# Isolated Configuration: Holistic Processing Beyond the Realm of Faces

Yuxuan Zeng[1] (zeng.774@osu.edu) Zitong Lu[1] Ren E Hentz[1] David E Osher[1]

The Ohio State University, Department of Psychology

## Abstract

Holistic processing has traditionally been regarded as unique to faces, attributed to the spatial layout across facial features and their variability across individuals. However, whether such processing can occur in stimuli other than faces has remained an open and contentious question. Here, we show that novel non-face stimuli, which are designed to be differentiated strictly by configural properties, can drive holistic processing effects, while similar non-configural stimuli do not generate these effects. Using novel, abstract stimuli that isolate configural information, we systematically examined four effects typically associated with holistic face perception: the inversion, part-whole, composite, and misalignment effects. We observed all four effects with these stimuli, suggesting that this special form of visual processing extends beyond the realm of faces, and primarily requires that stimuli contain configural information for accurate classification. These effects manifested with minimal training, demonstrating that the visual system can rapidly integrate spatial relationships. Furthermore, convolutional neural networks, trained to categorize these stimuli, also show these effects, and like humans, these effects are only observed with the configural stimuli, and not with the non-configural stimuli. Our results challenge the traditional view of holistic processing as face-specific, and propose that configural processing itself serves as a primary driver of holistic perception and may be a fundamental visual process of its own.

## Introduction

In visual perception, configuration refers to the spatial relationships between features within a stimulus (Tanaka & Sengco, 1997). The concept of configuration has predominantly been examined within the context of face recognition. Since Galton (1879) posited that face perception transcends the mere aggregation of individual features, the role of configuration has been a focal point of extensive discussion. This sort of holistic processing is often regarded as a distinctive behavioral marker of face perception (Jiang et al., 2011; Yovel & Kanwisher, 2004, 2005).

Despite its well-established role in face recognition, the exclusive nature of holistic processing for faces has been hotly debated (Freiwald et al., 2016; Piepers & Robbins, 2012). Prior research suggests that it can also develop through extensive experience with other complex objects like birds or cars (Gauthier et al., 2000), or even novel objects like greebles (Gauthier & Tarr, 1997). However, two main concerns remain unresolved: First, it has been argued that humans might automatically recruit face-specific strategies with these stimuli due to their animate and face-like appearance (Brants et al., 2011; Y. Xu, 2005; Y. Xu et al., 2005). This would suggest that

participants are relying on face-like processing mechanisms, rather than genuine configural processing (Bilalić et al., 2011; Bilalić et al., 2016; Gauthier et al., 2000; Wong et al., 2009). Second, many non-face configuration studies may unintentionally rely on feature-based recognition instead of true configural processing. For example, for a bird expert, identifying a bird species may depend on specific color patches or distinctive markings, rather than the spatial relationships among features. This reliance on features makes it unclear whether these findings genuinely reflect configural processing or merely a byproduct of salient feature detection.

To directly address this, we created two sets of stimuli: (1) *configuration-only stimuli* that contain identical features and their recognition depends strictly on the configuration of those features (Fig. 1a), and (2) *feature-only stimuli* that are recognized by their features and whose configuration is completely nondiagnostic (Fig. 1b). By comparing responses to these two stimulus sets, we can isolate the impact of configuration, allowing us to answer the core question: do novice observers holistically process non-face stimuli that are purely configural by design?

To comprehensively test the role of configuration in non-face recognition, we adapted three well-established tasks commonly used to measure holistic face perception: the inversion, part-whole, and composite tasks (Fig. 2a). These tasks have been widely used to illustrate how face processing relies on unique configuration properties. For example, although *inversion* generally increases the difficulty of recognizing objects, the disproportionate increase in difficulty for face identification (Yin, 1969) was initially attributed to a "special factor involving only the faces." Later, it was found that recognizing the *whole* (e.g., the face of Brad Pitt) does not necessarily imply recognizing its *parts* (e.g., Brad Pitt's eyes), suggesting that the processing of configuration information is more crucial than just feature information (Tanaka & Farah, 1993; Tanaka & Sengco, 1997). The *composite task* provides further evidence: merging parts of different stimuli (e.g., the top half of one stimulus and the bottom half of another) creates difficulty in recognizing the old parts due to the integration of the old configuration into the new one (Young et al., 1987). We employed these tasks to test whether similar holistic effects can occur using our abstract, non-face stimuli (Fig. 2b).

Additionally, the *misalignment* task, typically used as a control in composite effect studies (Young et al., 1987), is also noteworthy. Misalignment changes the configuration of an object without affecting its feature information, making it an effective method for detecting configuration information, although its isolated effect has not been previously tested in face recognition. By applying the misalignment task to our stimuli, we aim to further confirm the presence of holistic processing in non-face configurations.

We hypothesize that configuration-only stimuli will produce effects akin to faces for each of these tests of holistic processing, while the feature-only stimuli will not, in novice participants who learn to recognize these stimuli during a brief session (roughly 30 minutes total for learning both configuration and feature-only stimuli). We also expect that a deep neural network model,

trained to recognize these stimuli, will demonstrate the same findings. These results would demonstrate that holistic processing effects are not specific to faces, nor expertise, but are rather a byproduct of recognizing stimuli that contain significant configural information. Note that this does not imply that faces are not special stimuli, as faces indeed contain a substantial amount of configural information, which is arguably very rare among the visual objects that humans encounter in the natural world.

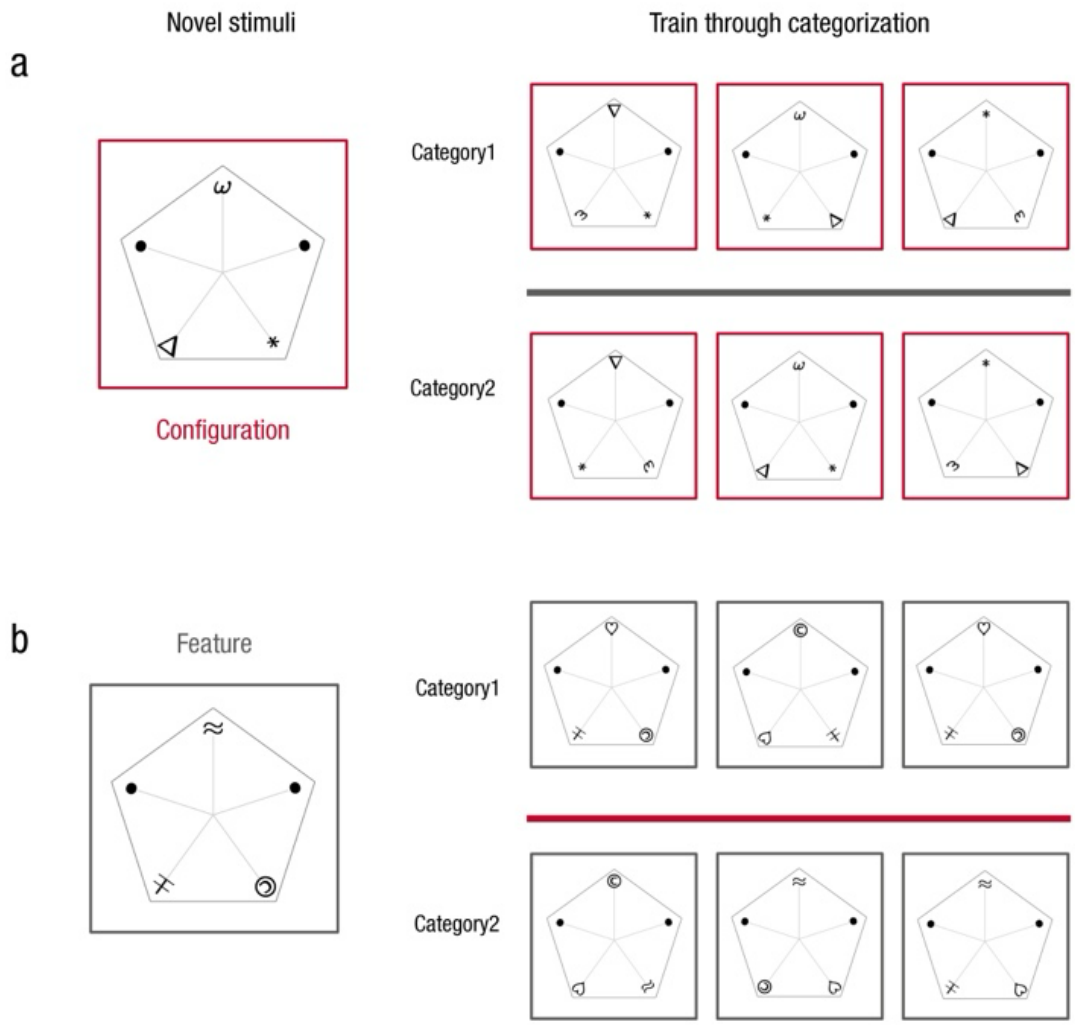



**Figure 1.** Novel stimuli used in the experiment: Configuration vs. Feature conditions.

(a) Configuration Condition (red): Left panel shows an example stimulus, where three symbols ("∇", "*", "ω") are arranged within a pentagon. Right panel displays all six exemplars, grouped into two categories based on the spatial arrangement of these symbols. The position of each symbol is balanced across categories, ensuring that no single symbol or its location determines category membership. Categorization relies on the overall configuration rather than individual symbols.

(b) Feature Condition (gray): Left panel shows an example stimulus, where each stimulus contains three symbols chosen from a set of four ("©", "±", "♡", "≈"). Right panel presents a subset of six exemplars (out of 24 total). Categorization depends on the presence of specific symbol pairs—"©" and "±" indicate Category 1, while "♡" and "≈" indicate Category 2. The third symbol in each stimulus is a distractor, randomly drawn from the opposite category. Spatial arrangement is irrelevant for categorization.

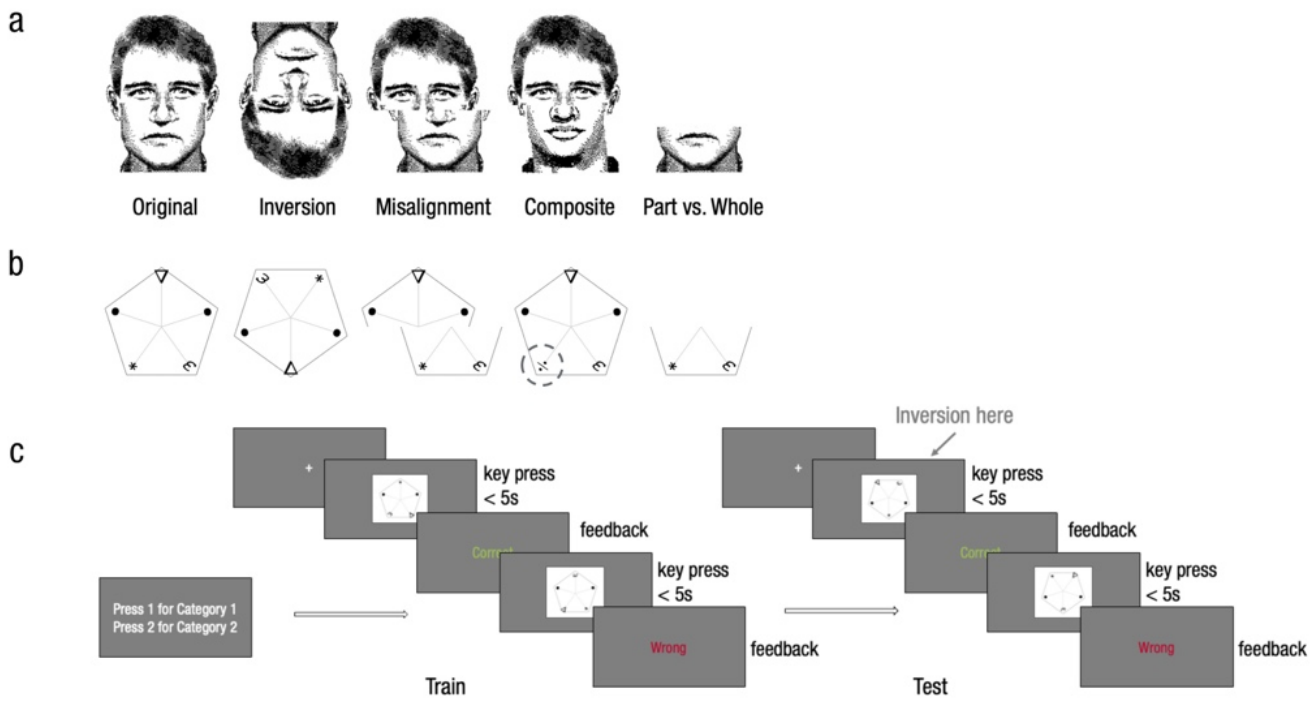



**Figure 2.** Experimental design and task structure.

(a) Holistic perception tasks: Commonly used tasks in face perception research—Inversion, Misalignment, Composite, and Part vs. Whole—illustrate how disrupting facial configuration affects recognition. These tasks were adapted for non-face stimuli in this study.

(b) Adaptation to novel stimuli: The same task principles were applied to our abstract stimuli, testing whether holistic processing effects extend beyond face recognition.

(c) Trial structure: Participants first completed a training phase (300 trials), categorizing stimuli by pressing "1" for Category 1 or "2" for Category 2. Each stimulus was presented for up to 5 seconds or until a response, followed by a 500ms intertrial interval with feedback (correct/incorrect). In the test phase (30 trials per task), participants categorized altered stimuli (e.g., inverted stimuli) using the same learned rule. They were informed of the transformation

(e.g., "These stimuli are inverted") before beginning the test phase, but categorization relied on previously learned configuration.

## Experiment 1: Inversion Effect

Configural processing, typically studied within the context of face recognition, is examined here using abstract, non-face stimuli to determine whether such processing can occur independently of facial features.

### Methods

**Participants**: Forty-five undergraduates from The Ohio State University participated in this experiment. They were recruited through the Research Experience Program. Only those achieving an 80% accuracy rate in the preliminary training were included in the analysis, leaving 20 participants in the final sample (individualized participants' learning performance see Supplemental Fig. 2a).

**Stimuli**: Stimuli were created using MATLAB and displayed on a standard 23.8-inch Dell monitor with a resolution of 1920x1080 pixels. Each stimulus consisted of an enclosed pentagon with five symbols (1167x875 pixels, see Fig. 1). Two fixed symbols ("●") served as placeholders for positional uniformity across conditions, while the remaining three symbols determined group classification. No symbols were shared between the configural and featural conditions to ensure distinct grouping criteria:

- **Configural Condition**: Three symbols "∇", "*", "ω" used in all stimuli (Feature number determined by a pilot study, see Supplemental Fig. 1). Six unique stimuli were created by varying the spatial arrangement of these symbols within the pentagon. Grouping based on the relative spatial location of "∇", "*", "ω". We balanced each symbol in each position for their grouping, such that no single symbol's presence or position could reliably indicate group membership (Fig. 1a). In other words, the only way for participants to be able to accurately classify these stimuli is to learn the spatial relationships among symbols.
- **Featural Condition**: These stimuli were arranged in the identical pentagonal arrangement as the configuration-only stimuli. Each stimulus consisted of three symbols selected from four symbols "©", "±", "♡", "≈", and arranged in distinct spatial configurations, resulting in 24 unique combinations (4 possible symbol combinations × 6 spatial arrangements). Spatial arrangement is irrelevant for grouping. Grouping based on specific symbol pairs. "©" and "±" showing simultaneously indicated Group 1, "♡" and "≈" showing simultaneously indicated Group 2. The 3rd symbol is a distractor randomly choose from the opposite group (Fig. 1b). In other words, the only way for participants to

be able to accurately classify these stimuli is to learn which features must be present in each group.

**Design and Procedure**: The study employed a within-subject design, with participants completing both the configural and featural condition tasks in a randomized order. Within each condition, participants first completed a training phase with 300 trials to learn the categorization rules through feedback learning. In each trial (Fig 2c), participants were shown a single stimulus and were instructed to decide whether they belong to "group 1" or "group 2" based on the trial and error, responding via keyboard press. Stimuli were presented on the screen for up to 5 seconds or until a key response was made. If they responded within the time limit, feedback ("correct" or "wrong") appeared immediately; if no response was made, "no response" was displayed. A fixation cross ("+") was shown for 0.5 seconds between trials.

After training, participants proceeded to the testing phase, which was identical to the training task except the stimuli were rotated 180 degrees (Fig 2b), and this phase lasted 30 trials. Participants were instructed to perform the task as before, and that the stimuli would be inverted (noting that they are rotated and not flipped). Accuracy and reaction times from the final 30 trials of the training phase served as baseline measures to compare against the results of the testing phase.

**Results**

Upon completing the 300 training trials, participants' performance in the configural and featural conditions was consistently high, with average accuracy rates for the last 30 trials of 96.0% (SE = 1.1%) and 95.7% (SE = 1.0%), respectively. A paired t-test confirmed no significant difference in accuracy between the configural and featural conditions, $t(38) = 0.22$, $p = 0.826$. Similarly, reaction times (RT) showed no significant difference between the two conditions, $t(38) = -0.35$, $p = 0.731$. Thus, any testing effects are unlikely to be due to differences in task difficulty.

In the testing phase, the RT for inverted stimuli in the configural condition significantly increased from the baseline, $t(19) = 4.07$, $p < 0.001$, indicating a pronounced inversion effect. This elongation in RT demonstrates the sensitivity of orientation changes for the configural stimuli. In contrast, the featural condition did not show a significant change in RT, $t(19) = -0.67$, $p = 0.513$, suggesting that featural processing is less affected by orientation (Fig. 3c).

Accuracy followed a pattern similar to RT; in the configural condition, accuracy significantly decreased for inverted stimuli compared to upright stimuli, $t(19) = -2.78$, $p = 0.013$, highlighting the disruption caused by inversion to configural recognition. Conversely, in the featural condition, the accuracy difference between inverted and upright stimuli was not significant, $t(19) = -0.89$, $p = 0.385$, reinforcing the notion that featural processing maintains stability across orientations (Fig. 3d).

An independent samples t-test assessing the differential impacts of inversion on configural versus featural processing revealed significant differences in both RT and accuracy, $t(38) = 3.87$, $p < .001$ and $t(38) = 2.38$, $p = 0\ .024$, respectively (Fig. 3c and 3d).

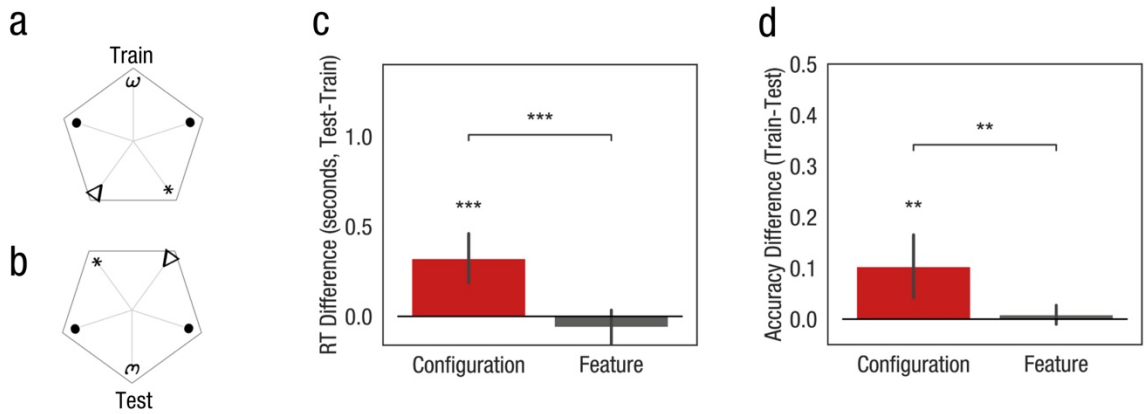



**Figure 3.** Inversion effect on reaction time and accuracy.

(a, b) Stimulus example: Illustration of a training stimulus (a) and its corresponding inverted version in the test phase (b).

(c) RT difference: Mean RT difference (test phase – last 30 trials of training) for the configural and featural conditions. RT significantly increased from baseline in the configural condition ($p < 0.001$), while no significant change was observed in the featural condition (*n.s.*). The RT difference between the two conditions was also significant ($p < 0.001$).

(d) Accuracy difference: Mean accuracy difference (last 30 trials of training – test phase). Accuracy significantly decreased in the configural condition ($p < 0.01$), while no significant change was observed in the featural condition (*n.s.*). The inversion effect was significantly larger for the configural than the featural condition ($p < 0.01$).

Error bars represent standard errors of the mean (SEM). Statistical significance: n.s. (not significant), ** $p < 0.01$, *** $p < 0.001$.

## Experiment 2: Part-Whole Effect

Experiment 2 builds on the initial demonstration of configural processing in non-facial objects from Experiment 1, focusing on the part-whole effect to explore further the capabilities of configural processing.

**Methods**

**Participants**: A new group of forty undergraduate students from the Ohio State University participated in this study, recruited through the Research Experience Program. The inclusion criteria required achieving at least 80% accuracy in the final 30 of 300 training trials, leaving 20 participants in the final sample (individualized participants' learning performance see Supplemental Fig. 2b).

**Stimuli and Procedure**: The stimuli and procedure for Experiment 2 mirrored those of Experiment 1, with a key variation involving the use of only the bottom half of each stimulus in the test phase. During the test phase, for the configural condition, the top half was cropped, leaving only the bottom half containing the remaining two of the three original symbols ("∇", "*", "ω", see Fig. 4b). Categorization is based on the configural relationship of these symbols, and all combinations of two symbols in the bottom half provide sufficient information to differentiate between categories. For the featural condition, the bottom half of each stimulus was presented, such that they still contained the symbols diagnostic of group membership. As in Experiment 1, participants underwent 300 training trials to familiarize themselves with the stimuli based on configural or featural properties, followed by 30 testing trials where the parts of the stimuli were categorized using the same rules as during the training phase. Participants were instructed to perform the task as before, and that the stimuli would be cropped leaving 2 of the original symbols.

**Results**

Upon completion of the 300 training trials, all twenty participants exceeded the 80% accuracy threshold. The mean accuracy rate was 94.8% (SE = 1.1%) in the configural condition and 97.0% (SE = 1.1%) in the featural condition. A t-test confirmed no significant difference in accuracy rates between the configural and featural conditions, $t(38) = -1.38, p = 0.177$. Similarly, RT showed no significant difference between the two conditions, $t(38) = -0.47, p = 0.643$.

In the testing phase, significant differences emerged between part and whole stimuli, particularly in the configural condition. Reaction times for part stimuli were significantly longer than for whole stimuli, $t(19) = 4.44, p < 0.001$. This increase in RT suggests that participants engaged in more holistic processing, which is challenged when only part of the configuration is visible. Conversely, in the featural condition, we observed a significant reduction in RT for part stimuli compared to whole stimuli, $t(19) = -7.26$, $p < 0.0001$, suggesting that isolating features simplifies the task due to reduced complexity and a narrower focus on specific elements (Treisman & Gelade, 1980) (Fig. 4c).

Accuracy in the configural condition dropped significantly for part stimuli compared to whole configurations, $t(19) = -2.40$, $p = 0.027$, further corroborating the sensitivity of configural processing to disruptions in holistic perception. Meanwhile, accuracy in the featural condition showed no significant change, $t(19) = 1.00$, $p = 0.333$, possibly due to ceiling effects (Fig. 4d).

Furthermore, the configural and featural tasks significantly differed, both in RT, t(38) = 8.44, p<0.0001, and accuracy, t(38) = 2.59, p =0.013 (Fig. 4c and 4d).

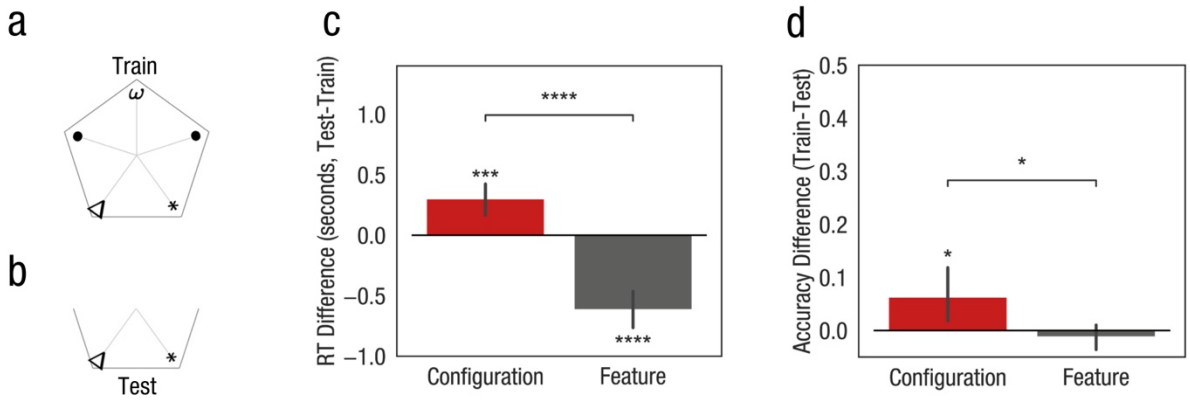



**Figure 4.** Part-whole effect on reaction time and accuracy.

(a, b) Stimulus example: Illustration of a training stimulus (a) and its corresponding part-based version in the test phase (b).

(c) Reaction time (RT) difference: Mean RT difference (test phase – last 30 trials of training) for the configural and featural conditions. RT significantly increased from baseline in the configural condition ($p < 0.001$), while it significantly decreased in the featural condition ($p < 0.0001$). The RT difference between the two conditions was also significant ($p < 0.0001$).

(d) Accuracy difference: Mean accuracy difference (last 30 trials of training – test phase). Accuracy significantly decreased in the configural condition ($p < 0.05$), while no significant change was observed in the featural condition (*n.s.*). The accuracy difference between the two conditions was also significant ($p < 0.05$).

Error bars represent standard errors of the mean (SEM). Statistical significance: *n.s.* (not significant), * $p < 0.05$, *** $p < 0.001$, **** $p < 0.0001$.

## Experiment 3: Composite Effect

Experiment 3 delves into the composite effect, a phenomenon where modifications to a part of a configuration (e.g. the top half of a face) lead to the perception of an overall holistic change, despite the remaining parts being unchanged.

**Methods**

The same 20 participants from Experiment 2 participated in this experiment. In the both conditions, one of the three symbols was randomly replaced with a new symbol not seen during the training phase (Fig. 5b). The procedure followed the same general structure as in previous experiments, and the testing phase consisted of 30 trials. Participants were instructed to perform the task as before, and that one of the symbols would be switched with a new one but use the remaining 2 for categorization.

**Results**

In the testing phase, RT for configural tasks showed a notable increase from the training phase, $t(19) = 11.38$, $p < 0.0001$, indicating substantial processing challenges with the composite version of the configuration-only stimuli (Fig. 5c). Conversely, RT in the featural tasks decreased, $t(19) = -2.73$, $p = 0.013$, similar to the parts task in Experiment 2.

We also observed a significant drop in accuracy for the configural task, $t(19) = -2.65$, $p = 0.016$, suggesting that the composite versions of the configural stimuli were more difficult to recognize. This was not statistically significant for the featural task, $t(19) = 0.35$, $p = 0.726$ (Fig. 5d).

The effects were significantly different between tasks as well, for both RT ($t(38) = 8.44$, $p < 0.0001$), and accuracy ($t(38) = 2.65$, $p = 0.012$; Fig. 5c and 5d).

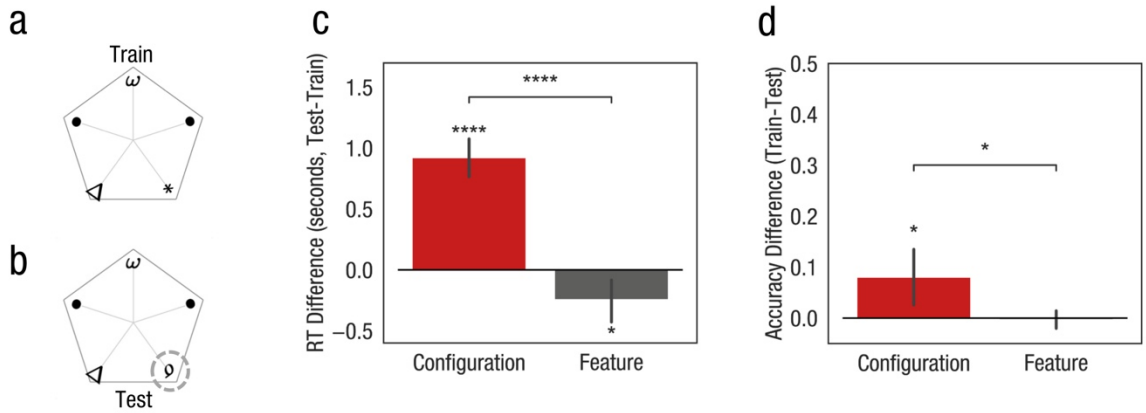



**Figure 5**. Composite effect on reaction time and accuracy.

(a, b) Stimulus example: Illustration of a training stimulus (a) and its corresponding composite-altered version in the test phase (b).

(c) Reaction time (RT) difference: Mean RT difference (test phase – last 30 trials of training) for the configural and featural conditions. RT significantly increased in the configural condition ($p < 0.0001$), while it significantly decreased in the featural condition ($p < 0.05$). The RT difference between the two conditions was also significant ($p < 0.0001$).

(d) Accuracy difference: Mean accuracy difference (last 30 trials of training – test phase). Accuracy significantly decreased in the configural condition ($p < 0.05$), while no significant change was observed in the featural condition (*n.s.*). The accuracy difference between the two conditions was also significant ($p < 0.05$).

Error bars represent standard errors of the mean (SEM). Statistical significance: *n.s.* (not significant), * $p < 0.05$, ****$p < 0.0001$.

## Experiment 4: Misalignment Effect

Experiment 4 explores the implications of stimulus misalignment on configural processing. By misaligning the top and bottom halves of configural and feature-based stimuli, we aim to assess whether such misalignment disrupts holistic processing specifically for the configural stimuli only.

### Methods

The same 20 participants from Experiment 1, having met our learning criteria, engaged in this study. They underwent 30 test trials both for misaligned configural and featural stimuli (Fig. 6b). All other methods not mentioned here remain consistent with the procedures described in previous experiments. Participants were instructed to perform the task as before, but with stimuli that are misaligned.

### Results

Misaligned stimuli produced a significant increase in RT for the configural condition, $t(19) = 4.02$, $p < 0.001$, but not for the featural condition, $t(19) = 0.70$, $p = 0.491$, underscoring the distinct impact of misalignment on configural processing (Fig. 6c). Accuracy data did not show significant differences post-misalignment in either condition (configural $t(19) = -0.49$, $p = 0.632$; featural $t(19) = 1.28$, $p = 0.218$) (Fig. 6d). It is possible that a larger misalignment may led to a stronger effect in accuracy.

We also observed significant differences between the configural and featural conditions in RT. The misalignment effect was significantly larger for the configural condition than the featural condition, $t(38) = 2.59$, $p = 0.014$, indicating a greater impact of misalignment on configural processing. We did not observe a task difference in accuracy, $t(38) = 0.92$, $p = 0.362$ (Fig. 6c and 6d).

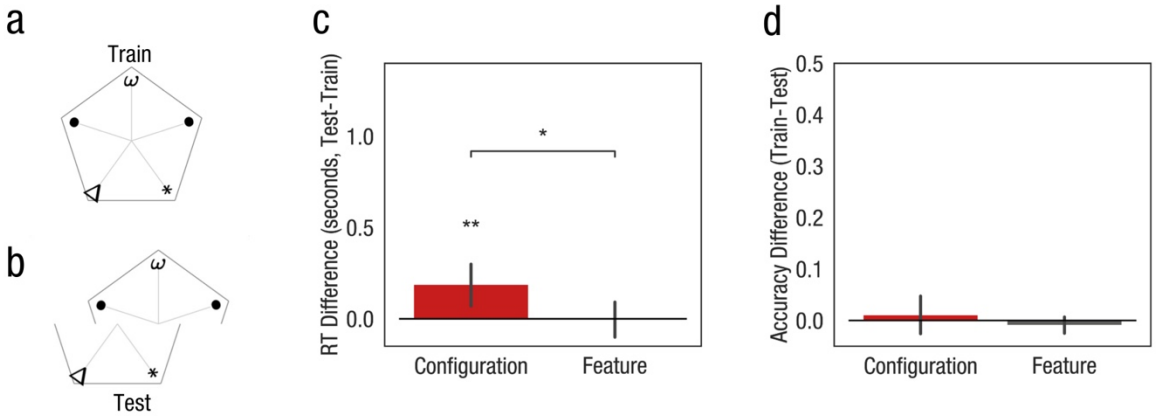



**Figure 6**. Misalignment effect on reaction time and accuracy.

(a, b) Stimulus example: Illustration of a training stimulus (a) and its corresponding misaligned version in the test phase (b).

(c) Reaction time (RT) difference: Mean RT difference (test phase – last 30 trials of training) for the configural and featural conditions. RT significantly increased in the configural condition ($p < 0.01$), while no significant change was observed in the featural condition (*n.s.*). The RT difference between the two conditions was significant ($p < 0.05$).

(d) Accuracy difference: Mean accuracy difference (last 30 trials of training – test phase). Accuracy changes were not significant in either condition (*n.s.*), and no significant difference was observed between the configural and featural conditions.

Error bars represent standard errors of the mean (SEM). Statistical significance: *n.s.* (not significant), * $p < 0.05$, ** $p < 0.01$.

## Experiment 5: Machine Vision

In recent years, convolutional neural networks (CNNs) have achieved remarkable success in object recognition, which also raises questions about whether such models can capture processing mechanisms similar to those observed in human cognition. While CNNs are designed to excel in complex pattern recognition, it remains an open question whether they can replicate same configuration effects as we observe here in human visual perception tasks.

### Methods

**Stimuli:** To generate a large number of images based on the original exemplars used in human behavior tasks to train and test CNN models, we implemented data augmentation by applying adding 5% random pixel noise. We generated 720 training images for each category (1440 images for configural task and 1440 images for featural task). Accordingly, we also generated

corresponding upright (same data augmentation processing on the original exemplars but different from the images in the training set), inverted, part, composite, and misaligned images for the model test.

**Model and training procedure**: We utilized pre-trained ResNet-50 models (He et al., 2015) on ImageNet (Russakovsky et al., 2015) to perform transfer learning based on our stimuli (for other ResNet models results, see supplemental material Fig. 3), allowing them to learn the configuration and feature-based categorical classification tasks, resulting in a separate set of configuration-directed and feature-directed models. In the training of transfer learning, the models were trained using the standard categorical cross entropy loss with a static learning rate of 0.0002 and the Adam optimizer (Kingma & Ba, 2014). We trained 20 configuration-directed and 20 feature-directed ResNets to simulate 20 individuals' behaviors.

**Model evaluations**: We tested our configuration-directed and feature-directed ResNets on corresponding five types of test sets respectively. All ResNets achieved an accuracy of over 95% on the training stimuli. We computed the difference of classification accuracy as a counterpart measure for human accuracy, and we used classification probability as a counterpart measure similar to human RT, for the inverted, part, composite, or misalignment conditions.

**Results**

To evaluate whether ResNet models exhibited similar task effects as humans, we computed accuracy and probability differences between the upright condition and each test condition for both the configuration-directed and feature-directed models.

We observed several significant effects in model accuracy across conditions. The Inversion task (Fig. 7b) showed a significant decrease in accuracy for both the configuration ($t(19) = 103.92$, $p < 0.00001$) and feature conditions ($t(19) = 13.04$, $p < 0.001$), with a significant difference between conditions ($t(38) = 8.67$, $p < 0.00001$, Bonferroni-Holm corrected). In the Misalignment task, accuracy significantly decreased in both the configuration ($t(19) = 3.12$, $p = 0.006$) and feature conditions ($t(19) = 6.27$, $p < 0.001$), but the difference between conditions was not significant ($t(38) = -1.22$, $p = 0.918$, Bonferroni-Holm corrected). In the Composite task, accuracy significantly decreased in the configuration condition ($t(19) = 89.50$, $p < 0.00001$), while no significant change was observed in the feature condition ($t(19) = -1.20$, $p = 0.246$). The difference between conditions was significant ($t(38) = 88.02$, $p < 0.00001$, Bonferroni-Holm corrected). In the Part-Whole task, accuracy differences were not significant in either the configuration ($t(19) = 1.44$, $p = 0.166$) or feature conditions ($t(19) = -1.20$, $p = 0.246$), and the difference between conditions was also not significant ($t(38) = 1.52$, $p = 0.548$, Bonferroni-Holm corrected).

For probability differences (Fig. 7a), the Inversion task showed a significant decrease in both the configuration ($t(19) = 11.15$, $p < 0.001$) and feature conditions ($t(19) = 12.47$, $p < 0.001$), with no significant difference between conditions ($t(38) = 2.43$, $p = 0.080$, Bonferroni-Holm corrected).

In the Misalignment task, probability significantly decreased in both the configuration ($t(19) = 11.22$, $p < 0.001$) and feature conditions ($t(19) = 19.88$, $p < 0.00001$), and the difference between conditions was not significant ($t(38) = 2.00$, $p = 0.213$, Bonferroni-Holm corrected). In the Composite task, probability significantly decreased in the configuration condition ($t(19) = 13.42$, $p < 0.001$) but increased in the feature condition ($t(19) = -9.58$, $p < 0.001$). The difference between conditions was significant ($t(38) = 16.48$, $p < 0.00001$, Bonferroni-Holm corrected). In the Part-Whole task, probability significantly decreased in the configuration condition ($t(19) = 4.39$, $p < 0.001$), while it significantly increased in the feature condition ($t(19) = -8.58$, $p < 0.001$). The difference between conditions was significant ($t(38) = 7.57$, $p < 0.00001$, Bonferroni-Holm corrected).

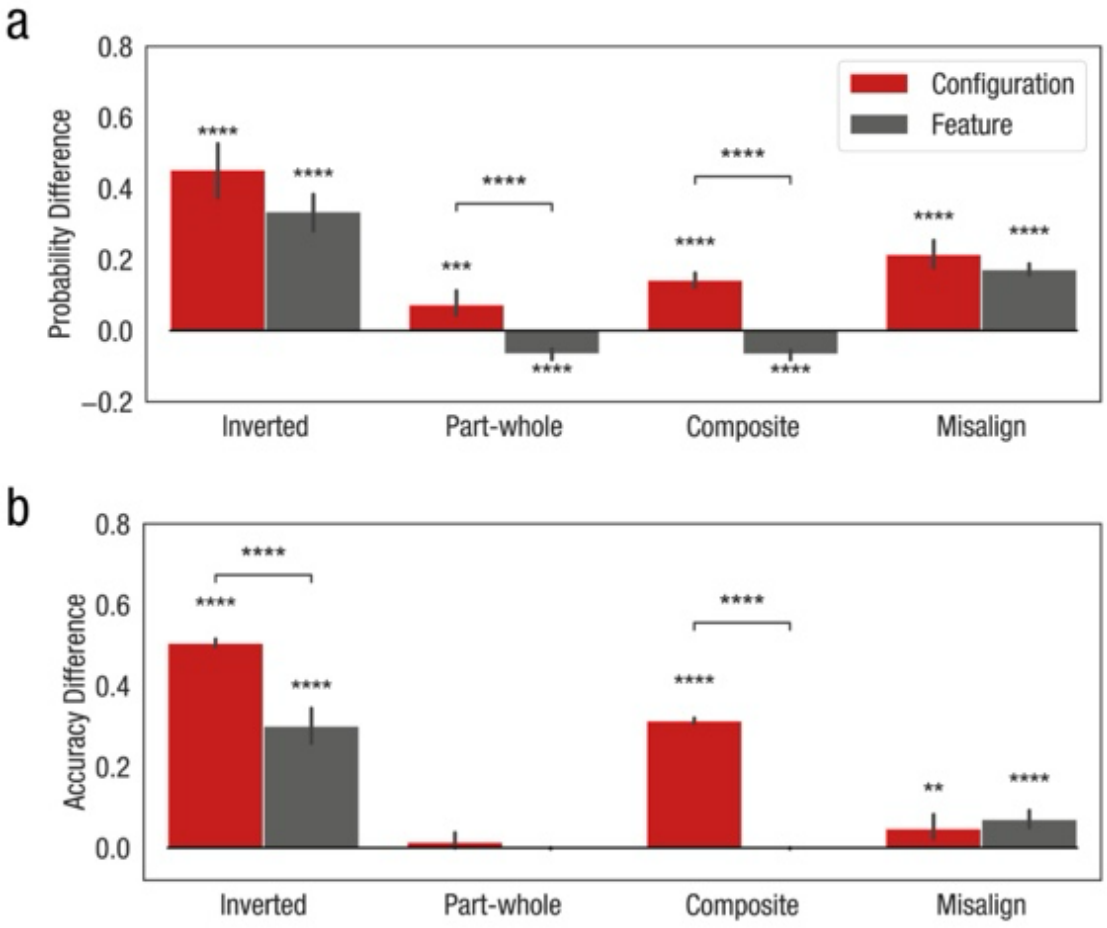



**Figure 7**. ResNet model performance across tasks.

(a) Probability difference: Mean probability difference (test phase – training phase) for the configuration (red) and feature (gray) conditions across four tasks (Inversion, Part-Whole, Composite, and Misalignment). Probability significantly decreased in both conditions for

Inversion ($p < 0.001$) and Misalignment ($p < 0.001$) tasks, while Composite ($p < 0.001$) and Part-Whole ($p < 0.001$) tasks showed opposite patterns between conditions. The difference between conditions was significant for Composite ($p < 0.0001$, Bonferroni-Holm corrected) and Part-Whole tasks ($p < 0.00001$, Bonferroni-Holm corrected), but not for Inversion or Misalignment tasks (*n.s.*).

(b) Accuracy difference: Mean accuracy difference (test phase – training phase) for the configuration (red) and feature (gray) conditions. Accuracy significantly decreased in both conditions for Inversion ($p < 0.0001$) and Misalignment ($p < 0.01$) tasks, while only the configuration condition showed a significant decrease in the Composite task ($p < 0.0001$). No significant accuracy differences were observed for the Part-Whole task (*n.s.*). The difference between conditions was significant for Inversion ($p < 0.0001$, Bonferroni-Holm corrected) and Composite tasks ($p < 0.0001$, Bonferroni-Holm corrected).

Error bars represent standard errors of the mean (SEM). Statistical significance: *n.s.* (not significant), ** $p < 0.01$, *** $p < 0.001$, ****$p < 0.0001$.

## Discussion

We systematically tested all major holistic processing effects—the inversion (Yin, 1969), part-whole (Tanaka & Farah, 1993), composite (Young et al., 1987), and misalignment effects—using non-face stimuli that isolated configural and featural information. Across all tasks, holistic processing emerged exclusively for configuration-based stimuli, while feature-based stimuli did not elicit the same effects. Notably, while each of these effects has been studied extensively in face perception, few studies have systematically examined all hallmark holistic effects within a single experimental framework, even in research on faces (e.g., Boutet et al., 2021; Rezlescu et al., 2017). Importantly, these behavioral findings were consistent with computational modeling: a deep neural network trained on real-world images exhibited similar holistic patterns, reinforcing the robustness of these effects.

Here, we not only included all key holistic tasks but also demonstrated that configuration alone is sufficient to elicit holistic processing in non-face stimuli. This challenges the long-standing view that holistic processing is unique to face perception (Richler et al., 2011; Wang et al., 2012) and that these effects are specifically tied to face perception mechanisms (Farah et al., 1998; Maurer et al., 2002; McKone et al., 2007). Instead, our findings support the idea that holistic effects arise as a consequence of configural processing itself, irrespective of stimulus category. Similar findings have been reported in studies showing that holistic processing can be driven by object-based spatial relationships rather than expertise alone (Zhao et al., 2016).

Our results demonstrate that holistic processing is not exclusive to faces but is triggered by configuration itself. By isolating configural structure and removing featural cues, we showed that holistic effects emerge purely from spatial relationships, rather than any intrinsic face-specific

properties. However, faces do inherently contain strong configural properties (Diamond & Carey, 1986; Rhodes et al., 1993), which may be rare among other object categories (Robbins & McKone, 2007). This rarity could explain why holistic effects are predominantly observed in face perception, reinforcing the idea that faces are not unique in being processed holistically but instead represent an optimal case of configural processing.

While our findings establish configuration as a primary driver of holistic processing, expertise has also been proposed as a contributing factor (Gauthier & Tarr, 1997; Richler et al., 2011; Wong et al., 2009). In our experiments, participants exhibited robust holistic effects for configural stimuli despite minimal training (~15 minutes), in contrast to the extensive training typically required to develop expertise-based holistic effects—ranging from hours in laboratory settings to decades in professional domains (Diamond & Carey, 1986; Gauthier & Tarr, 1997). This suggests that extensive expertise is not a prerequisite for holistic processing to emerge, further supporting our claim that configuration alone is sufficient to trigger holistic perception. However, we do not rule out the possibility that configuration and expertise contribute to holistic processing through distinct mechanisms. Future research should explore whether extensive training can induce holistic effects in featural stimuli, which would provide further insight into why faces are perceived as special—not because they uniquely elicit holistic processing, but because they exist at the intersection of strong configural properties and lifelong expertise.

Interestingly, we observed same holistic effects in CNNs, further confirming that isolated configuration serves as a key componence for the holistic processing. Similar to human studies, previous research has only examined holistic processing in faces (Dobs et al., 2023; Peng et al., 2021; Tian et al., 2022; Wirth et al., 2024), whereas our study is the first to systematically evaluate holistic processing effects for non-face stimuli in CNN. This suggests that configuration may serve as a universal factor across visual systems – both human and artificial – that induces holistic processing effects. Furthermore, some recent studies have employed reverse engineering approaches as a novel perspective to gain insights into brain visual mechanisms (Huang et al., 2022; Kanwisher et al., 2023; Lu & Ku, 2023; S. Xu et al., 2021). Given that CNNs exhibit human-like behavior in this task, future research could leverage CNNs as a tool for manipulating and analyzing the underlying mechanisms, offering a potential avenue to infer the origins of holistic processing.

One potential limitation of our study is the lack of individuation of stimuli. Holistic processing is often associated with fine-grained discrimination among individual exemplars, particularly in faces (Kanwisher, 2000). Our stimuli were designed to isolate configural information, but participants were not required to learn or differentiate individual exemplars. It remains possible that individuation plays a critical role in the development of holistic processing, which future studies should investigate by directly comparing individuation-based and category-based learning. Additionally, we did not collect eye-tracking data, which could provide deeper insight into attentional strategies and feature integration during holistic processing (Kundel et al., 2007; Zhong et al., 2023). Since holistic perception may be linked to distinct gaze patterns, future

research using eye-tracking could clarify how visual information is prioritized when processing configural versus featural stimuli.

Finally, face processing deficits are well-documented in clinical populations, including prosopagnosia and autism spectrum disorders, where impairments in configural processing have been suggested as a key factor (Behrmann et al., 2006; Palermo et al., 2011). Given our findings that holistic processing is driven by configuration rather than face specificity, stimuli that isolate configural and featural information may offer a useful approach for assessing these deficits and investigating their underlying mechanisms. Moreover, recent evidence suggests that enhancing configural information improves object perception in computational models (Jang et al., 2024), further supporting the importance of configural processing as a fundamental visual mechanism.

To conclude, our findings provide strong evidence that configural properties are a key driver of holistic processing. By showing that holistic effects arise purely from spatial relationships in the absence of featural cues, we challenge the notion that holistic processing is exclusive to face perception. These results underscore the need for a broader investigation of configural processing as a fundamental mechanism in the visual system (Zachariou et al., 2014). More broadly, our findings establish a framework for understanding how spatial relationships shape perception beyond the domain of faces.

## Reference


Behrmann, M., Avidan, G., Leonard, G. L., Kimchi, R., Luna, B., Humphreys, K., & Minshew, N. (2006). Configural processing in autism and its relationship to face processing. *Neuropsychologia*, *44*(1), 110–129. https://doi.org/10.1016/j.neuropsychologia.2005.04.002

Bilalić, M., Grottenthaler, T., Nägele, T., & Lindig, T. (2016). The Faces in Radiological Images: Fusiform Face Area Supports Radiological Expertise. *Cerebral Cortex*, *26*(3), 1004–1014. https://doi.org/10.1093/cercor/bhu272

Bilalic, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many Faces of Expertise: Fusiform Face Area in Chess Experts and Novices. *Journal of Neuroscience*, *31*(28), 10206–10214. https://doi.org/10.1523/JNEUROSCI.5727-10.2011

Boutet, I., Nelson, E. A., Watier, N., Cousineau, D., Béland, S., & Collin, C. A. (2021). Different measures of holistic face processing tap into distinct but partially overlapping mechanisms. *Attention, Perception, & Psychophysics*, *83*(7), 2905–2923. https://doi.org/10.3758/s13414-021-02337-7

Brants, M., Wagemans, J., & Op De Beeck, H. P. (2011). Activation of Fusiform Face Area by Greebles Is Related to Face Similarity but Not Expertise. *Journal of Cognitive Neuroscience*, *23*(12), 3949–3958. https://doi.org/10.1162/jocn_a_00072

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107–117. https://doi.org/10.1037/0096-3445.115.2.107

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

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482–498. https://doi.org/10.1037/0033-295X.105.3.482

Freiwald, W., Duchaine, B., & Yovel, G. (2016). Face Processing Systems: From Neurons to Real-World Social Perception. *Annual Review of Neuroscience*, *39*(1), 325–346. https://doi.org/10.1146/annurev-neuro-070815-013934

Galton, F. (1879). Composite Portraits, Made by Combining Those of Many Different Persons Into a Single Resultant Figure. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *8*, 132. https://doi.org/10.2307/2841021

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. https://doi.org/10.1038/72140

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" Expert: Exploring Mechanisms for Face Recognition. *Vision Research*, *37*(12), 1673–1682. https://doi.org/10.1016/S0042-6989(96)00286-6

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1512.03385

Huang, T., Song, Y., & Liu, J. (2022). Real-world size of objects serves as an axis of object space. *Communications Biology*, *5*(1), 749. https://doi.org/10.1038/s42003-022-03711-3

Jang, H., Sinha, P., & Boix, X. (2024). *Configural processing as an optimized strategy for robust object recognition in neural networks* (arXiv:2407.19072). arXiv. https://doi.org/10.48550/arXiv.2407.19072

Jiang, F., Blanz, V., & Rossion, B. (2011). Holistic processing of shape cues in face identification: Evidence from face inversion, composite faces, and acquired prosopagnosia. *Visual Cognition*, *19*(8), 1003–1034. https://doi.org/10.1080/13506285.2011.604360

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*(8), 759–763. https://doi.org/10.1038/77664

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

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization* (Version 9). arXiv. https://doi.org/10.48550/ARXIV.1412.6980

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology*, *242*(2), 396–402. https://doi.org/10.1148/radiol.2422051997

Lu, Z., & Ku, Y. (2023). Bridging the gap between EEG and DCNNs reveals a fatigue mechanism of facial repetition suppression. *iScience*, *26*(12), 108501. https://doi.org/10.1016/j.isci.2023.108501

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, *11*(1), 8–15. https://doi.org/10.1016/j.tics.2006.11.002

Palermo, R., Willis, M. L., Rivolta, D., McKone, E., Wilson, C. E., & Calder, A. J. (2011). Impaired holistic coding of facial expression and facial identity in congenital prosopagnosia. *Neuropsychologia*, *49*(5), 1226–1235. https://doi.org/10.1016/j.neuropsychologia.2011.02.021

Peng, Y., Lee, H., Shu, T., & Lu, H. (2021). Exploring biological motion perception in two-stream convolutional neural networks. *Vision Research*, *178*, 28–40. https://doi.org/10.1016/j.visres.2020.09.005

Piepers, D. W., & Robbins, R. A. (2012). A Review and Clarification of the Terms "holistic," "configural," and "relational" in the Face Perception Literature. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00559

Rezlescu, C., Susilo, T., Wilmer, J. B., & Caramazza, A. (2017). The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1961–1973. https://doi.org/10.1037/xhp0000400

Rhodes, G., Brake, S., & Atkinson, A. P. (1993). What's lost in inverted faces? *Cognition*, *47*(1), 25–57. https://doi.org/10.1016/0010-0277(93)90061-Y

Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Holistic Processing Predicts Face Recognition. *Psychological Science*, *22*(4), 464–471. https://doi.org/10.1177/0956797611401753

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, *103*(1), 34–79. https://doi.org/10.1016/j.cognition.2006.02.008

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Tanaka, J. W., & Farah, M. J. (1993). Parts and Wholes in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*, *46*(2), 225–245. https://doi.org/10.1080/14640749308401045

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, *25*(5), 583–592. https://doi.org/10.3758/BF03211301

Tian, F., Xie, H., Song, Y., Hu, S., & Liu, J. (2022). The Face Inversion Effect in Deep Convolutional Neural Networks. *Frontiers in Computational Neuroscience*, *16*, 854218. https://doi.org/10.3389/fncom.2022.854218

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual Differences in Holistic Processing Predict Face Recognition Ability. *Psychological Science*, *23*(2), 169–177. https://doi.org/10.1177/0956797611420575

Wirth, B. E., Knoblauch, A., & Müller, P. (2024). *Face Processing in Humans and CNNs: Comparing the Reliance on Holistic and Local Feature-Based Information*. https://doi.org/10.31234/osf.io/2xjcm

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for Facelike Expertise With Objects: Becoming a Ziggerin Expert—but Which Type? *Psychological Science*, *20*(9), 1108–1117. https://doi.org/10.1111/j.1467-9280.2009.02430.x

Xu, S., Zhang, Y., Zhen, Z., & Liu, J. (2021). The Face Module Emerged in a Deep Convolutional Neural Network Selectively Deprived of Face Experience. *Frontiers in Computational Neuroscience*, *15*, 626259. https://doi.org/10.3389/fncom.2021.626259

Xu, Y. (2005). Revisiting the Role of the Fusiform Face Area in Visual Expertise. *Cerebral Cortex*, *15*(8), 1234–1242. https://doi.org/10.1093/cercor/bhi006

Xu, Y., Liu, J., & Kanwisher, N. (2005). The M170 is selective for faces, not for expertise. *Neuropsychologia*, *43*(4), 588–597. https://doi.org/10.1016/j.neuropsychologia.2004.07.016

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational Information in Face Perception. *Perception*, *16*(6), 747–759. https://doi.org/10.1068/p160747

Yovel, G., & Kanwisher, N. (2004). Face Perception. *Neuron*, *44*(5), 889–898. https://doi.org/10.1016/j.neuron.2004.11.018

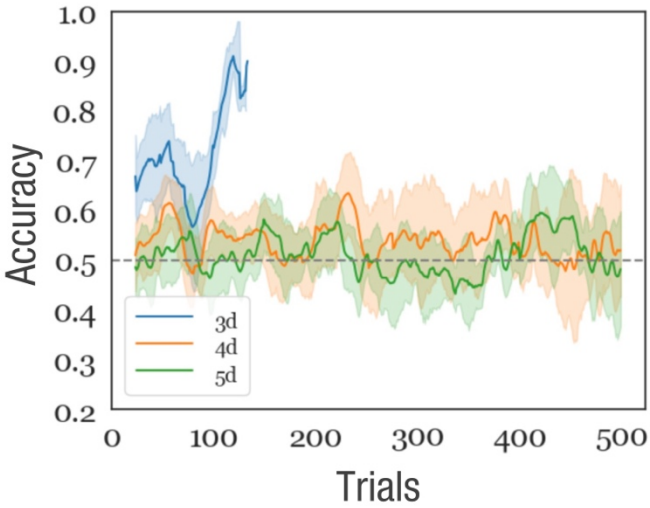
Yovel, G., & Kanwisher, N. (2005). The Neural Basis of the Behavioral Face-Inversion Effect. *Current Biology*, *15*(24), 2256–2262. https://doi.org/10.1016/j.cub.2005.10.072

Zachariou, V., Klatzky, R., & Behrmann, M. (2014). Ventral and Dorsal Visual Stream Contributions to the Perception of Object Shape and Object Location. *Journal of Cognitive Neuroscience*, *26*(1), 189–209. https://doi.org/10.1162/jocn_a_00475

Zhao, M., Bülthoff, H. H., & Bülthoff, I. (2016). Beyond Faces and Expertise: Facelike Holistic Processing of Nonface Objects in the Absence of Expertise. *Psychological Science*, *27*(2), 213–222. https://doi.org/10.1177/0956797615617779

Zhong, N., Hsiao, J. H., Zhou, G., & Hayward, W. G. (2023). Association of idiosyncratic eye-movement patterns with holistic processing of faces as measured by the composite face effect and the face inversion effect. *Visual Cognition*, *31*(7), 485–500. https://doi.org/10.1080/13506285.2024.2315783

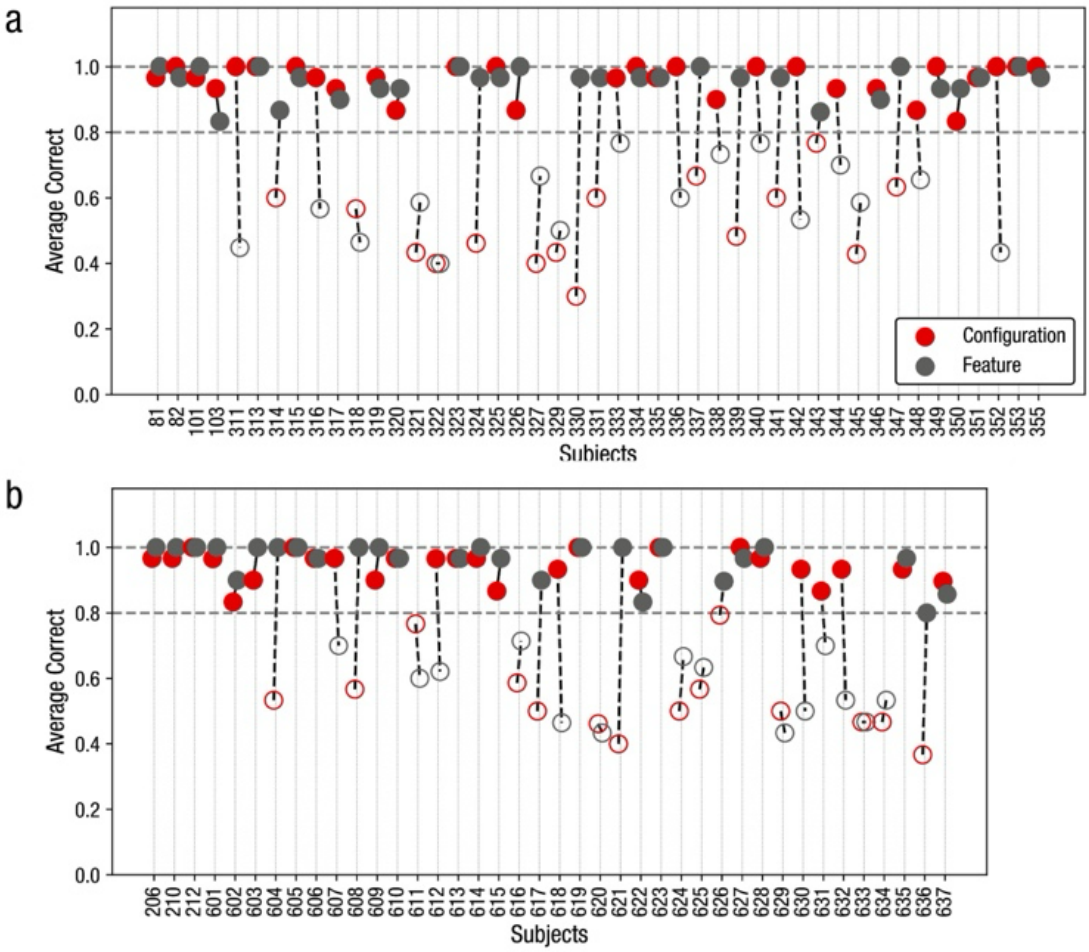
## Supplemental material




**Supplemental Figure 1.** Pilot Study on Configural Categorization Learning

Example stimuli used in the pilot study, testing participants' ability to categorize 3-dimensional (3D), 4-dimensional (4D), and 5-dimensional (5D) configural stimuli. The dashed circles indicate the dimensions manipulated in each condition but were not visible to participants during the task. (b) Learning performance across 500 trials, measured as categorization accuracy for each dimensional condition. Shaded regions represent standard error.

Thirteen participants were recruited to evaluate their ability to categorize stimuli based on configural relationships. Each participant was trained to categorize stimuli into two predefined groups. While participants successfully learned the 3D stimuli, reaching over 90% accuracy within approximately 120 trials, no learning trend was observed for the 4D or 5D stimuli, even after extensive training. Based on these results, we used 3D stimuli in the main study to ensure learnability while preserving the configural nature of the task.



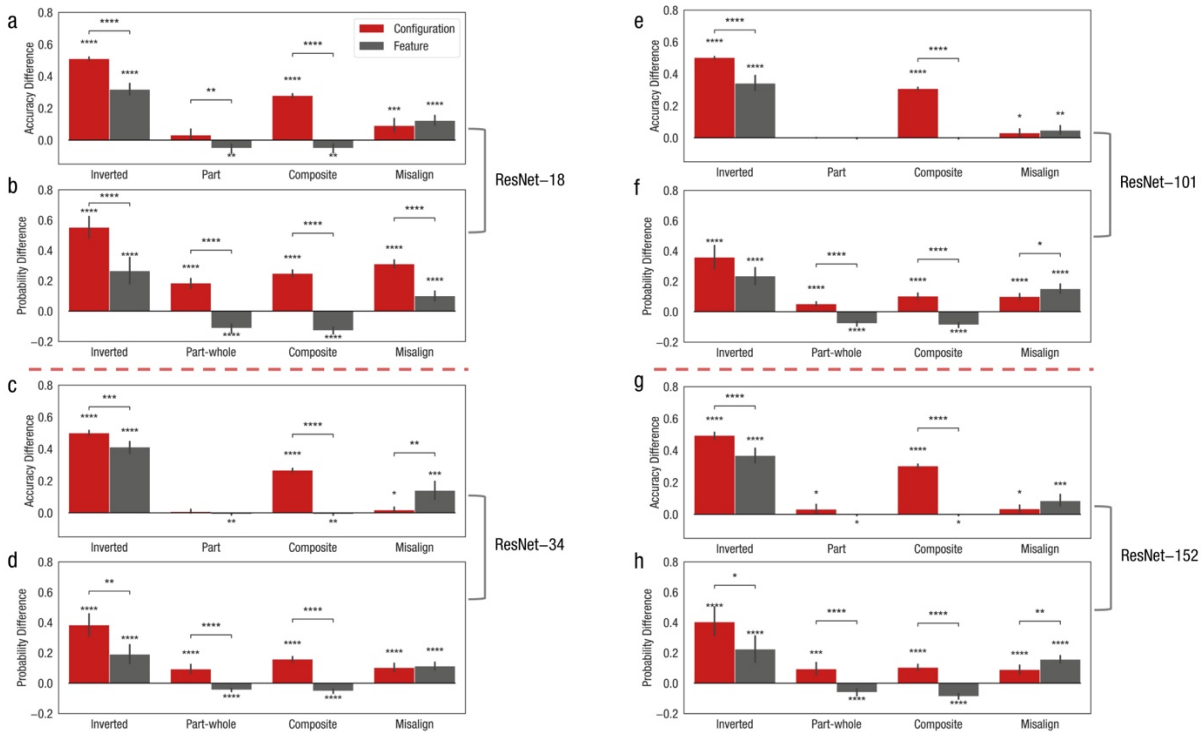

**Supplemental Figure 2.** Participant Training Performance.

(a) and (b) show individual participant training performance for the configural and featural categorization tasks. The x-axis represents participant IDs, while the y-axis indicates average categorization accuracy. Red circles denote performance in the configural condition, and gray circles represent performance in the featural condition. Solid circles indicate participants who achieved ≥80% accuracy in both tasks, while open circles indicate accuracy below 80%. Each participant's configural and featural accuracy rates are connected by a line; dashed lines indicate that at least one task fell below the 80% inclusion criterion.

(a) Participants recruited for Experiment 1 and Experiment 4.

(b) Participants recruited for Experiment 2 and Experiment 3.

Only participants who met the ≥80% accuracy criterion in both conditions were included in the final analysis.




**Supplemental Figure 3.** Performance of ResNet Models on Configural and Featural Stimuli.

This figure presents the results of four ResNet architectures (ResNet-18, ResNet-34, ResNet-101, and ResNet-152) trained on our stimulus set, evaluating their ability to replicate human-like holistic effects. Each model was tested on four tasks: inversion, part-whole, composite, and misalignment.

(a, b) Accuracy differences (a) and probability differences (b) between training and test conditions for ResNet-18.

(c, d) Accuracy (c) and probability (d) differences for ResNet-34.

(e, f) Accuracy (e) and probability (f) differences for ResNet-101.

(g, h) Accuracy (g) and probability (h) differences for ResNet-152.

In all plots, red bars represent configuration-based stimuli, and gray bars represent feature-based stimuli. Bars indicate the difference between training and task performance (training – task). Statistical analyses include: 1) One-sample t-tests comparing accuracy and probability differences against zero (uncorrected). 2) Condition-wise t-tests comparing configural vs. featural conditions for each task (Bonferroni-Holm corrected).