyck, L. E., & Gruber, W. R. (2023). Modeling Biological Face Recognition with Deep Convolutional Neural Networks. *Journal of Cognitive Neuroscience*, *35*(10), 1521–1537. https://doi.org/10.1162/jocn_a_02040

Vizioli, L., Rousselet, G. A., & Caldara, R. (2010). Neural repetition suppression to identity is abolished by other-race faces. *Proceedings of the National Academy of Sciences*, *107*(46), 20081–20086. https://doi.org/10.1073/pnas.1005751107

Walker, P. M., & Hewstone, M. (2006). A perceptual discrimination investigation of the own-race effect and intergroup experience. *Applied Cognitive Psychology*, *20*(4), 461–475. https://doi.org/10.1002/acp.1191

Walker, P. M., Silvert, L., Hewstone, M., & Nobre, A. C. (2008). Social contact and other-race face processing in the human brain. *Social Cognitive and Affective Neuroscience*, *3*(1), 16–25. https://doi.org/10.1093/scan/nsm035

Walker, P. M., & Tanaka, J. W. (2003). An Encoding Advantage for Own-Race versus Other-Race Faces. *Perception*, *32*(9), 1117–1125. https://doi.org/10.1068/p5098

Wan, L., Crookes, K., Reynolds, K. J., Irons, J. L., & McKone, E. (2015). A cultural setting where the other-race effect on face recognition has no social–motivational component and derives entirely from lifetime perceptual experience. *Cognition*, *144*, 91–115. https://doi.org/10.1016/j.cognition.2015.07.011

Wang, A., Sliwinska, M. W., Watson, D. M., Smith, S., & Andrews, T. J. (2023). Distinct patterns of neural response to faces from different races in humans and deep networks. *Social Cognitive and Affective Neuroscience*, *18*(1), nsad059. https://doi.org/10.1093/scan/nsad059

Wang, M., & Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, *429*, 215–244. https://doi.org/10.1016/j.neucom.2020.10.081

Wong, H. K., Estudillo, A. J., Stephen, I. D., & Keeble, D. R. T. (2021). The other-race effect and holistic processing across racial groups. *Scientific Reports*, *11*(1), 8507. https://doi.org/10.1038/s41598-021-87933-1

Wong, H. K., Stephen, I. D., & Keeble, D. R. T. (2020). The Own-Race Bias for Face Recognition in a Multiracial Society. *Frontiers in Psychology*, *11*, 208. https://doi.org/10.3389/fpsyg.2020.00208

Xiong, Z., Wang, Z., Du, C., Zhu, R., Xiao, J., & Lu, T. (2018). An Asian Face Dataset and How Race Influences Face Recognition. In R. Hong, W.-H. Cheng, T. Yamasaki, M. Wang, & C.-W. Ngo (Eds.), *Advances in Multimedia Information Processing – PCM 2018* (Vol. 11165, pp. 372–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-00767-6_35

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). *Learning Face Representation from Scratch* (No. arXiv:1411.7923). arXiv. https://doi.org/10.48550/arXiv.1411.7923

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2012). Perception and Motivation in Face Recognition: A Critical Review of Theories of the Cross-Race Effect. *Personality and Social Psychology Review*, *16*(2), 116–142. https://doi.org/10.1177/1088868311418987

Yovel, G., Grosbard, I., & Abudarham, N. (2022). *Deep learning models of perceptual expertise support a domain-specific account*. https://doi.org/10.1101/2022.12.01.518342

Zeman, A., Leers, T., & De Beeck, H. O. (2022). *Mooney Face Image Processing in Deep Convolutional Neural Networks Compared to Humans*. Bioinformatics. https://doi.org/10.1101/2022.03.21.485240

Zhao, M., Hayward, W. G., & Bulthoff, I. (2014). Face format at encoding affects the other-race effect in face memory. *Journal of Vision*, *14*(9), 6–6. https://doi.org/10.1167/14.9.6