The Mechanism of Average Representation Based on Synthetic Average Stimuli: Evidence from Average Facial Attractiveness

Authors: Tian Xinran, Wenxia Hou, Ou Yuxiao, Yi Bing, Chen Wenfeng, Shang Junchen, Chen Wenfeng, Shang Junchen

Date: 2021-02-18T00:00:00+00:00

Abstract

Humans can rapidly extract statistical information from ensembles to form average representations. Regarding the generation mechanism of average representations, researchers have proposed two viewpoints: integrating ensemble members to synthesize an average stimulus, or computing the mean of feature values across ensemble members. In previous studies, the results from these two approaches—the feature values of synthesized average stimuli and the computed mean of member feature values—were similar, making it difficult to distinguish between the two viewpoints. Since the mean of attractiveness ratings for multiple faces differs from the attractiveness rating of an average face synthesized from these faces, the present study employed classic ensemble discrimination tasks (Experiments 1 and 2) and attractiveness evaluation tasks (Experiments 3 and 4) to provide supporting evidence for the view that average representation generation originates from synthesized average stimuli. The four experiments respectively employed large-capacity and small-capacity face ensembles to investigate the formation mechanism of average representations. The results revealed that both large and small ensembles formed average stimuli, and that the evaluation and processing of average representations may rely more on synthesized average stimuli rather than simply averaging member feature values; additionally, ensemble attractiveness exhibited an overestimation phenomenon, though this overestimation occurred less frequently in small ensembles, indicating that the effect of average stimuli is influenced by ensemble size. This study provides novel evidence for the formation mechanism of ensemble average representations and the generation mechanism of overestimation phenomena in facial ensemble attractiveness.

Full Text

Average Percept in Ensemble Perception is Based on Morphed Average Object: Evidence from Average Facial Attractiveness

Xinran Tian[1], **Wenxia Hou**[1], **Yuxiao Ou**[1], **Bing Yi**[1], **Wenfeng Chen**[1], ***Junchen Shang***[2] [1]Department of Psychology, Renmin University of China, Beijing 100872, China (wchen@ruc.edu.cn) [2]School of Psychology, Liaoning Normal University, Dalian 116029, China (junchen_{20081}@163.com)

Abstract: [**Objective**] Previous research demonstrated that ensemble perception of groups can be formed rapidly by extraction of the average of high-level complex features. However, it is unclear whether the average percept is the outcome of extraction from the characteristic value of the average stimulus (for example, average face) created from group members, or from calculation of the average value of group members' characteristic values. The above two values were confused with each other in prior research, since most average values of group members are similar to the characteristic value of the average stimulus. However, the attractiveness rating of the average face created from a group of faces is usually systematically higher than the mean value of attractiveness ratings of this group of faces. Therefore, it is easier to explore how the ensemble coding of crowd face attractiveness (i.e., group attractiveness) is formed by comparing the attractiveness of the average face with the mean value of attractiveness rating of a group of faces. This could provide a useful approach to explore how the average percept is formed.

[**Methods**] The present study used the average discrimination paradigm (Experiment 1 & 2) and the scoring paradigm (Experiment 3 & 4) to clarify the mechanism of the formation of average percept by comparing the group attractiveness with the attractiveness of average face. To tackle this issue, whether the average face was presented in the group of faces or not was manipulated (conditions: Avg vs. NoAvg). Group size was also manipulated to explore whether group size modulated the formation of average percept. In the average discrimination paradigm, a group of faces served as group stimuli to be compared with the probe face for attractiveness. Participants were asked to judge which is more attractive between the group stimuli and the probe face. In the scoring paradigm, participants were asked to rate the attractiveness of group stimuli, the average face created from the group, and each face of the group in an isolated manner. Each group consisted of twelve faces (in Experiments 1 and 3) or four faces (in Experiments 2 and 4). There were two kinds of groups: one in which all group members are original faces, without the average face, and another in which an average face morphed from other original faces was included in the group.

[**Results**] In Experiment 1, the proportions for judging the probe average face as more attractive than group attractiveness in the Avg condition were simi-

lar to the NoAvg condition. In Experiment 2, when the set size was four, the proportions for judging the probe average face as more attractive than group attractiveness were significantly higher in the NoAvg condition. Moreover, in Experiment 3, the ratings for group attractiveness were not significantly different between Avg and NoAvg conditions. This may indicate that the group attractiveness is based on the average face which was created from group members rather than the mean value calculated from group members' attractiveness. In addition, the diffusion model analysis showed that the coding time was longer for the NoAvg condition, which indicated that the formation of average face needed cognitive resource. In Experiment 4, when the set size was four, the attractiveness rating of the average face was significantly higher than group ratings for the two kinds of groups. The different results in different group sizes may be interpreted as the outcome of weakened average percept caused by the salient individual face representations in small groups. This was evident from several analyses: (1) group attractiveness and the attractiveness of morphed average face decreased with smaller set size (Experiment 4); (2) when the probe face was morphed average face, the proportion for judging probe face as more attractive than group attractiveness was greater, compared with the condition when the probe was a new face whose attractiveness was similar to the morphed average face (Experiment 2); (3) the performance for the hypothesized condition with average percept included in the set is between the conditions with/without real average face included (Experiment 2-4). In addition, comparing with Experiment 1, the information accumulation speed in Experiment 2 is slower, and the processing time of group attractiveness is longer, reflecting the disturbance of the individual face representation.

[**Conclusions**] Group attractiveness is based on the morphed average face, and the ensemble percept relies on the extraction from the average stimulus created from the group.

Keywords: average representation; face attractiveness; average face

1 Introduction

Our visual system receives massive amounts of information every moment, much of which is highly structured. This structured information is often similar and exists in the form of ensembles. People can perform perceptual averaging on these ensembles, extracting the average representation of all members within a set with considerable precision (Alvarez, 2011; Haberman & Whitney, 2012; Whitney & Yamanashi Leib, 2018). This process involves low-level features such as size, orientation, brightness, and position (Alvarez & Oliva, 2008; Ariely, 2001; Bauer, 2009; Parkes et al., 2001), as well as high-level social information including facial identity, gender, and expression (Haberman & Whitney, 2007; Haberman et al., 2015; Li et al., 2016). Many studies have focused on how average representations are generated in the brain: whether through integrating ensemble members to form a representation of an average stimulus or through calculating the mean value of individual members' features (Maule & Franklin,

2015; Whitney & Yamanashi Leib, 2018). In previous research, average representations have typically been measured using the mean value of ensemble members, implicitly assuming that average representation is equivalent to the ensemble members' average value. However, because the characteristic value of an ensemble' s average stimulus is often very similar to the mean of members' characteristic values, this assumption cannot serve as evidence to distinguish whether average representation formation results from creating a representation of an average stimulus in the brain or from computing the mean value of ensemble members. Therefore, the mechanism underlying average representation formation remains an unresolved issue. One approach to solving this problem is to separate the characteristic value of an ensemble' s average stimulus from the mean of ensemble members' characteristic values. Facial attractiveness is particularly suitable for addressing this issue because the attractiveness of an average face is typically higher than the mean attractiveness of the individual faces that compose it (Carragher et al., 2018; Komori et al., 2009). To effectively distinguish between the representation of an average stimulus and simple mean value calculation, this study investigates whether an average stimulus representation is formed during perceptual averaging by leveraging the difference between ensemble facial attractiveness average representation and the mean value of all faces' attractiveness in the ensemble.

1.1 The Debate on Average Representation Formation Mechanisms

Previous research has proposed two main explanations for how average representations are formed: holistic encoding based on distributed attention and individual encoding based on focused attention. The holistic encoding view posits that the visual system processes ensemble stimuli in parallel, enabling accurate representation of the ensemble average while preventing accurate representation of individual members within the ensemble (Ariely, 2001). In contrast, the individual encoding view suggests that the visual system concentrates limited attentional resources on a small sample extracted from the ensemble, processes these samples in detail, and then infers the ensemble' s average representation through mean value calculation of the sample information (de Fockert & Marchant, 2008; Myczek & Simons, 2008).

From the perspective of general visual processing, visual information is processed hierarchically. The debate between holistic and individual encoding can be partly reduced to the priority of holistic versus individual visual processing levels during average representation formation. Recently, the Reverse Hierarchy Theory of visual processing (Hochstein & Ahissar, 2002; Hochstein et al., 2015) has proposed that holistic and individual processing exhibit a reverse hierarchical relationship, where statistical representations, as high-level constructs built through rapid bottom-up processes, precede the detection of individual representations. Reverse Hierarchy Theory suggests that conscious perception of holistic representations (such as scene gist) begins in high-level cortex and is a perceptual process based on inputs from lower-level cortex. In the initial

stage of visual processing, we can only consciously detect holistic representations of visual scenes (such as gist) and cannot detect the antecedent details (i.e., individual details that constitute the holistic representation) of high-level holistic representations. Following this prioritized hierarchical processing, the visual system then directs attention to specific low-level cortical processing units to extract local detail information. In other words, holistic representations in high-level cortex return to local processing in a top-down manner (reverse hierarchy return) to confirm (or correct) the initial holistic representation estimates (Hochstein et al., 2015). Therefore, according to Reverse Hierarchy Theory, average representations are initially formed by the brain integrating coarse individual information rather than through mean value calculation based on precise individual representations, though they may be corrected by individual representations in later processing stages. However, this remains a theoretical inference that requires more direct evidence.

1.2 The Group Attractiveness Effect and Average Face Attractiveness

Similar to other facial features such as expression and identity, scholars have previously speculated that the average attractiveness of a face ensemble should equal the average of each face' s attractiveness (Abbas & Duchaine, 2008; Brady & Alvarez, 2015; Haberman & Whitney, 2012). Early studies found that the attractiveness of an ensemble composed of three young male faces with different attractiveness levels was exactly equal to the average attractiveness of the three individuals (Anderson, 1965; Anderson et al., 1973).

However, researchers have also found different results. Van Osch et al. (2015) systematically manipulated ensemble size and discovered that when the number of faces in an ensemble exceeded six, the ensemble' s attractiveness rating was significantly higher than the mean of members' ratings. This phenomenon is called the group attractiveness effect. The group attractiveness effect can also occur in small face ensembles under specific conditions. For example, Willis (1960) found that when an ensemble contained only two or three faces, the ensemble' s attractiveness rating was more extreme than the members' average, with high-attractiveness ensembles receiving higher ratings than the mean of individual members.

Van Osch et al. (2015) proposed that the possible mechanism underlying the group attractiveness effect is that perceptual processing of face ensembles forms a representation of the average face. That is, the average attractiveness representation of a face ensemble is not the average of members' attractiveness values but rather results from participants morphing and fusing all faces in the ensemble into a new average face, which then influences the evaluation of ensemble facial attractiveness. Facial attractiveness is strongly correlated with facial averageness (O' Toole et al., 1999; Rhodes et al., 2001), meaning that the attractiveness of an average face is enhanced due to the average properties it carries, making it higher than the average attractiveness of the faces that compose it (Carragher et al., 2018).

Van Osch et al.'s (2015) explanation regarding average faces aligns with the view that stimulus ensembles form representations of average stimuli but does not reconcile the contradiction between findings of ensemble attractiveness equaling the average (Anderson et al., 1973; Luo & Zhou, 2018) and the group attractiveness effect. We argue that this is not contradictory. Anderson et al. (1973) used small face ensembles (N = 3 or 4), while the group attractiveness effect emerged in larger ensembles (N ≥ 8) (Van Osch et al., 2015). This difference in ensemble size may cause inconsistent results: relative to large face ensembles, when ensembles are small, processing resources are sufficient for precise processing of individual members, making individual member representations more salient and more likely to interfere with average face representation. Li et al. (2016) provided supporting evidence for this view: under limited processing resources, average representations have an advantage over individual representations, with lower precision for individual representations; but when processing resources are relatively abundant, individual representation precision increases while average representation precision decreases. According to Reverse Hierarchy Theory, average representation formation does not require precise individual representations, whereas numerical calculation requires high individual representation precision. Additionally, the attractiveness of average faces is also affected by the number of faces in the ensemble, with average faces synthesized from small ensembles having lower attractiveness than those from large ensembles (Langlois & Roggman, 1990). Indeed, Van Osch et al. (2015) also found that when ensemble size decreased, the probability of the group attractiveness effect occurring dropped substantially. This may be because the average face formed by small ensembles has relatively low attractiveness or is subject to interference, or because no average face is formed and processing relies on mean value calculation instead. Therefore, the group attractiveness effect is reduced in small ensembles, but its mechanism requires clarification.

1.3 Research Questions and Hypotheses

Currently, the mechanism of ensemble average representation formation remains at the theoretical level without direct evidence. The synthesis of an average stimulus, as a possible mechanism, can better explain theoretical issues and experimental phenomena related to ensemble representation processing. First, it provides a solution to the debate between holistic and individual encoding of average representations. Second, it offers empirical support for the Reverse Hierarchy Theory of visual processing. Finally, it provides an empirical explanation for the group attractiveness effect in face ensembles. This study uses the average discrimination task (Experiments 1 and 2, mean discrimination paradigm, Haberman & Whitney, 2009) and the attractiveness rating task (Experiments 3 and 4) to explore the mechanism of the group attractiveness effect by comparing ensemble attractiveness with average face attractiveness, thereby clarifying whether an average stimulus representation is formed during perceptual averaging. The average discrimination task requires participants to perform perceptual comparison between a single stimulus and the ensemble average representation,

using responses to the ensemble average representation after perceptual comparison as the dependent variable to infer whether an average representation exists. The attractiveness rating task requires participants to rate the attractiveness of the ensemble as a whole and the average stimulus, directly reflecting the perceptual process of average representation. We created conditions where the ensemble contained an average face. If an average face is formed during ensemble processing, then whether the ensemble originally contained an average face should have no impact on results. If no average face is formed, then including an average face in the ensemble should facilitate the average discrimination process or enhance ensemble attractiveness. Experiments 1 and 3 compare the proportions of keypress responses between conditions with and without average face stimuli in large ensembles to provide more direct evidence for average face formation. Experiments 2 and 4 examine whether the relationship between ensemble average face and ensemble attractiveness changes across different ensemble sizes to provide experimental data supporting the debate between average representation formation and mean value calculation. Additionally, diffusion model analysis is used to investigate the ensemble processing procedure, providing evidence regarding information processing for Experiments 1 and 2. Based on different views of average representation formation mechanisms, the following predictions can be made for experimental results:

(1) If average representation is calculated through the mean value of ensemble members' attractiveness, then due to the high attractiveness of average faces, ensembles containing average face stimuli should be more attractive than those without, being closer to average face attractiveness. Consequently, when an ensemble contains an average face, ensemble attractiveness should increase (Experiments 3 and 4), the tendency to judge ensemble attractiveness as higher should increase in the average discrimination task, and the proportion of judging the probe average face as more attractive should decrease, without being affected by ensemble size (Experiments 1 and 2).

(2) If average representation is generated through forming a representation of an average stimulus, then whether the ensemble contains an average face stimulus should have no effect on either the average discrimination task or the rating task (Experiments 1 and 3). Moreover, in small ensembles where the average face is subject to interference, ensemble attractiveness containing average face stimuli should be closer to average face attractiveness (Experiment 4), thereby reducing the proportion of judging the probe average face as more attractive (Experiment 2).

(3) Based on Hypothesis 2, if reduced group attractiveness effect in small ensembles is due to interference with the average face, then the difference between ensemble attractiveness and average face attractiveness should be larger under conditions without average face stimuli (Experiment 4), resulting in a higher proportion of judging the probe average face as more attractive in small ensembles without average face stimuli (Experiment 2).

(4) If the reduction of the group attractiveness effect in small ensembles is due to relatively lower attractiveness of small ensemble average faces, then the decreased attractiveness of small ensemble average faces leads to smaller differences between ensemble attractiveness and average face (comparison between Experiments 3 and 4), thereby resulting in a lower proportion of judging the probe average face as more attractive in small ensembles (comparison between Experiments 1 and 2).

Experiment 1

Experiment 1 employed an average discrimination task, requiring participants to choose the more attractive option between ensemble attractiveness and average face attractiveness, manipulating whether the average face appeared in the ensemble to determine whether an average face was formed.

2.1 Method

(1) Participants Using G*Power with statistical power = 0.8, medium effect size f = 0.25, and repeated measures with 2 levels of the independent variable (ensemble type), the minimum estimated sample size was N = 34. We recruited 34 students from Renmin University of China (18 female) with a mean age of 20.75 years (SD = 2.02). All were right-handed with normal or corrected-to-normal vision. All participants signed informed consent, and the experiment was approved by the Ethics Committee of the Department of Psychology at Renmin University of China.

(2) Experimental Materials To generate sufficient between-group differences, we selected internet materials as faces with extremely high and low attractiveness. Hair, neck, and ears were removed, facial contours were cropped into ellipses, and images were converted to grayscale for standardization. Some face materials were selected from the female-neutral emotion database in the Chinese Facial Affective Picture System (Wang & Luo, 2005). Images selected from this system and internet materials together comprised the original materials. All materials were rated by 20 Chinese university students (10 female, mean age = 20.54 years, SD = 2.17) on attractiveness level and emotional valence (both on 101-point scales). Materials rated as non-neutral (significantly different from the neutral score of 50) were excluded. The selected materials had valence ratings (M = 49.94, SD = 0.77) that did not differ significantly from neutral, $t(19) = 0.35$, $p = 0.732$. Thirty original faces were selected, including 6 internet materials (4 in the high-attractiveness group, 2 in the low-attractiveness group) and 24 images from the Chinese Facial Affective Picture System.

Based on attractiveness ratings, original materials were divided into high, medium, and low attractiveness groups, with 10 images in each group. The mean scores were 76.33, 43.12, and 22.42 for high, medium, and low attractiveness groups, respectively. Repeated measures ANOVA testing differences among the three groups revealed significant differences, $F(2, 38) = 148.64$, $p <$

0.001, $p^2 = 0.89$. Pairwise comparisons were all significant: low vs. medium attractiveness groups, $t(19) = 7.91$, $p < 0.001$, Cohen' s d = 3.63; medium vs. high attractiveness groups, $t(19) = 9.77$, $p < 0.001$, Cohen' s d = 4.48; low vs. high attractiveness groups, $t(19) = 15.92$, $p < 0.001$, Cohen' s d = 7.30. We then created ensembles composed of faces from different attractiveness groups. Each ensemble contained either 11 or 12 images, with 60 ensembles total in Experiment 1. In any single ensemble, faces from different attractiveness groups were used with varying frequencies, but across all ensembles used in the experiment, faces from different attractiveness groups were used equally often.

Subsequently, we used the face morphing software Abrosoft FantaMorph (Abrosoft Fantamorph 5.4.8, www.fantamorph.com) to create average faces from different face ensembles, yielding 365 images. This software can blend two faces according to a specified ratio by annotating facial features with numerous keypoints (e.g., position, size, and curvature of mouth corners) and then averaging these keypoints to synthesize images. For example, to create an average face from four original faces, we paired the original faces, synthesized them at 50:50 ratios, then synthesized the resulting two images again at 50:50, equivalent to each original face contributing 25% to the final synthesized face. To create an average face from three original faces, we controlled each face' s contribution ratio at 33.3%.

All face images were rated again for attractiveness by 20 students from Renmin University of China (10 female, mean age = 20.35 years, SD = 2.03) to obtain pre-rated scores.

All experimental materials were presented on a 24-inch Dell monitor with a resolution of 1920$×$1080 and a gray background. Participants sat approximately 70 cm from the screen.

(3) Experimental Design A single-factor within-subjects design was used, with ensemble type (ensemble without average face G1 vs. ensemble with average face G2) as the independent variable. Dependent variables were the proportion of judgments that the average face was more attractive and diffusion model parameters (information accumulation rate v, threshold separation a, and non-decision processing time t0; see Results section for details).

(4) Experimental Procedure The experiment employed an average discrimination task. Ensemble stimuli were presented first, followed by probe stimuli. Probe stimuli included the ensemble average face, ensemble member faces, and non-member non-average faces. Because the size relationships with the average face and ensemble mean value were uncertain, making result interpretation difficult, the latter two stimulus types served as filler stimuli (control conditions) in this study.

Regarding ensemble stimulus types, an ensemble composed of 12 original faces represented the "ensemble without average face G1" condition. When 11 original faces composed the ensemble and the average face of ensemble members was added as a new member, this represented the "ensemble with average face G2"

condition. In ensembles containing the average face, its position was randomized. For probe stimulus types, presenting the average face of ensemble members represented the "ensemble average face" condition. When the ensemble stimulus contained the average face, this meant the average face appeared twice. The "present one ensemble member" condition involved presenting one member face from the ensemble excluding the average face. The "new face" condition involved presenting a face that did not appear in the ensemble stimulus.

In each trial, participants first fixated on a central point for 1000 ms, then viewed the ensemble stimulus presented on screen for 2000 ms, followed by a blank screen for 500 ms, and finally a probe face that remained until response. Participants pressed the F or J key to judge which was more attractive—the ensemble stimulus' s overall attractiveness or the probe stimulus' s individual attractiveness. There were 180 trials total, with conditions randomly intermixed. Participants took a break every 60 trials. For probe faces that were new faces or ensemble members, half had pre-rated attractiveness higher than the mean attractiveness of ensemble members, and half lower (see Figure 1 [Figure 1: see original paper]).

Ensemble stimuli were presented in a 4$×$3 matrix. Single face images subtended 5.69° × 6.53° of visual angle. Probe stimulus materials were single faces presented at the center of the screen, with image sizes identical to ensemble member stimuli.

(5) Diffusion Model According to Reverse Hierarchy Theory, although average face formation is rapid, it still requires information accumulation, potentially producing decision response differences between ensembles with and without average faces. To examine such possible differences, we analyzed response decision information using the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) to decompose different cognitive processes, integrating both response time distributions and response accuracy results to further analyze the perceptual mechanism of ensemble facial attractiveness with and without average faces. This model can map decomposed cognitive processes to different model parameters (Voss et al., 2013).

The diffusion model' s basic assumption is that in rapid two-choice tasks, information accumulates continuously from a starting point until reaching a response threshold, at which point the response is activated. The basic diffusion model (Ratcliff, 1978) has four parameters (see Figure 2 [Figure 2: see original paper]): (1) drift rate, denoted as v, indicating the rate of information accumulation; (2) threshold separation, denoted as a, indicating the amount of information needed for decision-making; (3) starting point, denoted as z, indicating pre-decision bias; and (4) duration of non-decisional processes, denoted as t0, including time for encoding and response execution that is not part of decision-making.

Figure 2 shows the diffusion model (translated from Ratcliff & McKoon, 2008, Figure 2), displaying three example paths. Information accumulates gradually from the starting point (z) at an average rate (v) until reaching the threshold

for Response A (a) or Response B (0). Due to random noise, these paths vary across trials.

2.2 Results

(1) Button Response Results We calculated response accuracy for control condition stimuli based on pre-rated attractiveness scores to ensure participants understood the task and responded correctly, and to verify whether pre-rated attractiveness scores were applicable to our participants. We computed the mean attractiveness of ensemble members based on pre-rated scores and compared these with pre-rated probe face attractiveness to determine correct responses. Results showed overall accuracy of 84.72% for probe stimuli that were new faces or ensemble members, significantly higher than chance, $t(33) = 28.21$, $p < 0.001$, 95% CI = [0.32, 0.37], Cohen' s d = 9.82, indicating that participants' attractiveness judgments were consistent with pre-ratings.

Using pre-rated scores, we calculated the mean attractiveness of all ensemble members in ensembles without average faces in Experiment 1 (M1 = 49.19). We then hypothetically calculated the mean member attractiveness assuming the ensemble synthesized an average face and included its attractiveness in the ensemble member average (M2 = 50.49). The difference between M1 and M2 indicated that synthesizing an average face increased the ensemble attractiveness average, $t(19) = 22.82$, $p < 0.001$, 95% CI = [1.14, 1.37], Cohen' s d = 10.47.

We analyzed the proportion of actual responses judging the probe average face as more attractive. Regardless of whether the ensemble contained the average face, the proportion judging the average face as more attractive (G1 without average face: 84.03%; G2 with average face: 83.55%) was significantly higher than chance (50%), $t(33) = 8.16$, $p < 0.001$, 95% CI = [0.25, 0.42], Cohen' s d = 2.84; $t(33) = 10.31$, $p < 0.001$, 95% CI = [0.27, 0.40], Cohen' s d = 3.59. When the probe stimulus was the average face, the proportion judging the average face as more attractive did not differ significantly between ensemble types (see Figure 3 [Figure 3: see original paper]), $t(33) = 0.11$, $p = 0.912$, 95% CI = [-0.10, 0.11], indicating that the presence or absence of an average face had no significant effect on perceptual discrimination of ensemble attractiveness.

(2) Diffusion Model Analysis Although the proportion analysis indicated no significant effect of average face presence on ensemble attractiveness discrimination, it remained unclear whether there were effects on decision processes (e.g., discrimination time, decision criteria). Here we used the hierarchical diffusion model (HDM; Vandekerckhove et al., 2011) for model fitting. HDM analysis advantages include accounting for individual differences between participants when calculating model parameters. Model upper and lower bounds were set as correct and incorrect responses, respectively. For probe stimuli that were average faces, judging the probe as more attractive was set as the correct response. Since no pre-existing response bias existed for correct versus incorrect responses, the model set the starting point (z) to a/2. Other parameters were

allowed to vary with experimental variables (ensemble type, probe stimulus type). Through HDM fitting, we obtained drift rate v, threshold separation a, and non-decision processing time t0 for each participant under each condition for statistical analysis (see Figure 4 [Figure 4: see original paper]).

The diffusion model fits each participant individually. Generally, if model fit parameter R-hat is less than 1.05 (Vehtari et al., 2019), the fit is considered good. One-sample t-tests on Experiment 1' s fit results revealed that all fit parameters (M = 1.00) were significantly less than 1.05, indicating good model fit.

Parameter-based t-tests showed no significant differences in information accumulation drift rate v or threshold separation a between ensembles with and without average faces, $t(33) = 0.48$, $p = 0.632$; $t(33) = 1.72$, $p = 0.096$, indicating that these did not affect discrimination between ensemble attractiveness and average face. However, non-decision processing time t0 was affected by whether the ensemble contained an average face, with shorter t0 required for ensembles containing the average face, $t(33) = 2.57$, $p = 0.015$, 95% CI = [0.01, 0.06], Cohen' s d = 0.90.

2.3 Discussion

Experiment 1' s results indicate that the group attractiveness effect indeed occurs for ensembles of 12 faces. This effect stems from the higher attractiveness of synthesized average faces and influences participants' comparison between ensemble attractiveness and probe average faces, making results no different from conditions that actually contained average faces. The pattern of results supports Hypothesis 2 of synthesized average stimuli and does not support Hypothesis 1 of mean value calculation. Therefore, average discrimination does not simply compute the mean of original ensemble members but rather forms an ensemble average face. Due to the high attractiveness of average faces, the mean of all members in ensembles with average faces is higher than in ensembles without average faces, being closer to the probe average face and thus making the probe and ensemble attractiveness less distinguishable. If average discrimination were completed through mean value calculation of ensemble members (i.e., member average as the comparison standard), then the proportion judging the average face as more attractive should be lower in conditions with average faces than without. However, results contradicted this prediction, showing no significant difference in proportions between ensemble types. Therefore, it is more likely that when perceiving ensemble stimuli, participants synthesize members into a new face (i.e., a highly attractive average face), significantly enhancing the attractiveness of ensembles without average faces. In ensembles containing the average face, the average face formed from the entire ensemble should be equivalent to the average face formed from the 11 original faces—namely, the average face already present in the ensemble. Thus, the appearance of the average face in the ensemble should not significantly increase ensemble attractiveness, and no difference should exist in proportions selecting the ensemble as more attractive

between conditions with and without average faces.

Diffusion model analysis results show that when no actual average face is presented, encoding and other non-decision processes take longer, indicating that forming an average face during average discrimination requires processing time and resources. This process is rapid (approximately 400 ms), thus having little impact on decision-making, resulting in no significant difference in decision information accumulation speed compared to when an actual average face is presented.

In summary, Experiment 1' s results demonstrate that ensemble attractiveness judgments involve forming an average face, leading to the group attractiveness effect. What factors might influence average face formation? Recent research has found that averaging of high-level features such as facial expression is a capacity-limited process (Ji et al., 2018) constrained by processing resources (Li et al., 2016). Is facial attractiveness averaging also affected by capacity? Previous results suggest yes. For example, Van Osch (2015) found that the group attractiveness effect rarely occurred in ensembles of 4-6 faces. Studies finding ensemble attractiveness equivalent to member averages also used small ensembles (Anderson, 1965; Anderson et al., 1973). Does this indicate that no average face is formed in small ensembles, and that small ensemble facial attractiveness judgments involve mechanisms different from large ensembles? Alternatively, average faces may form but be subject to interference. To separate these possibilities, Experiment 2 examined the relationship between small ensemble attractiveness and average faces.

Experiment 2

Experiment 2 used ensembles of four faces with the same experimental design and procedure as Experiment 1 to examine how capacity affects the relationship between ensemble attractiveness and average faces.

3.1 Method

(1) Participants Using G*Power with statistical power = 0.8, medium effect size f = 0.25, and repeated measures with 2 levels of the independent variable (ensemble type: ensemble without average face G1 vs. ensemble with average face G2), the minimum estimated sample size was N = 34. We recruited 35 students from Renmin University of China, excluding one who misremembered key assignments, leaving 34 valid participants (17 female) with a mean age of 20.68 years (SD = 2.27). All were right-handed with normal or corrected-to-normal vision.

(2) Experimental Materials The same as Experiment 1, but ensemble stimuli contained only four images. For ensemble stimulus types, an ensemble composed of four original faces represented the "ensemble without average face G1" condition. When three original faces composed the ensemble and the average face of ensemble members was added as a new member, this represented

the "ensemble with average face G2" condition. Probe stimulus types included ensemble average face, ensemble member faces, and non-member non-average faces, with the latter two serving as control stimuli.

Ensemble stimuli were presented in a 2$×$2 matrix. Image size was 8.19° × 9.43° of visual angle. Probe stimulus materials were single faces presented at the center of the screen, with image sizes identical to ensemble stimulus sizes.

(3) Experimental Design The same as Experiment 1. After completing the main task, each participant also rated the attractiveness of each image.

(4) Experimental Procedure The same as Experiment 1.

3.2 Results

(1) Button Response Results Based on pre-rated scores, response accuracy for probe stimuli that were new faces or ensemble members was 84.17%, significantly higher than chance, $t(33) = 16.84$, $p < 0.001$, 95% CI = [0.31, 0.39], Cohen' s d = 5.83, confirming that participants understood and performed the task correctly.

Similar to Experiment 1, we calculated the mean attractiveness of all ensemble members in ensembles without average faces when the probe was the average face (M1 = 47.82). We also hypothetically calculated the member mean assuming the ensemble generated an average face and included it in the ensemble member average (M2 = 49.73). The difference between M1 and M2 indicated that average faces also increased small-capacity ensemble attractiveness averages, $t(29) = 6.68$, $p < 0.001$, 95% CI = [1.44, 2.47], Cohen' s d = 2.48. Based on participants' post-task ratings, average face attractiveness (M = 55.18, SD = 11.02) was higher than the mean attractiveness of ensemble member faces (M = 51.71, SD = 11.76), $t(33) = 2.35$, $p = 0.020$, 95% CI = [0.51, 7.05], Cohen' s d = 0.820.

Comparing average face attractiveness between Experiment 2 and Experiment 1 when the average face was the probe revealed that average faces formed from small ensembles were less attractive (57.20 vs. 65.61), adjusted $t(41.7) = 100.61$, $p < 0.001$, 95% CI = [8.26, 8.60], Cohen' s d = 24.53. Cross-experiment comparison of the difference between average face (probe) and ensemble mean (9.51 vs. 16.43) showed that the difference between small ensemble average face and ensemble mean was smaller in Experiment 2, adjusted $t(53.8) = 112.13$, $p < 0.001$, 95% CI = [6.70, 6.94], Cohen' s d = 27.53. Consequently, the proportion judging the probe average face as more attractive should decrease in small ensembles. Results confirmed this: the proportion judging the average face as more attractive was significantly lower in Experiment 2 than Experiment 1 (66.57% vs. 83.79%), adjusted $t(63) = 3.37$, $p = 0.001$, 95% CI = [0.07, 0.27], Cohen' s d = 0.85.

Statistical tests showed that participants tended to judge average faces as more attractive than ensembles. The proportion judging the probe average face as

more attractive was significantly higher than chance, $t(33) = 4.60$, $p < 0.001$, 95% CI = [0.09, 0.24], Cohen' s d = 1.60. Moreover, for proportions judging the average face as more attractive, the condition without average faces was significantly higher than the condition with average faces, $t(33) = 3.77$, $p = 0.001$, 95% CI = [0.03, 0.12], Cohen' s d = 1.31 (see Figure 3).

Thus, small ensemble processing results indeed differed from large ensembles. To explore whether this difference arose from different processing mechanisms or interference with the average face, we tested whether an average face was formed in ensembles without average faces. We selected new face probe stimuli with attractiveness similar to average faces (means of 54.80 vs. 54.11, $t(46) = 0.18$, $p = 0.859$). However, the proportion judging probe attractiveness as higher when the probe was an average face (69.12%) remained higher than in the new face condition (52.01%), $t(33) = 4.84$, $p < 0.001$, 95% CI = [10.21%, 24.88%], Cohen' s d = 1.69. Even when we further selected new face stimuli with higher attractiveness than average faces (72.33 vs. 60.86, $t(42) = 3.85$, $p < 0.001$, 95% CI = [5.54%, 17.49%], Cohen' s d = 1.19), the proportion judging probe attractiveness as higher when the probe was an average face (71.01%) remained higher than in the new face condition (61.31%), $t(33) = 2.62$, $p = 0.013$, 95% CI = [2.24%, 17.13%], Cohen' s d = 0.91. This indicates that during discrimination, the average face did not appear as a new face but was more likely formed during ensemble presentation.

We also calculated accuracy rates Acc1 and Acc2 using actual ensemble member average M1 and hypothetical ensemble member average M2 (assuming average face formation) as comparison standards. Because the difference between average face ratings and ensemble means was large (mean difference = 10.40), creating ceiling effects, we only selected trials with differences below 10 to calculate Acc1 and Acc2. If an average face was formed, then ensemble attractiveness would increase, reducing the difference with the probe average face, thereby decreasing accuracy (Acc2 lower than actual average accuracy Acc1) and making it closer to accuracy for ensembles actually containing average faces (Acc0). t-tests showed no significant differences between Acc2 (50.34%) and Acc1 (53.03%), $t(33) = 1.18$, $p = 0.249$, Cohen' s d = 0.42, or between Acc2 and Acc0 (45.79%), $t(33) = 1.46$, $p = 0.154$, Cohen' s d = 0.51. Trend analysis with the three accuracy rates as three levels of a single factor revealed a linear decreasing trend from Acc1 to Acc2 to Acc0, $F(1, 33) = 4.21$, $p = 0.048$, $p^2 = 0.11$. These results indicate that although the average face effect detected in small ensembles without average faces could not fully equal the condition containing average faces, synthesized average faces still functioned to some extent.

(2) Diffusion Model Analysis Experiment 2 also used hierarchical diffusion model fitting to obtain drift rate v, threshold separation a, and non-decision processing time t0 for each participant under each condition for statistical analysis. All fit parameters (M = 1.00) were significantly less than 1.05, indicating good model fit.

Parameter-based t-tests showed that ensemble presence/absence of average face

did not affect threshold separation a or non-decision processing time t0 in discriminating ensemble attractiveness from average face, t(32) = -0.63, p = 0.533; t(32) = 0.72, p = 0.095. However, information accumulation drift rate v was affected by whether the ensemble contained an average face, with slower information accumulation in ensembles without average faces, t(33) = -4.775, p < 0.001, 95% CI = [-0.63, -0.25], Cohen' s d = 1.66 (see Figure 4).

Statistical results from hierarchical diffusion models differed between experiments (see Figure 4). For non-decision processing time t0, there was an interaction between ensemble presence/absence of average face and ensemble size, F(1, 63) = 14.03, p < 0.001, p^2 = 0.18. Simple effects analysis found that in ensembles containing average faces, Experiment 1' s t0 was significantly shorter than Experiment 2' s t0, t(63) = -2.568, p = 0.013, 95% CI = [-0.09, -0.01], Cohen' s d = 0.65. For drift rate v, there was also an interaction between ensemble presence/absence of average face and ensemble size, F(1, 63) = 9.63, p = 0.003, p^2 = 0.13. Simple effects analysis revealed that regardless of whether ensembles contained average faces, Experiment 1' s v was significantly greater than Experiment 2' s v, t(63) = 5.51, p < 0.001, 95% CI = [0.94, 2.00], Cohen' s d = 1.39; t(63) = 3.16, p = 0.002, 95% CI = [0.30, 1.34], Cohen' s d = 0.80.

Experiment 2 found that in conditions without average faces, the proportion judging the average face as more attractive was significantly higher than in conditions containing average faces, indicating that the presence of average faces in the ensemble significantly increased overall ensemble attractiveness. Thus, when ensemble size was four, participants' subjectively formed average face representation was suppressed or did not form. If average face representation did not form, then the average face as a probe should be similar to a new face. However, analyses showed that probe results for average faces and new faces were completely different. Therefore, small ensembles also formed average faces. These result patterns support Hypothesis 2 of synthesized average stimuli and do not support Hypothesis 1 of mean value calculation.

Although small ensembles formed average faces, the average discrimination response pattern differed from large ensembles in Experiment 1. This may be because average faces formed in small ensembles are more susceptible to interference, making ensemble attractiveness in ensembles containing average face stimuli closer to the average face, thereby reducing the proportion judging the average face as more attractive. This was also reflected in participants' response decision parameters: information accumulation speed was slower and required longer processing time compared to Experiment 1. Diffusion model results indicated that when actual average face input was present, discrimination decisions became easier, manifested as faster information accumulation. These results demonstrate that average faces were formed during average discrimination tasks but were subject to interference. This interference was also evident in the lower proportion judging average faces as more attractive in small ensembles compared to Experiment 1. Experimental results suggest two possible reasons: first, interference with small ensemble average faces (Hypothesis 3), which would

produce larger differences between ensemble attractiveness and average face in ensembles without average face stimuli, leading to higher proportions judging average faces as more attractive in small ensembles without average face stimuli; second, lower attractiveness of small ensemble average faces, creating smaller differences between ensemble and average face (Hypothesis 4).

Experiments 1 and 2 used relatively indirect average discrimination tasks. To provide more direct evidence, Experiments 3 and 4 employed rating tasks to further verify results from Experiments 1 and 2.

Experiment 3

Experiment 3 used large-capacity face ensembles in a rating task to provide more intuitive evidence for the relationship between ensemble attractiveness and average attractiveness under different capacities.

4.1 Method

(1) Participants Using G*Power with statistical power = 0.8, medium effect size f = 0.25, and single-factor 5-level design (rating type: mean of members in ensemble without average face M1, mean of members in ensemble without physical average face but including average face M2, ensemble without average face G1, ensemble with average face G2, average face Avg), the minimum estimated sample size was N = 21. We recruited 29 students from Renmin University of China, with 29 valid participants (15 female), mean age 22.14 years (SD = 3.17). All were right-handed with normal or corrected-to-normal vision.

(2) Experimental Materials The same as Experiment 1. Ensemble stimuli contained 12 faces presented in a 4$×$3 matrix. Single face images subtended 5.69° × 6.53° of visual angle. For rating types, an ensemble composed of 12 original faces represented the "ensemble without average face" condition. When 11 original faces composed the ensemble and the average face of ensemble members was added as a new member, this represented the "ensemble with average face" condition. Rating ensemble member faces and average faces separately constituted the "individual rating" condition.

(3) Experimental Design A single-factor 5-level within-subjects design was used (rating type: mean of members in ensemble without average face M1, mean of members in ensemble without physical average face but including average face M2, ensemble without average face G1, ensemble with average face G2, average face Avg). The dependent variable was participants' attractiveness ratings of target ensembles or faces.

(4) Experimental Procedure The experimental flow is shown in Figure 5 [Figure 5: see original paper]. A fixation point was presented for 500 ms, followed by a group of faces or a single face on screen. Participants rated the target'

s attractiveness on a 0-100 scale, where 0 represented lowest attractiveness and 100 represented highest attractiveness.

4.2 Results

Since the single-face rating task included both original faces and average faces from each ensemble, we used single-face ratings to calculate ensemble attractiveness averages. We computed the mean rating of ensemble members in ensembles without average faces (M1 = 47.31) and hypothetically calculated the member mean assuming the ensemble generated an average face (M2 = 48.78). We conducted ANOVA on M1, M2, ensemble attractiveness without average faces G1, ensemble attractiveness with average faces G2, and average face attractiveness Avg as five levels of rating type. Results showed a significant main effect of rating type, $F(4, 112) = 27.60$, $p < 0.001$, $p^2 = 0.50$. Multiple comparison results were as follows (see Figure 6 [Figure 6: see original paper]):

First, M2 was significantly greater than M1, $p < 0.001$, 95% CI = [1.22, 1.71], again confirming that synthesizing average faces enhances ensemble attractiveness averages. Second, ensemble attractiveness ratings without average faces G1 did not differ significantly from ensembles with average faces G2, $p = 0.532$, nor from M2, $p = 0.053$, but were greater than M1, $p = 0.011$, 95% CI = [1.26, 8.80]. Third, average face attractiveness was significantly higher than entire ensemble attractiveness G1, G2, and member averages M1, M2, all p' s ≤ 0.001. Trend analysis indicated an increasing trend from ensemble member averages to ensemble attractiveness to average face, $F(1, 28) = 62.82$, $p < 0.001$, $p^2 = 0.69$.

Additionally, we analyzed the difference between average face and ensemble attractiveness, finding no significant difference between ensembles with and without average faces, $t(28) = 0.19$, $p = 0.852$, again verifying Experiment 1' s finding that the proportion judging the probe average face as more attractive did not differ significantly between ensemble types.

Experiment 3' s rating task results essentially replicated Experiment 1' s findings. First, the rating task in large ensembles verified the group attractiveness effect, with ensemble attractiveness higher than the mean of member ratings. Second, average face attractiveness was greater than ensemble attractiveness, explaining why participants in Experiment 1 judged probe faces as more attractive. Third, whether ensembles contained average faces had no significant effect on ensemble ratings (supporting Hypothesis 2, not Hypothesis 1). Moreover, ensemble ratings without average stimuli G1 only approached ensemble member averages when average face formation was considered (i.e., M2). Finally, trend analysis and multiple comparisons showed that ensemble ratings without average stimuli were closer to results from conditions containing average faces. These results demonstrate that large ensembles indeed generate average faces.

Experiment 4

Experiment 4 used small-capacity face ensembles in a rating task to provide more intuitive evidence for the relationship between ensemble attractiveness and average attractiveness under different capacities.

5.1 Method

(1) Participants Using G*Power with statistical power = 0.8, medium effect size f = 0.25, and single-factor 5-level design, the minimum estimated sample size was N = 21. We recruited 31 students from Renmin University of China, excluding one with a rating range smaller than 10, leaving 30 valid participants (15 female), mean age 21.39 years (SD = 2.46). All were right-handed with normal or corrected-to-normal vision.

(2) Experimental Materials The same as Experiments 1 and 2. Ensemble stimuli contained four images presented in a 2$×$2 matrix. When rating ensemble average faces, stimuli were presented at the center of the screen, with single face images subtending 5.69° × 6.53° of visual angle. For rating types, an ensemble composed of four original faces represented the "ensemble without average face" condition. When three original faces composed the ensemble and the average face of ensemble members was added as a new member, this represented the "ensemble with average face" condition. Rating ensemble member faces and average faces separately constituted the "individual rating" condition.

(3) Experimental Design and Procedure The same as Experiment 3.

5.2 Results

Similar to Experiment 3, we used single-face ratings to calculate ensemble attractiveness averages. We computed the mean rating of ensemble members in ensembles without average faces (M1 = 47.87) and hypothetically calculated the member mean assuming the ensemble generated an average face (M2 = 50.32). We conducted ANOVA on M1, M2, ensemble attractiveness without average faces G1, ensemble attractiveness with average faces G2, and average face attractiveness Avg as five levels of rating type. Results showed (see Figure 6) a significant main effect of rating type, $F(4, 116) = 6.27$, $p < 0.001$, $p^2 = 0.18$. Multiple comparison results were as follows:

First, M2 was significantly greater than M1, $p < 0.001$, 95% CI = [1.82, 3.08], again confirming that synthesizing average faces enhances ensemble attractiveness averages. Second, ensemble attractiveness ratings without average faces G1 did not differ significantly from ensembles with average faces G2, $p = 0.110$, nor from M2, $p = 0.977$, nor from M1, $p = 0.504$. Third, average face attractiveness was significantly higher than entire ensemble attractiveness G1, G2, and member averages M1, M2, all p' s ≤ 0.007. Additionally, trend analysis indicated an increasing trend from ensemble member averages to ensemble attractiveness to average face, $F(1, 29) = 21.05$, $p < 0.001$, $p^2 = 0.42$.

Moreover, the difference between ensemble average face and overall ensemble attractiveness was larger in ensembles without average faces than in those with average faces (9.90 vs. 3.64), $t(29) = 6.40$, $p < 0.001$, 95% CI = [4.26, 8.26], Cohen' s d = 2.38, more intuitively confirming Experiment 2' s finding that the proportion judging probe attractiveness as higher decreased when average faces were included.

Comparing average face attractiveness between Experiment 4 and Experiment 3 showed that average faces formed from small ensembles were less attractive (60.11 vs. 65.24), $p = 0.004$, 95% CI = [2.76, 13.85], Cohen' s d = 0.94. Cross-experiment comparison of differences between average face and ensemble mean (6.76 vs. 12.55) revealed a possible trend: the difference between small ensemble average face and ensemble member average was smaller in Experiment 4, adjusted $t(35.649) = 1.72$, $p = 0.094$, 95% CI = [-1.06, 12.81], Cohen' s d = 0.07.

Experiment 4' s rating task results provided direct support for Experiment 2' s findings. First, in small ensembles, the group attractiveness effect was weakened, with no significant differences between ensemble attractiveness and ensemble member rating averages M1 and M2. Combined with Experiment 3' s results, this verifies previous conclusions: the group attractiveness effect is strong in large ensembles but weak in small ensembles. This supports the hypothesis that average faces are more susceptible to interference in small ensembles. Additionally, small ensemble average face attractiveness indeed decreased, with smaller differences from ensemble means, suggesting that relatively lower average face attractiveness in small ensembles is another possible reason for the reduced group attractiveness effect.

Second, similar to Experiment 3, whether ensembles contained average faces had no significant effect on ensemble ratings (supporting Hypothesis 2, not Hypothesis 1). Moreover, ensemble ratings without average stimuli G1 also did not differ from ensemble member averages M2 that included average faces. Trend analysis and multiple comparisons showed that ensemble ratings without average stimuli were closer to results from conditions containing average faces. These results indicate that small ensembles may also be influenced by generated average faces. Furthermore, the difference between ensemble attractiveness and average face attractiveness was larger in ensembles without average face stimuli (Hypothesis 3), reflecting that ensemble attractiveness containing average face stimuli was closer to average face attractiveness. This more intuitively demonstrates that in smaller ensembles, participants' subjectively formed average face representations were suppressed (combined with Experiment 2 results), and explains why the proportion judging probe attractiveness as higher decreased when average faces were included in Experiment 2.

The button response and diffusion model fitting results from Experiments 1 and 2, together with the rating results from Experiments 3 and 4, collectively demonstrate that when ensemble capacity is 12 faces, whether a highly attractive average face appears does not affect ensemble attractiveness ratings or average discrimination tasks, indicating that ensemble average faces are formed during representation. When ensemble capacity is four faces, the average face effect is weakened, possibly due to lower attractiveness of ensemble average faces and interference from individual representations.

6.1 The Group Attractiveness Effect in Face Ensembles

In Van Osch et al.' s (2015) study, using naturalistic materials (such as party photos) of several women, they found that in larger ensembles (photos with more people), ensemble attractiveness was higher than the average attractiveness of members.

Experiment 1' s average discrimination task found that when ensembles did not contain average faces, the proportion judging probe attractiveness as higher was similar to when ensembles contained average faces, indicating that average attractiveness was comparable across conditions and that ensemble attractiveness without average faces was higher than the mean of member faces, reflecting the group attractiveness effect from another perspective. Experiment 3 directly replicated the group attractiveness effect in large ensembles.

Experiment 2 also showed the group attractiveness effect, though less pronounced than Experiment 1. Similarly, Van Osch et al. (2015) rarely observed ensemble attractiveness exceeding member attractiveness averages in smaller ensembles. The present study provides both direct evidence for the group attractiveness effect through ratings and partial evidence inferred from the role of average faces in ensembles. Van Osch et al. (2015) used naturalistic materials with high ecological validity but had issues such as concentrated member attractiveness distributions and lack of representative high/low attractiveness faces. Our study improved upon this by using background-free ensembles composed of individually rated face stimuli with balanced numbers of different attractiveness levels, still obtaining similar conclusions. This indicates that the group attractiveness effect is relatively stable in large ensembles and that the phenomenon' s emergence is related to ensemble average face formation.

6.2 Mechanism of Average Representation Formation

By comparing and fitting actual average discrimination responses with response distributions predicted by different theoretical hypotheses, we found that the hypothesis assuming ensemble attractiveness includes average face contributions

fit better, supporting the view that average stimuli (such as average faces) are formed during average representation. In Experiment 1, when ensembles contained average faces, the proportion judging average faces as more attractive did not decrease, indicating that ensemble attractiveness with and without average face stimuli was equally close to average face attractiveness. Therefore, average representation is not obtained through mean value calculation of ensemble members. Similarly, both Experiment 3 and Experiment 4 found that average representation attractiveness was much higher than ensemble member averages. In Experiment 2' s small ensembles, different results for new faces versus average faces as probes indicated that average faces were processed before appearing as probes. This conclusion aligns with Ying et al. (2020), who found through an attractiveness adaptation aftereffect paradigm that adaptation aftereffects induced by a group of faces equaled those induced by the group' s average face, also supporting the formation of average stimuli in ensemble representation.

Notably, average stimulus formation also requires resource investment, as evidenced by longer encoding processing time when no average face was presented in Experiment 1. Huang (2015) found that priming effects were equal for object features and statistical representations, suggesting that statistical representation formation requires at least the same amount of attentional resources as individual object processing. Therefore, average representation and individual representation may have a competitive relationship in early stages (Li et al., 2016). Because there are insufficient cognitive resources to process all individuals in early stages, average representation is formed first. Bauer (2017), using line sets as ensemble stimuli and adding a digit memory task before average discrimination, found that high memory load conditions (4-7 random digits) were more conducive to forming average representations than low load conditions (a single 0). Cross-experiment comparisons between Experiments 1 and 2 also found that overall information accumulation speed was slower in small ensembles than in large ensembles.

Meanwhile, small ensembles also formed average representations requiring resource investment, as shown by slower information accumulation when no average face was presented in Experiment 2. That is, average stimulus formation occurs in both small and large ensembles. Differences in responses between ensemble sizes result from other factors rather than different processing mechanisms.

6.3 Relationship Between Ensemble Attractiveness and Average Face

Although ensemble attractiveness representation includes average face contributions, ensemble attractiveness is not completely equivalent to average face attractiveness. Both Experiments 1 and 2 found that participants tended to rate average faces as more attractive than ensembles. Moreover, average face ratings in Experiments 3 and 4 were also higher than ensemble ratings, indicating that ensemble attractiveness is based on average face formation, incorporating the average stimulus as an ensemble member before evaluating the ensemble as a

whole. Therefore, the group attractiveness effect does not entirely depend solely on average faces as Van Osch et al. (2015) inferred, but rather involves reading the average face representation, similar to expression ensemble processing (Haberman & Whitney, 2009).

A similar phenomenon is that participants often misidentify average representations as ensemble members in recognition tasks or perceive average faces as having the same identity as one member (Neumann et al., 2013). During judgment, participants may use a strategy: when processing entire ensemble attractiveness, they perceive the average face as a particular ensemble member. Since most other faces in the ensemble have lower attractiveness than the average face, they tend to judge the separately presented average face as more attractive.

Researchers have found that when multiple faces are present, individual faces are judged as more attractive than when presented in isolation (Walker & Vul, 2014), explained similarly as an averaging effect. Background faces bias face perception toward the group average. Therefore, the higher the attractiveness of background faces accompanying a target face, the higher the target face is rated (DeBruine et al., 2007; Perrett et al., 1994; Walker & Vul, 2014). Thus, average face formation may influence other ensemble members. Because average faces are highly attractive, other ensemble members' attractiveness is also elevated. This is one path through which average faces affect other faces and may also contribute to the group attractiveness effect.

6.4 Relationship Between Average Stimulus Representation and Ensemble Size

In small ensembles, facial attractiveness is generally rated as equal to the ensemble member average (Anderson, 1965; Anderson et al., 1973), and the group attractiveness effect disappears in small ensembles (Van Osch et al., 2015), indicating that average representation in small ensembles has specific characteristics.

On one hand, average stimuli generated by small ensembles have relatively low attractiveness, possibly resulting in smaller enhancement effects on ensemble attractiveness and thus reducing the group attractiveness effect. On the other hand, average representation may be subject to interference. Reverse Hierarchy Theory proposes (Hochstein et al., 2015) that holistic representations in high-level cortex return to local processing in a top-down manner, using local detail information to confirm (or correct) initial holistic representation estimates. This means average representation is corrected by individual representation in later processing stages. Li et al. (2016) also found that when presentation time is extended and cognitive resources increase, individual representation precision is enhanced.

Both theories received support from our results. In Experiment 2, the lower proportion judging average faces as more attractive, and Experiment 4' s comparison of differences between average face and ensemble attractiveness showing

smaller differences in small ensembles containing average faces, both indicate that lower average face attractiveness in small ensembles influences the effect. Meanwhile, Experiment 2' s lower proportion judging average faces as more attractive when average face stimuli were included indicates that direct input of average faces reduced the difference between ensemble attractiveness and average face compared to when participants generated average faces themselves, supporting the hypothesis that average faces in small ensembles were subject to interference. The slower information accumulation speed in small ensembles regardless of average face presence also supports Reverse Hierarchy Theory's claim that local detail information corrects and interferes with holistic representation.

7 Conclusion

(1) The group attractiveness effect in face ensembles is based on average face stimulus formation. (2) Average representation generation is based on average stimulus production. (3) Average stimuli are formed in both small and large ensembles. The absence of group attractiveness effect in small ensembles is due to interference with average stimuli and their relatively low attractiveness. (4) Average representation processing also requires certain cognitive resources.

References

Abbas, Z. A., & Duchaine, B. (2008). The role of holistic processing in judgments of facial attractiveness. *Perception*, 37, 1187–1196.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4), 392–398.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3), 122–131.

Anderson, N. H. (1965). Averaging versus adding as a stimulus combination rule in impression formation. *Journal of Experimental Psychology*, 70, 394–400.

Anderson, N. H., Lindner, R., & Lopes, L. L. (1973). Integration theory applied to judgments of group attractiveness. *Journal of Personality and Social Psychology*, 26, 400–408.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162.

Bauer, B. (2009). Does Stevens' s power law for brightness extend to perceptual brightness averaging? *Psychological Record*, 59(2), 171–185.

Bauer, B. (2017). Perceptual averaging of line length: Effects of concurrent digit memory load. *Attention, Perception & Psychophysics*, 79(8), 2510–2522.

Brady, T. F., & Alvarez, G. A. (2015). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working

memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 921–929.

Carragher, D. J., Lawrence, B. J., Thomas, N. A., & Nicholls, M. E. R. (2018). Visuospatial asymmetries do not modulate the cheerleader effect. *Scientific Reports*, 8(1), 2548.

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, 70(5), 789–794.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17), R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology*, 35, 718–734.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), *From Perception to Consciousness: Searching with Anne Treisman* (pp. 339–349). Oxford University Press.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, 144(2), 432–446.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 791–804.

Hochstein, S., Pavlovskaya, M., Bonneh, Y. S., & Soroker, N. (2015). Global statistics are not neglected. *Journal of Vision*, 15(4), 7.

Huang, L. (2015). Statistical properties demand as much attention as object features. *PLOS ONE*, 10(8), e0131191.

Ji, L., Chen, W., Loeys, T., & Pourtois, G. (2018). Ensemble representation for multiple facial expressions: Evidence for a capacity-limited perceptual process. *Journal of Vision*, 18(3), 17, 1–19.

Komori, M., Kawamura, S., & Ishihara, S. (2009). Averageness or symmetry: Which is more important for facial attractiveness? *Acta Psychologica*, 131, 136–142.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, 1(2), 115–121.

Li, H., Ji, L., Tong, K., Ren, N., Chen, W., Liu, C. H., & Fu, X. (2016). Processing of individual items during ensemble coding of facial expressions. *Frontiers in Psychology*, 7, 1332.

Luo, A. X., & Zhou, G. (2018). Ensemble perception of facial attractiveness. *Journal of Vision*, 18(8), 7, 1–19.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, 15(4).

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics*, 70(5), 772–788.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, 128(1), 56–63.

O' Toole, A. J., Price, T., Vetter, T., Bartlett, J. C., & Blanz, V. (1999). 3D shape and 2D surface textures of human faces: The role of "averages" in attractiveness and age. *Image and Vision Computing*, 18, 9–19.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4(7), 739–744.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922.

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological Science*, 7, 105–110.

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in non-western cultures: In search of biologically based standards of beauty. *Perception*, 30, 611–625.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16(1), 44–62.

Van Osch, Y., Blanken, I., Meijs, M. H. J., & Van Wolferen, J. (2015). A group' s physical attractiveness is greater than the average attractiveness of its members: The group attractiveness effect. *Personality and Social Psychology Bulletin*, 41(4), 559–574.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. (2019). Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC. *Bayesian Analysis*, Advance publication (2021), 28 pages.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, 60(6), 385–402.

Walker, D., & Vul, E. (2014). Hierarchical encoding makes individuals in a group seem more attractive. *Psychological Science*, 25(1), 230–235.

Wang, Y., & Luo, Y. J. (2005). Standardization and assessment of college students' facial expression of emotion. *Chinese Journal of Clinical Psychology*, 13(4), 396–398.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, 69(1), 105–129.

Willis, R. H. (1960). Stimulus pooling and social perception. *Journal of Abnormal and Social Psychology*, 60, 365–373.

Ying, H., Burns, E., Choo, A. M., & Xu, H. (2020). Temporal and spatial ensemble statistics are formed by distinct mechanisms. *Cognition*, 195.

Author Contributions: Wenfeng Chen: Conceptualized research questions and framework, designed research protocol, analyzed data, revised final manuscript; Xinran Tian, Wenxia Hou: Conducted experiments, analyzed data, drafted manuscript; Yuxiao Ou, Bing Yi: Collected data; Junchen Shang: Revised final manuscript.

Note: Figure translations are in progress. See original paper for figures.