Leveraging Vision Transformers to Propose a Context-Dependent Computational Mechanism for the Holistic Process of Faces

Srijani Saha (srijanisaha@g.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA

Talia Konkle (talia_konkle@harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA

George Alvarez (alvarez@wjh.harvard.edu)
Department of Psychology, Harvard University, Cambridge, MA, USA

Abstract

Holistic processing, or the integration of facial features to build an identity representation, has offered a clever solution to a critical problem - how can we seamlessly tell apart faces when there is little inter-class variability? While seminal work in psychology has demonstrated the behavioral consequences of holistic processing where individual features and identities appear different as a result of the facial context (e.g., the Composite Face Effect, Thatcher Illusion), it has been difficult to identify a computational mechanism that operationalizes these context-dependent perceptual effects. Here, we leverage the vision transformer's architecture to show how local perturbations in a face can update the representations of other face features, thereby affecting the identity representation. The interactions between the perturbed feature and the context updates the representation of the unchanged facial features and identity, the latter towards a different, new identity. The shift in identity primarily occurs when the local changes are naturalistic.

Introduction

Holistic face processing describes the well-known phenomenon where individual face parts are integrated into a unified percept of a face that goes beyond an independent sum of the parts (Tanaka & Farah, 1993; Young et al., 1987). A compelling demonstration occurs when changing a single feature—such as replacing a nose—shifts the perceived identity of the entire face, even though other features remain physically unchanged. This "part-whole effect" demonstrates, how a small local change can alter the overall appearance and identity of a face (Tanaka & Sengco, 1997; Rossion, 2013). However, it is unclear whether the change to a local feature (e.g., the nose) alters representations of other local features, or only the overall identity, and what mechanisms would support such contextual modulation.

Vision transformers (ViTs) provide a novel model architecture to explore potential computational mechanisms of holistic processing. Unlike CNNs that primarily capture local patterns, transformers employ self-attention mechanisms that allow for interactions between all image regions simultaneously (Dosovitskiy et al., 2021). Further, their patch-based architecture enables precise tracking of how local perturbations propagate through the network to influence local (patch-wise) representations, as well as overall global representations (e.g., the class identity token). Here, we systematically manipulate facial features within a single local patch, and track how this information influences the representation of the unchanged parts of the face and the "perceived" identity. To what extent do these models show signatures of holistic processing strategies?

Methods

We used a ViT-B32 pretrained on VGG Face 2 (Rodrigo et al., 2024) for its high face identification accuracy. Nine identities from CelebAMask-HQ (Liu et al., 2015; Karras et al., 2018; Lee et al., 2020) were selected, all with moderate face-classification confidence to allow measurable shifts in identity representation.

We next created a parametric sweep of images for each target identity, where we increasingly perturbed the nose. All nose modifications were constrained to a single patch. We created three kinds of sweeps: (1) we parametrically expanded noses at five levels, using MaskGAN (Lee et al., 2020), which generates naturalistic feature modifications; (2) we added Gaussian noise to the nose region; and (3) we scrambled the pixels in the nose region. The latter two control conditions were carefully calibrated to match the expanded nose manipulation, as measured by the change in embedding in the initial patch-embedding layer (before self-attention layers), ensuring that any differences in model behavior would be attributable to the type of modification (naturalistic vs. noise) rather than the magnitude of change. For each level of nose expansion, we created 7-10 matched versions of these control conditions to ensure robust comparison. Example images from these conditions are shown in **Figure 1**.

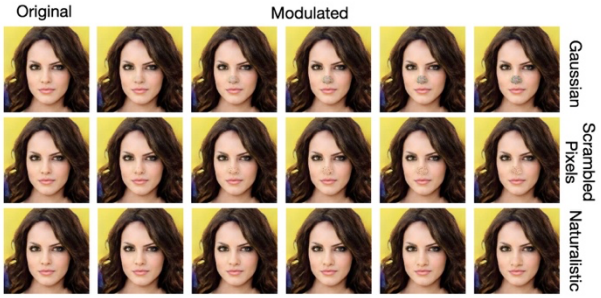

Figure 1: Example of 5 levels of modulations for a selected image from the CelebAMask-HQ dataset by adding gaussian noise, scrambling pixels and expanding nose within the nose patch

To investigate holistic processing in vision transformers, we conducted two analyses. First, we passed modified images (expanded nose, gaussian noise, or scrambled pixels) through ViT-B32 and measured confidence in target identity. To isolate contextual modulation effects, we used "model optogenetics": recording activations from modified images, then replaying them with modulated contextual patches but original nose-patch embedding at each step. At each stage, the class token aggregates information across patches, receiving original nose-patch embeddings but other patch-embeddings from the modulated image. This removes local effects of the perturbed patch at each processing stage - any identity confidence change must arise from how context was affected by interactions with the modified nose.

Results

The key results are shown in **Figure 2**. The difference in SoftMax probability of the target identity is plotted for the 5 parametric levels with increasing noise perturbations. In both analyses, we observe a substantial decrease in identity confidence for the natural nose-width manipulation across all modification types, indicating reduced confidence that the face belonged to the original person (area under the curve or AUC change=1.60 in Fig. 2A and 1.81 in Fig. 2B). This effect was significantly less pronounced in the noise conditions (Gaussian AUC = 3.44 in Fig. 2A and 3.47 in Fig 2B; Scrambled AUC= 3.39 in Fig. 2A and 3.45 in Fig. 2B).

Indeed, for natural modulations, by the third distortion level, nearly all the images are classified as a new identity; this same identity confusion was not true for the noisey-nose conditions. When modified-context, unmodified-nose activations were replayed (**Fig. 2B**), a similar trend was maintained, with more identity confusion for the natural modulation than in the noisy-nose conditions. Thus, the impact of the nose expansion on the identity can be accounted for nearly fully by the impact the nose expansion has on other patches. Together these results provide strong evidence for contextual modulation between local features (the eyes and mouth representations are affected by the nose representation) combining to modify the global identity representation.

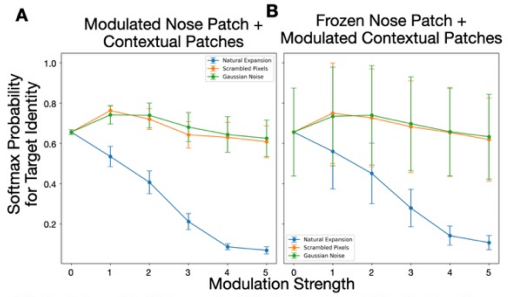

Figure 2 Trajectories of SoftMax probability for the target identity (n = 9) as a result of different types of modulation of the nose patch (natural expansion, scrambled pixels, adding gaussian noise) at different strengths. In **A** the entire modified image was passed through the ViT_B32 while in **B** the nose patch in the modified image was replaced with the original nose patch at every transformer block in the ViT_B32

We also directly examined how the representation of the patch embeddings changed in the context of nose-patch modulation. We found that the patch embeddings outside the nose had different embeddings in the context of the altered-nose patch. In other words, the very *same pixels outside the nose* "appear" different to the model induced by the altered-nose patches. This provides an intriguing signature of holistic processing, where the local features actually indeed 'look different' to the model due to self-attentional effects from other patches.

Discussion

Our findings reveal that vision transformers have signatures of holistic face processing: unchanged features are represented differently depending on facial context, with particularly strong shifts in identity when changes are naturalistic. The self-attention mechanism may parallel neural processes that integrate facial features into coherent percepts in human visual systems. This work provides a testable computational account of holistic processing and could help investigate other perceptual phenomena like the Composite Face Effect through systematic feature manipulations and may help to elucidate the mechanisms of holistic integration in human perception.

Acknowledgements

This work was supported by NSF PAC COMP-COG 1946308 to GAA.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *In International Conference on Learning Representations.*

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. *In International Conference on Learning Representations.*

Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5549-5558).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730-3738).

Rodrigo, M., Cuevas, C., & García, N. (2024). Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks. *Scientific Reports*, *14*(1), 21392.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139-253.

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.

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

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747-759.