communications biology
A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-07672-1

Configural processing as an optimized strategy for robust object recognition in neural networks

Check for updates

Hojin Jang[1,2], Pawan Sinha[2] & Xavier Boix[3]

Configural processing, the perception of spatial relationships among an object's components, is crucial for object recognition, yet its teleology and underlying mechanisms remain unclear. We hypothesize that configural processing drives robust recognition under varying conditions. Using identification tasks with composite letter stimuli, we compare neural network models trained with either configural or local cues. We find that configural cues support robust generalization across geometric transformations (e.g., rotation, scaling) and novel feature sets. When both cues are available, configural cues dominate local features. Layerwise analysis reveals that sensitivity to configural cues emerges later in processing, likely enhancing robustness to pixel-level transformations. Notably, this occurs in a purely feedforward manner without recurrent computations. These findings with letter stimuli successfully extend to naturalistic face images. Our results demonstrate that configural processing emerges in a naïve network based on task contingencies, and is beneficial for robust object processing under varying viewing conditions.

Configural processing refers to the perception of the spatial relationships among an object's components, which facilitates integrated and holistic recognition of objects. Early research has highlighted the importance of configural processing in object recognition. Biederman's Recognition-by-Components theory posits that object recognition is driven by identifying simple geometric components and their configurations[1]. Subsequent studies have demonstrated that configural processing extends to the recognition of complex objects and is particularly evident in expert systems. For instance, expert bird watchers and car enthusiasts can discern subtle differences within similar species or models due to enhanced configural processing abilities[2–4]. This evidence suggests that extensive experience with specific categories amplifies our capacity for holistic and configural object processing.

Among various object categories, humans are natural face experts, and this expertize is closely linked to configural processing. Research demonstrates that humans are highly sensitive to the spatial configurations of facial components[5–7]. Neurotypical individuals can easily identify subtle differences in interpupillary distance or philtrum length between two individuals with otherwise identical facial features such as eyes, nose, and mouth[8]. This holistic processing ability has often been suggested to underpin face specificity over other object categories[9–11]. Thus, the study of configural processing offers insights into our specialized perceptual skills and the foundations of expert visual recognition.

Despite extensive research, however, the functional benefits of configural processing are not fully understood. One might assume that focusing on individual local features could offer more advantages for expert recognition systems, yet evidence indicates the opposite. Understanding the advantages of configural processing will illuminate why experts adopt this strategy over a piecemeal approach.

The present study introduces a novel perspective driven by teleological considerations; why might a visual system develop a configural processing strategy? We hypothesize that prioritizing configural cues over local feature processing could be an ecologically driven strategy, optimizing recognition under diverse viewing scenarios. A few psychological studies provide the motivation for this hypothesis, indicating the crucial role of configural processing in facilitating face recognition under demanding conditions. For instance, McKone[12] found that configural processing remained consistent across views, while part-based processing was view-sensitive, implying configural processing's adaptive function in robust face recognition amidst varying local image details McKone[13]. further suggested that holistic processing is effective across a broad distance range Piepers and Robbins[14]. proposed that if the holistic representation of a face is based not on the shape of its features but on a relational structure anchored from the features' center points, then alterations in the face's second-order configuration would be minimal,

[1]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea. [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Artificial Intelligence Laboratory, Fujitsu Research of America, Silicon Valley, CA, USA. e-mail: hojin4671@korea.ac.kr; xboix@fujitsu.com

thus offering stable cues for recognition. While appealing, these proposals currently lack robust computational validation.

Recent progress in deep learning offers a viable framework to test various hypotheses in cognitive science. Studies indicate that deep neural networks, specifically those trained for object recognition tasks, serve as the most advanced models of the biological visual system[15,16], accurately predicting visual cortical responses in human and macaque brains[17–22]. Moreover, these models exhibit a substantial concordance with human face recognition behaviors[23–26] (see reviews[27,28]), thus offering a variety of testable hypotheses for object and face recognition research.

Neural network models tasked with object recognition present an opportunity to investigate their preference for local featural versus configural cues. Interestingly, current research suggests that these models often exhibit a preference for local feature-based processing over global processing[29,30]. More recent studies have reported that deep neural networks struggle with capturing configural cues in shape recognition tasks[31,32]. However, many of these studies do not clearly differentiate between local featural and configural processing, and both types of processing are often conflated in complex object recognition tasks, making it difficult to separate their roles.

In this study, we investigated whether deep learning models could come to utilize configural cues for recognition based on the contingencies of task demands during training and, if so, whether these models prioritized configural processing over local processing under dynamic viewing conditions. To this end, we created composite letter stimuli and compared various neural network models trained on tasks with either exclusively local or configural cues. We found that neural networks effectively learned to encode and generalize configural cues, demonstrating robustness to geometric transformations, such as rotation and scaling, comparable to that of local features. Furthermore, when both feature types were concurrently available, the models favored configural cues over local ones. A layer-by-layer analysis demonstrated an increasing sensitivity to configural cues compared to local features in higher network layers, which might explain their resilience to pixel-level changes. Notably, this configural processing did not seem to require recurrent computations. The results obtained using letter stimuli were also successfully replicated with real-world face images. In summary, our research demonstrates experience dependent genesis of configural processing strategies, and shows how they might contribute towards achieving robust and reliable recognition capabilities across diverse viewing conditions.

Results

Deep neural networks can effectively capture configural cues in recognition

To investigate the role of local and configural processing in recognition, it is crucial to separate one from the other. This study utilized the EMNIST dataset[33] to create composite patterns, where different letters represented individual features (Fig. 1A). In total, 9 letters were selected: B, D, L, M, N, P, R, T, and U. Subsequently, two distinct tasks were designed: one focusing on local featural processing ("local task") and the other concentrating on configural processing ("configural task"). In the local task, distinct classes

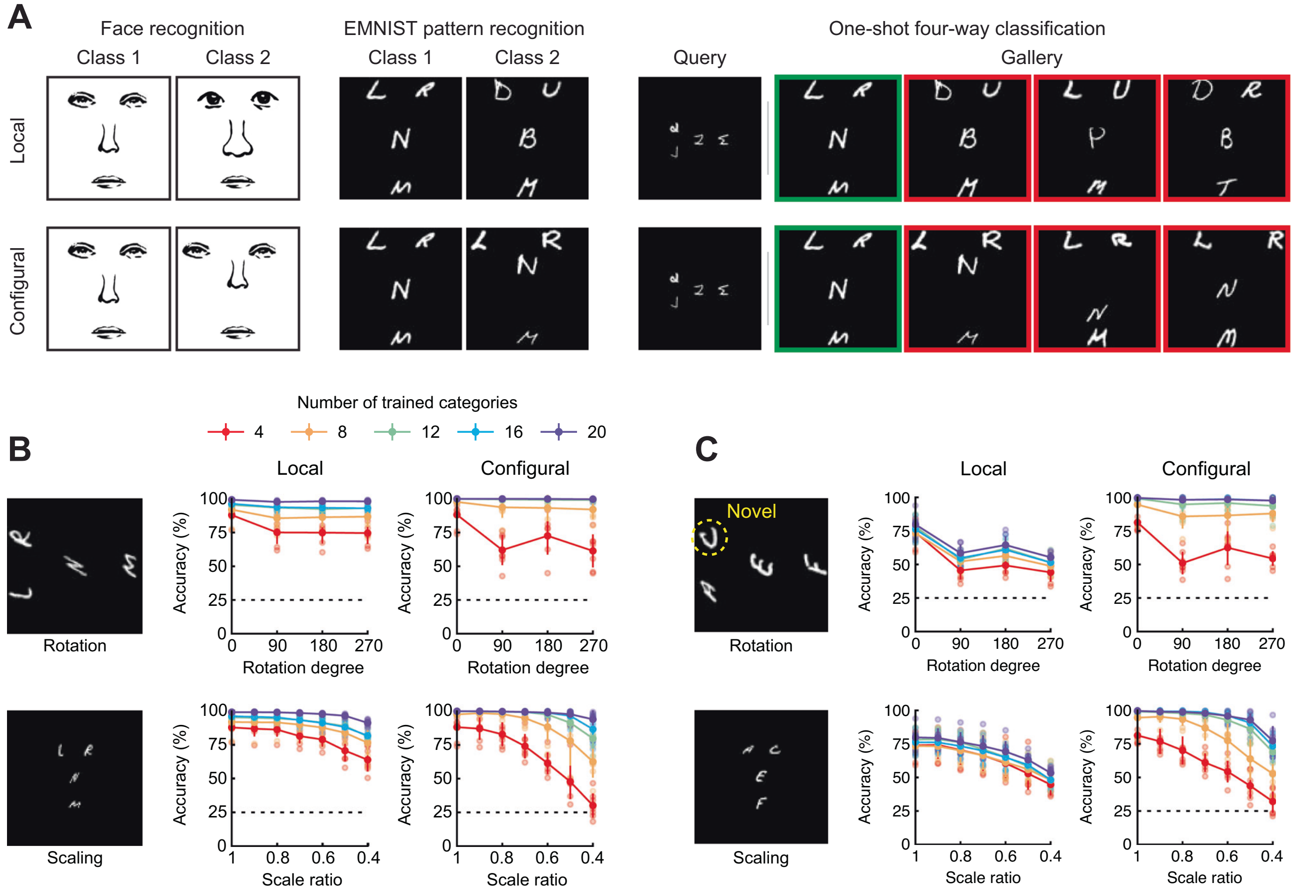

Fig. 1 | Visual stimuli, task conditions, and performance accuracy in local and configural tasks. **A** Conceptual illustrations (left) and actual representations (middle) of visual stimuli for local (top) and configural (bottom) tasks. Depiction of a one-shot four-way classification scenario with targets marked by green squares and distractors by red squares (right). **B** Performance accuracy for local and configural tasks under rotation (top) and scaling (bottom) transformations. The dashed line indicates chance level performance. Different colors represent the number of categories trained. Error bars indicate the standard deviation across seven neural networks. **C** Performance accuracy for local and configural tasks under rotation (top) and scaling (bottom) transformations, with patterns that included novel local features.

were characterized by a unique set of letters, sharing the same configuration. Conversely, in the configural task, all classes used an identical set of letters, but varied in their configurations. A total of 24 classes were generated for each task (Supplementary Fig. 1).

We evaluated two major types of neural network architectures, including feedforward and recurrent neural networks (detailed in the "Methods" section). To investigate the capability of these models in solving both tasks, we employed a one-shot four-way classification paradigm. In this paradigm, each classification task involves four galleries, with each gallery containing a single example from a specific class. A query image, which is from the same class as one of the galleries but transformed (e.g., through rotation or scaling), is presented to the model. The network's task is to correctly identify the class of the query image by comparing it to the examples in the galleries. To assess similarity and determine the correct class assignment, we used a distance metric, specifically the Euclidean distance between the query and gallery images. This approach was applied to both local and configural processing tasks (Fig. 1A). This one-shot learning paradigm allows for a more precise analysis of network generalization performance, particularly in how networks use local or configural cues to recognize novel classes. It also provides insights into how the diversity of training classes affects generalization[34]. We varied the number of training classes to include 4, 8, 12, 16, and 20 out of a total of 24, and subsequently evaluated the network's ability to generalize to four patterns previously unseen by the system. Note that the total number of training images remained the same across all conditions.

Figure 1B presents the accuracy levels for two tasks under two types of transformations: rotation (top) and scaling (bottom). The results revealed that the neural networks efficiently handled the local task, showing that local features were consistently effective for identification across various degrees of transformation. Consistent with the findings of Jang et al.[34], there was a noticeable improvement in generalization performance as the diversity of training classes expanded, even while the total amount of training images stayed the same. For the configural task, although initial performance was less than optimal with only four training classes, the networks quickly reached peak performance as the number of training classes increased, demonstrating effective use of configural cues. Collectively, these results underscore the efficacy of both configural and local feature cues within neural networks across different transformation conditions.

One could argue that our current methodological approach failed to clearly delineate local feature processing from configural processing. For instance, in the local task (Supplementary Fig. 1), networks could distinguish class 1 from class 2 based on the distinct letters 'M' and 'T'. However, they might also rely on the spatial arrangement of 'M' with the other three letters to identify class 1, and similarly for 'T' to identify class 2. This highlights the challenge of separating local and configural cues in pattern recognition. Further analysis supports this, as networks trained on the local task demonstrated some ability to generalize to the configural task (Supplementary Fig. 2A). In contrast, those trained on the configural task failed in their generalization to the local task. To address this issue, we implemented a strategy of randomly shuffling the locations of individual features during both training and evaluation phases. This approach ensured that the networks focused purely on the shape of local features without relying on their spatial relationships in the local task. By implementing this shuffling strategy, we were effectively able to differentiate local featural processing from configural processing, as the networks exhibited poor generalization between the local and configural tasks (Supplementary Fig. 2B). Furthermore, when the networks were tested under identical conditions to those used during training, they successfully addressed both local and configural tasks, reaffirming their capability to effectively utilize both local featural and configural cues.

Configural processing is independent of individual local features

Our previous results focused on the generalization capabilities of networks when presented with unseen letter patterns via a one-shot learning paradigm. We further escalated the challenge by assessing network performance on patterns composed of novel letters, specifically A, C, E, F, G, J, L, Q, and Y. Since these letters had not been previously encountered in a pattern by the networks, a decline in performance on the local task was anticipated. The primary objective of this investigation was to determine whether configural processing operates independently of local features or whether it is contingent upon a particular set of local features.

Figure 1C shows the accuracy performance across both tasks when utilizing patterns comprising novel letters. As expected, performance in the local task showed a significant drop; however, it remained above chance levels, suggesting that the network trained on local features could still encode local feature information for categorization, even with unfamiliar letters. Remarkably, performance on the configural task remained nearly unchanged with novel letters, highlighting the independence of configural processing from local features. Similar results were replicated with the shuffling strategy (Supplementary Fig. 3). This result indicates the robust nature of configural cues in maintaining consistent recognition performance under varying viewing conditions, especially in scenarios where local features might be new or degraded.

Configural cues are favored over local featural cues when both are concurrently available

In real-world scenarios where a vision system has access to both local featural and configural information, which cue predominates? To answer this question, we introduced an additional task named the "local plus configural task" (as shown in Fig. 2A), where different classes were characterized by their unique local features and distinct feature configurations. Under these circumstances, the networks could choose to leverage either local features or configural information, likely favoring the more beneficial approach to maximize their performance. Subsequently, the networks were evaluated across the three tasks, i.e., the local, configural, and local plus configural tasks.

Figure 2B presents the networks' performance on the three tasks. When evaluated on the local plus configural task, networks exhibited high accuracy performance as expected. Notably, when the networks were evaluated independently on local and configural tasks, a pronounced superior performance in the configural task was observed. This finding highlights the networks' preference for configural cues, indicating that configural information may offer greater benefits for maintaining consistent performance under various transformation conditions. Additionally, in a reverse approach, networks trained on either the local or the configural task were subsequently assessed using the identical testing regime, namely the local plus configural task. While both networks effectively utilized each feature type, the network trained on the configural task demonstrated slightly more robust performance (Fig. 2C). Taken together, these observations provide evidence that configural information plays an important role in maintaining robust recognition performance across diverse conditions.

To gain deeper insights, we performed an analysis at the single neuron level, examining the pattern of input images that elicited the strongest responses from individual neurons. Specifically, from the local plus configural task image set, we identified the top 20 images that maximally activated each neuron's firing (Fig. 3A). By assessing the categorical consistency across these high-response images per neuron, we could determine whether that neuron exhibited selectivity for local featural or configural cues. This sensitivity was then quantified on a scale from 0 to 1 (details in the "Methods" section). Figure 3B illustrates a histogram representing the distribution of neurons' sensitivities towards local or configural cues across hierarchical layers. Sensitivities were generally low in the early layers but increased in the higher layers, reflecting an enhanced category selectivity of neurons. Notably, neurons in the early layers tended to be more tuned to local features, but this preference gradually transitioned toward configural cues in the later layers. This progression from local to configural cue tuning across layers may explain why configural processing exhibited greater robustness to pixel-level transformations. Note that models trained on the local task showed a consistent bias toward local features across all layers with

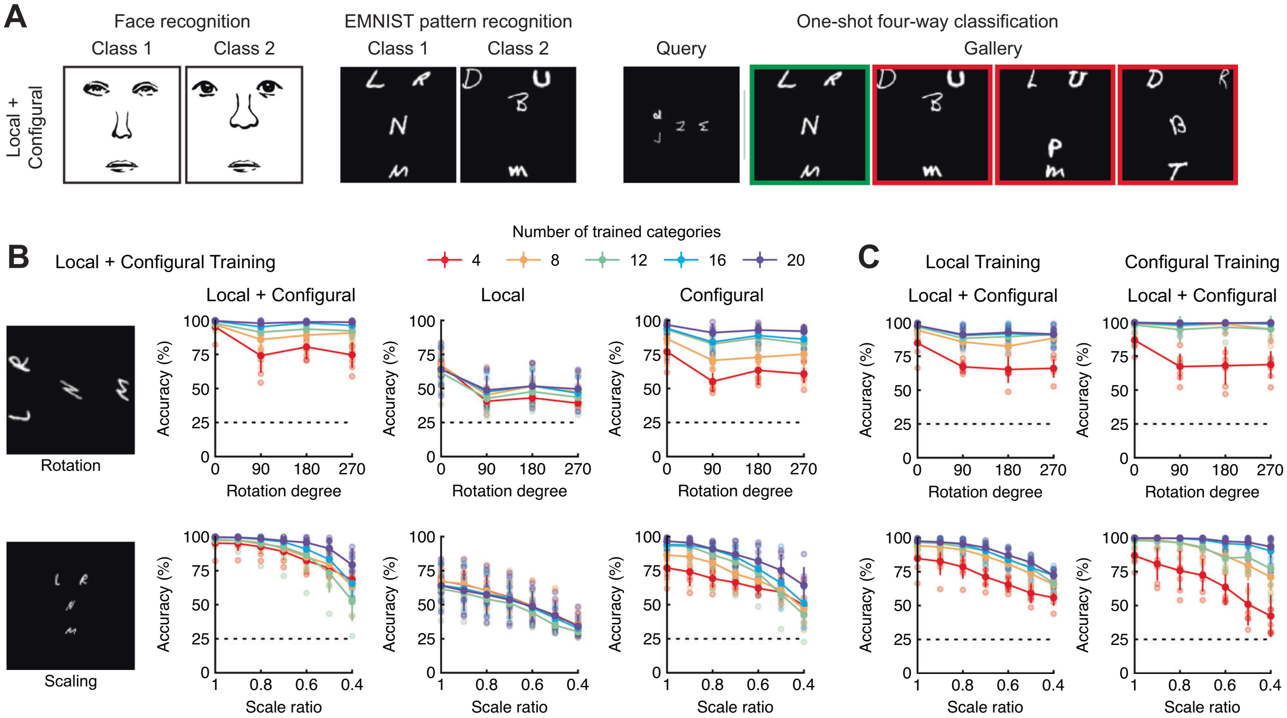

Fig. 2 | Performance accuracy and generalization in the local plus configural task. **A** Illustration of the local plus configural task. **B** Performance accuracy of networks trained on the local plus configural task, when tested on the local plus configural task (left), the local task (middle), and the configural task (right), following the figure conventions described in Fig. 1. Error bars indicate the standard deviation across seven neural networks. **C** Performance accuracy of networks trained on the local task and those trained on the configural task (left and right, respectively), each tested on the local plus configural task.

minimal configural sensitivity, whereas those trained on the configural task demonstrated pronounced configural sensitivity in deeper layers (Supplementary Fig. 4A, B).

To complement the single-neuron analysis, we conducted a representational similarity analysis to examine population-level trends in feature representation. Using test stimuli from the local plus configural task, we constructed a representational similarity matrix (RSM) across all stimuli and compared it to ideal local and configural models via Pearson's correlation (Fig. 3C). Our findings reveal that in the early layers, representations were more similar for stimuli sharing local features, whereas in higher layers, this pattern reversed, with stimuli sharing configural features exhibiting greater similarity. This population-level analysis aligns with the single-neuron findings, highlighting a shift from local to configural feature encoding across hierarchical layers. Taken together, these analyses show a hierarchical shift from local to configural encoding, emphasizing the increasing reliance on configural information for robust feature representation in higher layers.

Impact of network architectures and training loss functions

The local plus configural task's ability to reveal network biases toward local featural or configural cues could provide a framework for analyzing how different network architectures and training strategies influence a network's tendency to favor local or configural processing. Following the approach in Fig. 2B, networks trained on the local plus configural task, with varying architectures and loss functions, were assessed on either the local or the configural task. Here, instead of showing recognition performance across different rotation and scaling levels, we present the area under the performance curves as a function of the training classes.

To begin, we investigated the potential impact of recurrent computations, which play an important role in robust object recognition[35–37] and capturing long-range spatial dependencies[38]. Figure 4A shows the comparison between feedforward and recurrent neural networks concerning their preferences for one type of the cues. The results suggest that both architectures leaned more toward configural cues, without noticeable differences between feedforward and recurrent network types. This observation suggests a minimal role for recurrent computations in enhancing configural processing.

Further analysis focused on the transformer architecture, assessing its bias towards configural rather than local featural cues. Current literature highlights the capability of the transformer architecture in capturing long-range dependencies through attentional modules[39–42], thus we hypothesized a stronger bias to configural cues than in conventional convolutional models. Given the data-intensive nature of training transformers, a comparative analysis between ImageNet pretrained Vision Transformer and ResNet was performed. Consistent with our hypothesis, the transformer architecture demonstrated a stronger bias towards configural cues over local featural ones relative to the convolutional network architecture (Fig. 4B).

Additionally, we evaluated the impact of different training loss functions on a network's featural processing strategy. By contrasting the prototypical loss function with the conventional classification loss function, we observed that networks employing the standard classification loss function exhibited a stronger reliance on configural cues (Fig. 4C). This finding implies that configural processing can be effectively modulated by manipulating the loss function employed during training.

Generalization to real-world face stimuli

Our findings with the EMNIST dataset led us to ask whether our results would generalize to face stimuli. Specifically, we sought to determine if neural networks, when trained on naturalistic facial stimuli such as those in FaceScrub[43] under varying viewing conditions, i.e., rotation and scaling, would have a similar bias toward configural cues. To investigate this, we employed an open-source tool designed for crafting human avatar images[44]. This allowed us to generate two facial sets: one with fixed configural elements and varying local ones and vice versa (see Fig. 5A). The FaceScrub-trained models were then tested using a one-shot five-way classification task,

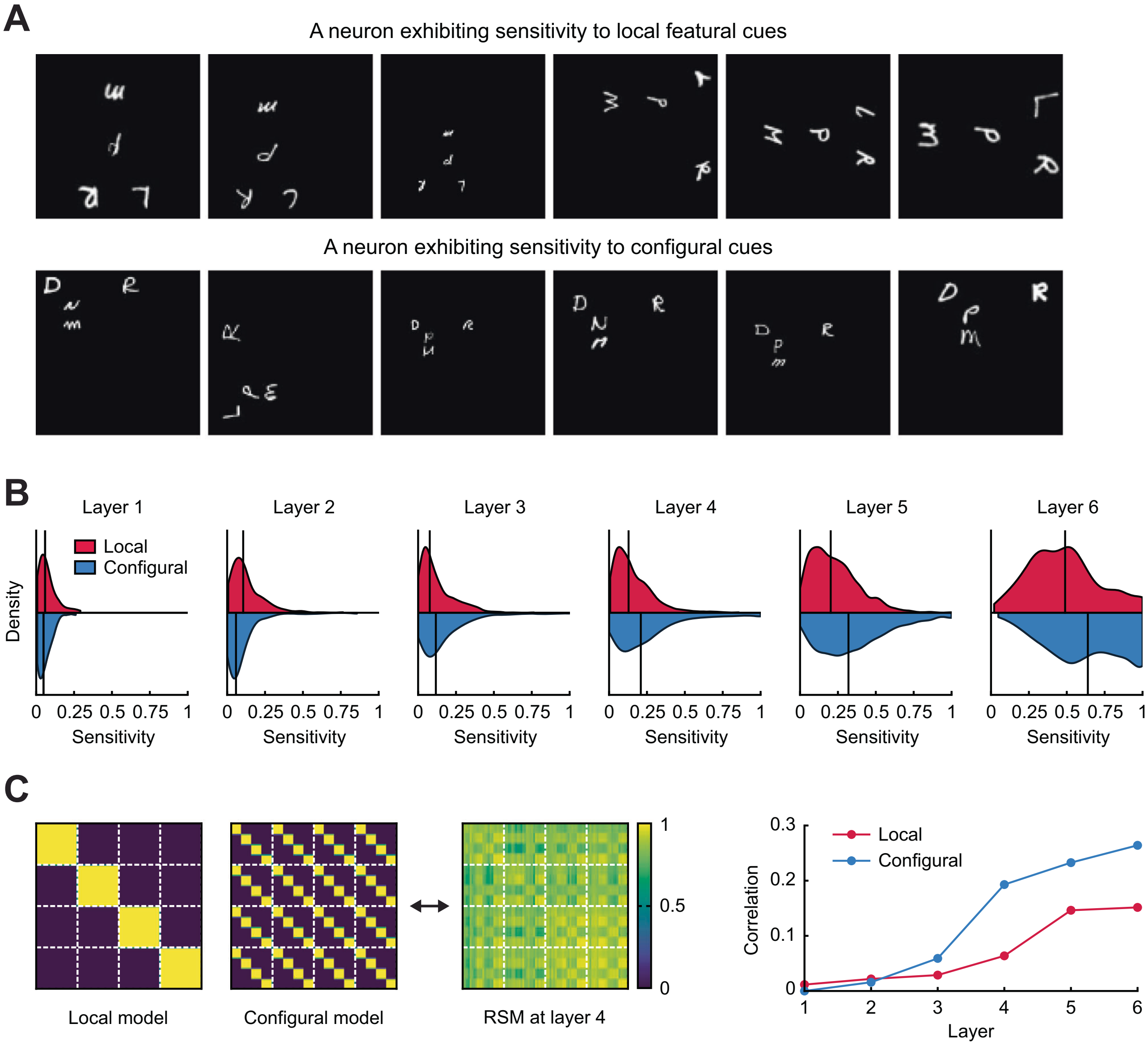

Fig. 3 | Layerwise neuronal sensitivity and representational structure in local and configural processing. **A** Top-6 images selected by a neuron sensitive to local featural cues (top) and another sensitive to configural cues (bottom). **B** Histograms displaying the sensitivity of individual neurons to local (red) and configural (blue) cues across the layers of ResNet50. **C** Representational similarity analysis of ResNet50 trained on the local plus configural stimuli, showing ideal local and configural models (left) and layerwise similarity plots (right). RSM representational similarity matrix.

where each query image was compared against five target classes. The goal of the network was to correctly identify the class of the face stimulus under the different transformations (rotation and scaling). This setup allowed us to examine the networks' ability to recognize faces based on their configural or local features in varying viewing conditions.

As illustrated in Fig. 5B, a distinct pattern emerged between the two face sets. Under rotation or scaling transformations, the networks were notably better at recognizing faces with unique configural aspects over those with distinct local features. This supports our primary hypothesis about the significance of configural cues for robust face recognition. As an additional control experiment, we evaluated the same performance metrics but utilized networks trained on ImageNet. This was aimed at discerning if the bias to configural cues emerges intrinsically from exposure to facial stimuli, or is merely an artifact of the specific face stimuli used here, thus establishing a baseline. Our findings indicated that ImageNet-trained networks did not exhibit a preference for configural cues, further bolstering our original hypothesis. In summation, our computational evidence underscores the pivotal role of configural processing in recognition and the importance of training with classes that require configural processing.

Discussion

In this study, we sought to test the hypothesis that configural processing may play a significant role in enhancing the robustness of visual recognition systems across varying viewing conditions. Although the existing literature offers some evidence supporting this hypothesis[12,13,45], such evidence lacks extensive validation through computational methods. As such, we leveraged deep neural network models to investigate the efficacy of configural cues in recognition by utilizing letter patterns subjected to geometric transformations, such as rotation and scaling.

Our results demonstrate that deep learning models effectively learn to discriminate between categories by leveraging configural cues, when these categories share the same local features (Fig. 1B). Moreover, our data reveal a specialized, independent mechanism for configural processing within these models, separate from the processing of local features (Fig. 1C), accentuating the critical role of configural cues in achieving reliable recognition across

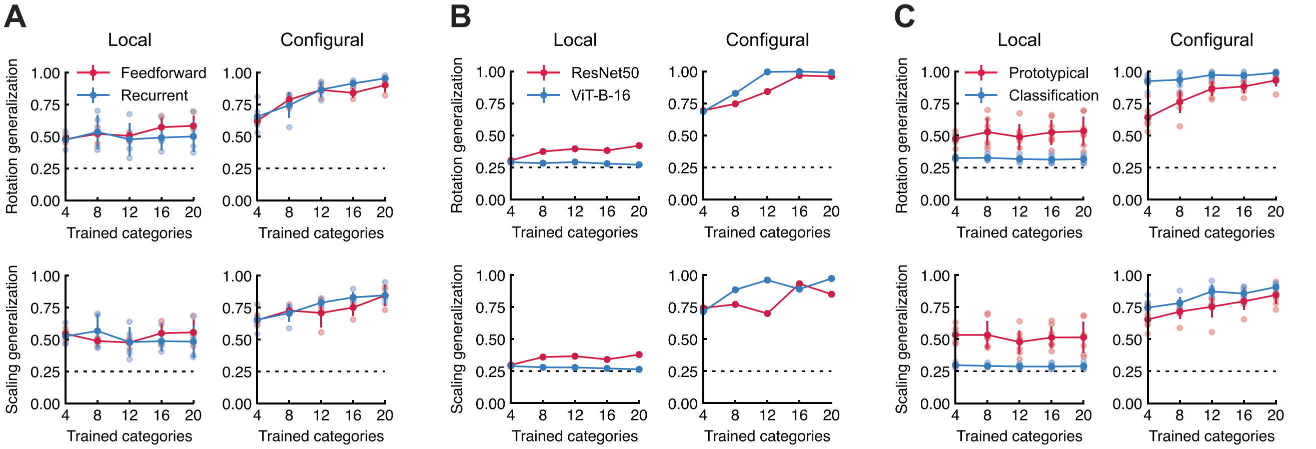

Fig. 4 | Comparative analysis of generalization performance across architectures and learning paradigms. Generalization performance across rotation and scaling transformations was evaluated for different neural network architectures and learning paradigms, including **A** feedforward networks ($n = 3$) and recurrent networks ($n = 4$), with error bars indicating the standard deviation across the networks in each group, **B** ResNet50 and ViT-B-16, and **C** networks trained with prototypical or classification loss functions, with error bars indicating the standard deviation across seven trained neural networks. All networks were trained on the local plus configural task and subsequently tested on the local (left) and configural (right) tasks, with the number of trained categories varying from 4 to 20.

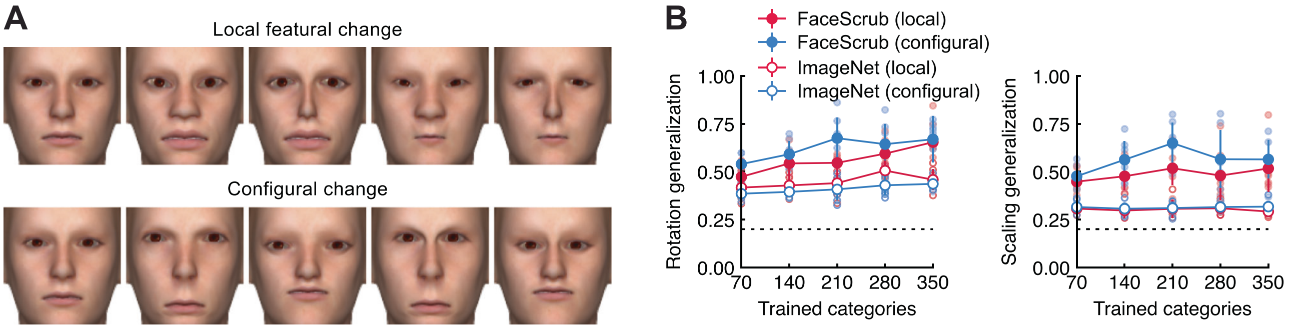

Fig. 5 | Generalization performance in facial stimuli with varying local and configural features. **A** Examples of five facial stimuli are shown in two rows: top with differing local features but identical configurations, and bottom with identical local features but different configurations. **B** Generalization performance comparison across rotation (left) and scaling (right) transformations in networks trained on FaceScrub (filled circles) versus ImageNet (open circles), testing on facial stimuli that either vary in local features but share configurations (red) or share local features but vary in configuration (blue). Error bars indicate the standard deviation across seven neural networks.

environments with varying local features. While previous research[31,32] has shown that networks trained to prioritize global configurations struggled to generalize, exhibiting performance degradation when local components were distorted or showing insensitivity to fragmented global configurations, our study highlights the ability of models to utilize configural cues effectively, particularly under controlled conditions of geometric transformations such as rotation and scaling. These specific transformations were chosen for their simplicity and relevance to naturalistic visual tasks, where objects frequently appear in varying orientations and sizes. Unlike alternative paradigms explored in earlier research, such as fragmented configurations, rotation and scaling preserve global spatial relationships, making them particularly suited for studying the robustness of configural processing in real-world scenarios.

In addition to rotation and scaling, viewpoint invariance is a critical aspect of object recognition, especially in naturalistic contexts where objects are observed from varying angles and perspectives. While our study focused primarily on rotation and scaling, extending these findings to viewpoint invariance is a logical and important next step. Viewpoint invariance has been a central focus in the object recognition literature, but it remains unclear to what extent configural cues contribute to this capability. Future research leveraging 3D stimuli and datasets with controlled viewpoint variations, such as those used in studies on Greeble objects[46], would help to validate and extend the applicability of our findings. By exploring a broader spectrum of transformation conditions, future research can better assess the generality and limitations of configural processing.

The benefits of configural processing are more apparent when networks have access to both configural and local cues under varying viewing conditions (Fig. 2). The preference for configural cues highlights their reliability for achieving optimal performance in such variable environments. This is in line with previous behavioral research, which has shown that configural processing of faces is robust and effective under varying viewing conditions[12,13]. Our additional analyses with face stimuli, crafted to differentiate between local and configural cues, further confirm the consistent benefit of configural processing. Notably, this advantage of configural processing was observed exclusively in neural networks trained specifically for face recognition and was not present in networks trained for general object recognition. This suggests that the preference for configural cues in face recognition networks may evolve as an adaptive response to extensive and varied exposure to facial stimuli, rather than being an innate feature of the face recognition task. This observation proposes an intriguing hypothesis: holistic face processing may have developed as a strategic adaptation to ensure consistent and reliable recognition performance under diverse environmental conditions. This interpretation resonates with a line of research indicating that holistic processing results from extensive experience with stimuli[4,46,47].

While our findings suggest that configural processing strategies may originate from extensive experience with an object under varying viewing conditions, we acknowledge other possibilities. One critical factor influencing the development of configural processing may be the specific task demands associated with recognizing certain object categories, such as faces. Facial recognition, for instance, may emphasize the spatial relationships between facial features, whereas general object categorization tasks, like those in ImageNet, are thought to rely more heavily on local features such as texture. Recent studies have shown that neural network models trained on faces exhibit more holistic processing compared to networks trained on non-face categories[48,49]. From this perspective, task demands may explain why networks trained on the FaceScrub dataset exhibited a pronounced configural processing bias across transformation levels, while ImageNet-trained networks did not, as shown in Fig. 5. To examine whether task demands alone drive configural processing or if extensive visual experience is also necessary, we conducted a supplementary analysis. Networks were trained exclusively on upright, full-scale face images from the FaceScrub dataset and subsequently tested on rotated and scaled faces. Under rotation conditions, the networks showed limited generalization and did not display a clear preference for configural cues over local features, closely aligning with the performance of ImageNet-trained networks (Supplementary Fig. 5). In contrast, under scaling transformations, FaceScrub-trained networks demonstrated still superior generalization and a strong preference for configural cues. These findings indicate that while task demands influence processing strategies, they are insufficient to establish robust, generalizable configural processing, particularly under rotation conditions.

Beyond visual experience and task demands, developmental trajectories may also play a crucial role in the emergence of configural processing strategies. Clinical research presents a hypothesis that early limitations in visual acuity due to retinal immaturities might contribute to the development of configural face processing. Empirical evidence from individuals with congenital cataracts, who typically do not experience blurred vision initially, has revealed significant deficits in configural face processing in later stages of life[8,50]. Moreover, recent computational investigations have highlighted the potential developmental advantages of this initial blurred vision phase in enhancing integrative face processing[51,52]. Future research should aim to disentangle the contributions of developmental factors, task demands, and diverse visual experiences to provide a more comprehensive understanding of the origins of configural processing.

What underlies the superiority of configural processing relative to local processing within neural networks? To address this question, we conducted a detailed examination at the unit and population levels of the network. We discovered that each unit was specialized to respond either to configural cues, local cues, or both, with this specialization varying by layer. Specifically, units in the lower layers were more responsive to local cues, while those in upper layers were more attuned to configural cues. At the population level, a similar trend was observed, with representational similarity shifting from local to configural features across layers. This layer-specific responsiveness suggests a hierarchical processing approach, where local and configural cues are handled at different stages of visual perception. Neuroscientific evidence supports this hierarchy, indicating that the fusiform face area is principally involved in configural processing, whereas earlier cortical stages focus on local feature processing[4,53–55]. We propose that configural cues processed in later stages are inherently less affected by local pixel variations, thereby enhancing recognition stability across diverse conditions.

One could argue that the inherently low variability of configural cues might also be a factor influencing the networks' preference for configural processing. To address this possibility and control for potential confounds, we introduced jitter to the positions of individual letters within the configurations. We found that, although this modification slightly reduced the networks' reliance on configural cues, they continued to leverage these cues effectively, underscoring their robust nature (Supplementary Fig. 6). While we do not rule out the possibility that the inherently low variability of configural cues contributes as a key factor, our findings suggest that even when variability is introduced, the structured patterns of spatial relationships in configural cues continue to provide both stability and reliability for recognition.

Are recurrent computations required for effective configural processing? While one might hypothesize that the integrative nature of recurrent computations is crucial for combining local features into holistic representations, our data do not support this view. We observed no significant difference in the processing of local versus configural information between recurrent and non-recurrent network architectures. This observation is corroborated by our layerwise analysis which demonstrates consistent unit preferences throughout the network layers, irrespective of recurrent cycles (Supplementary Fig. 7). Therefore, our results suggest that configural processing is likely driven by hierarchical computations rather than by recurrent dynamics, aligning with some studies suggesting that holistic processing may not depend on top-down attention[56,57]. Nevertheless, recurrent computations may play a more prominent role under conditions not explored in our study. Freiwald and Tsao[58], for instance, demonstrated the importance of recurrence in achieving view-invariant face representations, especially under conditions with significant variability in perspective. While our findings indicate that feedforward architectures are sufficient for addressing simpler transformations like scaling and rotation, the contribution of recurrence to configural processing in such more complex scenarios remains an open question.

Furthermore, our study indicates that the choice of network architectures and loss functions can impact a network's preference for local or configural cues. We showed that vision transformer models, known for their efficiency in capturing long-range dependencies[39–42], exhibited enhanced configural processing capabilities compared to convolutional neural networks. The transformer's self-attention mechanism allows for broad-scale integration of spatial cues, potentially facilitating more efficient configural visual information processing. Additionally, networks trained with prototypical loss functions, which focus on minimizing the distance to class prototypes, tend to favor local featural processing. Future investigations will be critical in elucidating the interplay between local and configural processing, thereby advancing the design of more robust and adaptable neural network models.

We also explored the effects of varying the number of training categories. An increase in category variety was found to improve the network's resilience to changes in rotation and scale, corroborating earlier research[34] that underlines the benefits of diverse category experiences for network robustness. Unlike the previous work, however, the current method assessed network performance using entirely novel categories not previously exposed to transformations within a one-shot learning framework, thereby removing potential category selection bias associated with prior exposure to the transformations.

In summary, this study provides valuable insights into the intersection of deep learning and psychological research, particularly in understanding configural processing as a robust strategy for visual recognition. By employing a reductionist approach, the study isolates and examines fundamental principles underlying human configural processing within a controlled and interpretable framework. This approach enables the systematic assessment of computational models that effectively leverage both local and configural cues for recognition. While the current methodology is rooted in controlled experimental settings, its findings establish a strong foundation for extending these models to more complex and naturalistic environments. By bridging these insights with real-world contexts, this line of inquiry moves closer to uncovering the fundamental mechanisms behind the remarkable efficiency and adaptability of human vision.

Methods

Visual stimulus set of letter patterns

In this study, we constructed a novel set of visual stimuli composed of patterns. Each category included four letters arranged into a composite pattern. To introduce variability within a single letter, we collected 500 distinct instances of each letter from the EMNIST database[33]. This diversity was crucial not only for providing a sufficient dataset for training neural network models but also for reflecting the natural variability in the appearance of individual features.

To systematically investigate the role of local and configural processing in recognition, we proposed two distinct tasks, as illustrated in Fig. 1A: the "local" and the "configural" tasks. In the local task, different categories shared identical configurations of individual letters but differed in their unique combinations of letters. We chose nine letters—B, D, L, M, N, P, R, T, and U—to generate a total of 24 distinct patterns. Conversely, in the configural task, while each category utilized an identical set of letters (L, R, N, and M), their configurations were unique. To prevent any overlap in the configurations, we segmented the entire input image (100 by 100 pixels) into 25 sections (20 by 20 pixels each) in a grid-like formation. Out of these, four were selected to create a total of 24 patterns. Details of the 24 categories are provided in Supplementary Fig. 1.

Additionally, to investigate whether configural processing is established on specific local features or is independent of them, we devised an additional stimulus set. This set was analogous to the original one but employed a different selection of nine letters—A, C, E, F, G, J, L, Q, and Y—for the local task. Similarly, for the configural task, we utilized distinct configural patterns of the same four letters, namely A, C, E, and F. This new set of stimuli was only utilized for evaluation purposes and was not incorporated into the training process.

Beyond the local and configural tasks, we introduced an additional composite stimulus set termed the "local plus configural task". This new task merged a category from the local task with one from the configural task, so that each category in this task presented both an exclusive set of local features and a distinct configuration. This experimental design was implemented to investigate which type of cue, local or configural, networks would prioritize when faced with both options simultaneously. Details and examples of the 24 categories are provided in Supplementary Fig. 1.

Neural network architectures

Our study conducted a comprehensive evaluation of various neural network architectures to verify the robustness and generalizability of our findings. This evaluation encompassed three feedforward convolutional neural networks: ResNet18, ResNet34, and ResNet50[59]. Concurrently, we analyzed four recurrent convolutional neural networks: CORnet-S[60], BLnet, BLTnet[36], and ConvLSTM[61]. For those recurrent models that incorporated batch normalization layers, we substituted these with layer normalization layers to mitigate potential batch effects. Each recurrent model underwent five recurrent cycles to allow for iterative integration of information. Detailed descriptions and methodologies of these models can be found in their respective original publications. Additionally, our investigation extended to vision transformers, specifically Vit-B-16[39]. Due to the extensive training data requirements of this model, we employed a version of Vit-B-16 pretrained on the ImageNet dataset[62]. This model was compared with a similarly pretrained version of ResNet50 on ImageNet, to explore their relative performances in the local and configural tasks.

In the "Results" section, Figs. 1, 2, and 5 present results averaged across all seven convolutional neural networks, including three feedforward and four recurrent models. Figure 3 focuses specifically on ResNet50. Figure 4 compares feedforward and recurrent networks as well as ResNet50 and the ViT-B-16 model.

Training procedures

The networks underwent a two-stage training process using PyTorch, version 2.0.1. Initially, they were pre-trained on the classification of a single letter to develop robust low-level representations. This stage provided 2468 samples per letter, with the training images resized to 20 × 20 pixels and randomly positioned within a larger input field of 100 × 100 pixels. The training was conducted using the stochastic gradient descent optimization algorithm for 200 epochs, with a fixed learning rate of 0.001, a batch size of 64, a weight decay of 0.0001, and a momentum value of 0.9.

Following the initial pre-training, the networks were further optimized to recognize patterns of letters in one of the specific stimulus sets: the local, configural, or local plus configural tasks. In this subsequent training phase, the input patterns experienced a combination of transformations, including translation within the input field, rotation, and scaling. The rotation involved altering the orientation of the entire input pattern by 0, 90, 180, and 270 degrees, with all constituent elements of the pattern being rotated simultaneously. Scaling changed the size of the original 100 × 100 pixel input images, with ratios progressively reducing from a full scale to 0.4.

The training followed the framework introduced by Snell et al.[63], where the model learns an embedding space to classify samples based on proximity to class prototypes. During training, episodes simulated few-shot tasks by randomly selecting $N = 4$ classes per episode, with each class represented by $n_s = 1$ support example and $n_q = 1$ query example. Class prototypes were calculated as the mean of the support embeddings:

$$\mathbf{p}_c = \frac{1}{n_s}\sum_{i=1}^{n_s} f_\theta\left(\mathbf{x}_{c,i}\right), \tag{1}$$

where f_θ denotes the embedding function, and $\mathbf{x}_{c,i}$ represents the i-th support example of class c. Query embeddings $f_\theta\left(\mathbf{x}_q\right)$ were classified based on the shortest Euclidean distance to prototypes:

$$d(\mathbf{x}_q, \mathbf{p}_c) = \| f_\theta\left(\mathbf{x}_q\right) - \mathbf{p}_c\|_2^2, \tag{2}$$

with class predictions made by minimizing this distance:

$$\hat{y}_q = \arg\min_c d(\mathbf{x}_q, \mathbf{p}_c). \tag{3}$$

The prototypical loss was designed to optimize the embedding space by minimizing distances to the correct prototype and maximizing distances to others:

$$L = -\frac{1}{Nn_q}\sum_{q=1}^{Nn_q} \log \frac{\exp\left(-d(\mathbf{x}_q, \mathbf{p}_{y_q})\right)}{\sum_{c=1}^{N} \exp\left(-d(\mathbf{x}_q, \mathbf{p}_c)\right)}. \tag{4}$$

The embedding network, pre-trained on single-letter classification, was fine-tuned with this loss function over episodes containing transformed stimuli (i.e., rotation and scaling). The network generalized to unseen classes during evaluation, where $n_s = 1$ support example per class was provided for one-shot classification.

Training was conducted over 25,000 episodes using the stochastic gradient descent algorithm, with a fixed learning rate of 0.00001, a weight decay of 0.0001, and a momentum of 0.9. For each network, the final layer was modified to output 512 embedding vectors, which were then used to compute the prototypical loss.

Additionally, the present study explored the impact of the number of training categories on network robustness, expanding on recent findings by Jang et al.[34]. We systematically varied the number of training categories from 4 to 20, reserving the remaining four of 24 total categories for evaluation. Importantly, unlike the previous work, the current approach offered a legitimate assessment of invariance to transformations on novel stimuli, avoiding potential bias where invariant features might be associated with specific categories.

Layerwise neural sensitivity analysis

In this analysis, the goal was to assess the sensitivity of individual neurons to local and configural cues. To achieve this, we selected the 20 images that elicited the strongest response from each neuron, using stimuli from the local plus configural task. We then evaluated the neuron's sensitivity to either local or configural cues by examining the consistency of categorical selections among these 20 images. This was quantified using entropy, which measures the uncertainty or randomness of these categorical selections.

Specifically, $p(x_i)$ represents the probability of a particular input feature x belonging to category i across the top 20 images that elicited maximum response from a given neuron. It is calculated as:

$$p(x_i) = \frac{n(x_i)}{\sum_j n(x_j)}, \tag{5}$$

where $n(x_i)$ is the frequency of input features x belonging to category i, and the denominator is the total frequency of all features across these top 20 images. The sensitivity metric S is then computed as:

$$S(X) = 1 - H(X)/\log_2 N, \tag{6}$$

where $H(X) = -\sum_{i=1}^{N} p(x_i)\log_2 p(x_i)$ represents the entropy and N is the number of categories (4 in our analysis). The sensitivity measure ranges from 0 to 1. A value of 1 signifies maximum sensitivity, where the neuron consistently identifies all 20 top images as belonging to a single category. This indicates that the neuron is highly specialized or tuned to the local or configural cues characterizing that specific category. Conversely, a value of 0 means that the neuron exhibites complete insensitivity, with its selections from the 20 top images showing no discernible pattern or preference for any specific category, suggesting a lack of sensitivity to either local or configural cues.

This layerwise neural sensitivity analysis was conducted on the ResNet50 model across five scenarios: (1) a model trained solely on the local task, (2) a model trained solely on the configural task, (3) a model trained on the combined local plus configural task, (4) a model trained on the FaceScrub dataset, and (5) a model trained on ImageNet. The findings from these analyses are illustrated in Fig. 3B and Supplementary Fig. 4. Additionally, the analysis was extended to examine sensitivity patterns across network layers and through recurrent cycles of ConvLSTM, as shown in Supplementary Fig. 7.

Visual stimulus set of faces

We sought to test whether the configural processing bias observed with the EMNIST dataset could be also applicable to facial recognition tasks using natural face images. To this end, we trained neural networks on the FaceScrub database[43], which included facial images altered through rotation and scaling, following the same training procedures as those used for EMNIST but using a one-shot five-way learning paradigm instead. The number of identities varied systematically across training sets, including 70, 140, 210, 280, and 350. Each network model was trained using a fixed learning rate of 0.00001 through stochastic gradient descent, incorporating a batch size of 64, weight decay of 0.0001, and momentum of 0.9. Additionally, networks were also trained on ImageNet, using 350 object categories selected at random from a set of 1000, to serve as a baseline for comparison.

To determine the local or configural processing bias of neural networks trained with real-world face images, we utilized MakeHuman[44] (http://www.makehumancommunity.org/), a tool for crafting human avatars, to generate specific sets of facial stimuli. Similar to the original approach, we developed two distinct types of stimuli: local and configural. For local feature recognition, we created faces with the same overall configuration but altered specific local features. More specifically, starting with an average face[64], modifications were made in aspects such as the scaling of right and left eyes, positioning of the eye corners, and mouth scaling both horizontally and vertically, ended up with four distinct facial stimuli (as seen in the top panel of Fig. 5A). Conversely, for the configural task, while keeping local features constant, we varied the overall configuration of facial elements. This included adjustments in the positions of both eyes (inward/outward and upward/downward), along with changes in the nose and mouth configurations (shown in the bottom panel of Fig. 5A). To introduce more variations in the facial images for each condition, we altered the yaws (from −5° to 4°) and pitches (from −5° to 4°), and adjusted the lighting conditions with three different light sources along the x-, y-, and z-axes. These sets of stimuli were used in evaluations of both face- and ImageNet-trained networks.

Statistics and reproducibility

All computations and visualizations were conducted using the Python scientific computing framework. Neural network models were trained and evaluated using PyTorch (version 2.1.1). The ResNet and ViT models were obtained from the Torchvision library (version 0.16.1). Other models, including CORnet-S, BLNet, BLTNet, and ConvLSTM, were either implemented from scratch or adapted from their original sources with modifications. These models are publicly available on the Open Science Framework (OSF) at https://osf.io/htduf/[65]. All visualizations were generated using Seaborn (version 0.11.2) and Matplotlib (version 3.5.2).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All stimulus sets employed in this study are available on the Open Science Framework[65]. The source data for the main figures are available in the Supplementary Data.

Code availability

The codes for training and analysis are available on the Open Science Framework [65].

Received: 10 September 2024; Accepted: 4 February 2025;
Published online: 07 March 2025

References

1. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115 (1987).
2. Tanaka, J. W. & Taylor, M. Object categories and expertise: Is the basic level in the eye of the beholder? *Cogn. Psychol.* **23**, 457–482 (1991).
3. Gauthier, I., Skudlarski, P., Gore, J. C. & Anderson, A. W. Expertise for cars and birds recruits brain areas involved in face recognition. *Nat. Neurosci.* **3**, 191–197 (2000).
4. Gauthier, I. & Tarr, M. J. Unraveling mechanisms for expert object recognition: bridging brain activity and behavior. *J. Exp. Psychol.* **28**, 431 (2002).
5. Young, A. W., Hellawell, D. J. & Hay, D. C. Configurational information in face perception. *Perception* **16**, 747–759 (1987).
6. Tanaka, J. W. & Farah, M. J. Parts and wholes in face recognition. *Q. J. Exp. Psychol. Sect. A* **46**, 225–245 (1993).
7. Maurer, D., Le Grand, R. & Mondloch, C. J. The many faces of configural processing. *Trends Cogn. Sci.* **6**, 255–260 (2002).
8. Le Grand, R., Mondloch, C. J., Maurer, D. & Brent, H. P. Early visual experience and face processing. *Nature* **410**, 890–890 (2001).
9. Farah, M. J., Wilson, K. D., Drain, M. & Tanaka, J. N. What is" special" about face perception? *Psychol. Rev.* **105**, 482 (1998).
10. Goffaux, V., Gauthier, I. & Rossion, B. Spatial scale contribution to early visual differences between face and object processing. *Cogn. Brain Res.* **16**, 416–424 (2003).
11. Goffaux, V. & Rossion, B. Faces are" spatial"–holistic face perception is supported by low spatial frequencies. *J. Exp. Psychol.* **32**, 1023 (2006).
12. McKone, E. Configural processing and face viewpoint. *J. Exp. Psychol.* **34**, 310 (2008).
13. McKone, E. Holistic processing for faces operates over a wide range of sizes but is strongest at identification rather than conversational distances. *Vis. Res.* **49**, 268–283 (2009).
14. Piepers, D. W. & Robbins, R. A. A review and clarification of the terms “holistic,” “configural,” and “relational” in the face perception literature. *Front. Psychol.* **3**, 559 (2012).
15. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).

16. Yamins, D. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
17. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
18. Yamins, D. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
19. Güçlü, U. & Van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
20. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
21. Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl Acad. Sci. USA* **115**, E9015–E9024 (2018).
22. Jang, H. & Tong, F. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nat. Commun.* **15**, 1989 (2024).
23. Hill, M. Q. et al. Deep convolutional neural networks in the face of caricature. *Nat. Mach. Intell.* **1**, 522–529 (2019).
24. Grossman, S. et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
25. Higgins, I. et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* **12**, 6456 (2021).
26. Jozwik, K. M. et al. Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proc. Natl Acad. Sci. USA* **119**, e2115047119 (2022).
27. O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q. & Chellappa, R. Face space representations in deep convolutional neural networks. *Trends Cogn. Sci.* **22**, 794–809 (2018).
28. O'Toole, A. J. & Castillo, C. D. Face recognition by humans and machines: three fundamental advances from deep learning. *Annu. Rev. Vis. Sci.* **7**, 543–570 (2021).
29. Geirhos, R. et al. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Int. Conf. Learn. Repr.* https://doi.org/10.48550/arXiv.1811.12231 (2019).
30. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
31. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Local features and global shape information in object classification by deep convolutional neural networks. *Vis. Res.* **172**, 46–61 (2020).
32. Baker, N. & Elder, J. H. Deep learning models fail to capture the configural nature of human shape perception. *Iscience* **25**, 104913 (2022).
33. Cohen, G., Afshar, S., Tapson, J. & Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, 2921–2926 (IEEE, 2017).
34. Jang, H. et al. Robustness to transformations across categories: is robustness driven by invariant neural representations? *Neural Comput.* **35**, 1910–1937 (2023).
35. Wyatte, D., Curran, T. & O'Reilly, R. C. The limits of feedforward vision: recurrent processing promotes robust object recognition when objects are degraded. *J. Cogn. Neurosci.* **24**, 2248–2261 (2012).
36. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 278016 (2017).
37. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
38. Sundaram, S., Sinha, D., Groth, M., Sasaki, T. & Boix, X. Recurrent connections facilitate symmetry perception in deep networks. *Sci. Rep.* **12**, 20931 (2022).
39. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. Learn. Repr.* https://doi.org/10.48550/arXiv.2010.11929 (2021).
40. Naseer, M. M. et al. Intriguing properties of vision transformers. *Adv. Neural Inf. Process. Syst.* **34**, 23296–23308 (2021).
41. Tuli, S., Dasgupta, I., Grant, E. & Griffiths, T. L. Are convolutional neural networks or transformers more like human vision? In *Proc. Annual Meeting of the Cognitive Science Society*, Vol 43 https://doi.org/10.48550/arXiv.2105.07197 (2021).
42. Mao, X. et al. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 12042–12051 (2022).
43. Ng, H.-W. & Winkler, S. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, 343–347 (IEEE, 2014).
44. Bastioni, M., Re, S. & Misra, S. Ideas and methods for modeling 3d human figures: the principal algorithms used by makehuman and their implementation in a new approach to parametric modeling. In *Proceedings of the 1st Bangalore annual compute conference*, 1–6 (2008).
45. Jarudi, I. N. et al. Recognizing distant faces. *Vis. Res.* **205**, 108184 (2023).
46. Gauthier, I. & Tarr, M. J. Becoming a "greeble" expert: exploring mechanisms for face recognition. *Vis. Res.* **37**, 1673–1682 (1997).
47. Diamond, R. & Carey, S. Why faces are and are not special: an effect of expertise. *J. Exp. Psychol.* **115**, 107 (1986).
48. Tong, F. & Jang, H. Convolutional neural networks optimized for face recognition reveal a computational basis for holistic face processing. *J. Vis.* **22**, 4185–4185 (2022).
49. Yovel, G., Grosbard, I. & Abudarham, N. Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proc. R. Soc. B* **290**, 20230093 (2023).
50. Grand, R. L., Mondloch, C. J., Maurer, D. & Brent, H. P. Impairment in holistic face processing following early visual deprivation. *Psychol. Sci.* **15**, 762–768 (2004).
51. Vogelsang, L. et al. Potential downside of high initial visual acuity. *Proc. Natl Acad. Sci. USA* **115**, 11333–11338 (2018).
52. Jang, H. & Tong, F. Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *J. Vis.* **21**, 6–6 (2021).
53. Yovel, G. & Kanwisher, N. G. The neural basis of the behavioral face-inversion effect. *Curr. Biol.* **15**, 2256–2262 (2005).
54. Schiltz, C. & Rossion, B. Faces are represented holistically in the human occipito-temporal cortex. *Neuroimage* **32**, 1385–1394 (2006).
55. Liu, J., Harris, A. & Kanwisher, N. Perception of face parts and face configurations: an fmri study. *J. Cogn. Neurosci.* **22**, 203–211 (2010).
56. Boutet, I., Gentes-Hawn, A. & Chaudhuri, A. The influence of attention on holistic face encoding. *Cognition* **84**, 321–341 (2002).
57. Norman, L. J. & Tokarev, A. Spatial attention does not modulate holistic face processing, even when multiple faces are present. *Perception* **43**, 1341–1352 (2014).
58. Freiwald, W. A. & Tsao, D. Y. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851 (2010).
59. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).
60. Kubilius, J. et al. Brain-like object recognition with high-performing shallow recurrent anns. *Adv. Neural Inf. Process. Syst.* **32**, 12805–12816 (2019).

61. Shi, X. et al. Convolutional lstm network: a machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **28**, 802–810 (2015).
62. Deng, J. et al. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255. https://api.semanticscholar.org/CorpusID:57246310 (2009).
63. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **30**, 4080–4090 (2017).
64. Hays, J., Wong, C. & Soto, F. A. Faret: a free and open-source toolkit of three-dimensional models and software to study face perception. *Behav. Res. Methods* **52**, 2604–2622 (2020).
65. Jang, H., Sinha, P. & Boix, X. Configural processing as an optimized strategy for robust object recognition in neural networks [dataset]. *Open Science Framework*. https://osf.io/htduf/ (2024).

Acknowledgements

This research was partially supported by grants from Korea University (K2404751) and the National Research Foundation of Korea, funded by the Korean government (RS-2024-00451866). Additional support was provided by Fujitsu Research and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100009. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Author contributions

H.J. and X.B. conceived of and designed the study, H.J. trained neural network models and performed all data analyses, H.J., P.S., and X.B. wrote the paper together.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07672-1.

Correspondence and requests for materials should be addressed to Hojin Jang or Xavier Boix.

Peer review information *Communications Biology* thanks Connor Parde and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jasmine Pan.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.