Emergence of Interpretable Functional Specialization in Neural Networks Trained on Facial Expression and Identity Recognition

Yanis Tazi [1 2] **Michael Berger** [2] **Winrich A. Freiwald** [2 3]

Abstract

Facial expression and identity recognition are essential cognitive processes underpinning daily life and social relationships. Despite their importance, the biological neural mechanisms and brain regions associated with these processes are yet to be fully understood. Convolutional Neural Networks (CNNs), a reasonable model of the biological visual system, are commonly employed in facial recognition tasks. This research investigates how CNNs develop functional differentiation when simultaneously trained on tasks of facial identity and expression. Our results indicate that a specialized model exclusively trained on a single task underperforms on the other task, while a joint model trained simultaneously on both tasks performs at least as well as the specialized model for each task. For interpretation, we used class activation maps. These helped illustrate how the joint model distinguishes different facial attributes for recognizing expression and identity, revealing a functional segregation within the network. This differentiation becomes particularly apparent in the final stages of the convolutional filter processing hierarchy, where task-specific features emerge. In sum, our study presents an interpretable artificial neural network-based framework for facial processing, delivering valuable insights for developing effective neurobiological support systems for individuals with related facial recognition impairments. Code is accessible on Github.

1. Introduction and related work

Faces convey important information about identity, emotional expressions, and social traits, and the human visual system can process this information quickly (Mckone et al., 2009; Susskind et al., 2008; Frith, 2009; Adolphs, 2006; Zadra & Clore, 2011; Plutchik, 2001; Katana et al., 2019; Bar et al., 2006; Willis & Todorov, 2006). Among these skills, recognition of facial identity and facial expression are crucial for social communication as they allow us to identify people, understand their emotional state, and respond accordingly. Facial expression recognition involves decoding a person's emotional state based on the expression of the face, while facial identity recognition identifies individuals by their unique facial features.

Yet, the biological processing mechanisms for these abilities in the brain are still debated. Majority of research advocates the parallel processing of facial expression and identity, each supported by unique neural and cognitive mechanisms (Bruce & Young, 1986; Sergent et al., 1994; Haxby JV, 2000; Winston et al., 2004). Facial expression recognition is supported by brain regions such as the amygdala and the insula, while facial identity recognition is supported by brain regions such as the fusiform gyrus and the superior temporal sulcus. This theory is supported by studies on patients with facial impairments, with impairment in one task not affecting the other (Tranel, 1998; Andrew W. Young & Hay, 1993; Hornak, 1996). However, recent research has suggested some overlap between these two biological processes, with some identity and expression information found in shared brain regions (Dobs, 2018), and in separable face patches of the same region (Yang & Freiwald, 2021).

CNNs are popular as a model of the visual system due to their similarities with how the brain processes visual information (Yamins & DiCarlo, 2016). While CNNs have been widely used for facial expression and facial identity recognition, these tasks have traditionally been studied separately (Mellouk & Handouzi, 2020; Tazi et al., 2022).

The conventional approach is to train each task separately, however the brain processes visual information from multiple tasks at once. Recently, researchers have started to investigate to which extent neural networks exhibit a degree of specialization. Schwartz et al. 2023 found that two separate networks, one trained on identity and one on expression resulted in more orthogonal features in deeper layers, suggesting subspace disentanglement. Dobs et al. 2022 demonstrated that a convnet trained jointly on identity recognition and object detection segregated into two com-

[1]Tri-Insititutional Computational Biology, and Medicine PhD Program: Cornell University, MSKCC, Rockefeller [2]Laboratory of Neural Systems, The Rockefeller University [3]Center for Brain, Minds, and Machines, MIT. Correspondence to: Yanis Tazi <yat4003@med.cornell.edu>.

The 2023 ICML Workshop on Computational Biology, https://icml.cc/.

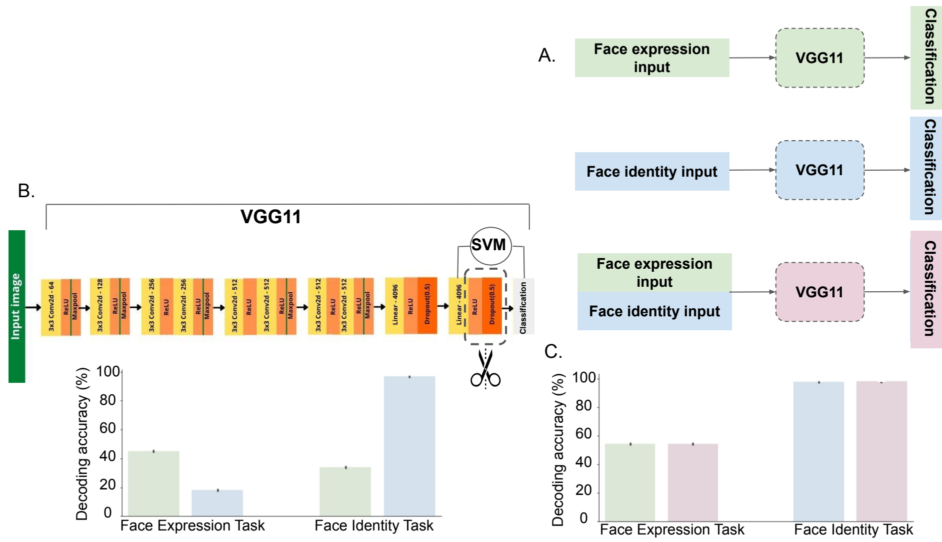

Figure 1. Evidence of functional specialization of facial expression and facial identity in a joint model. (**A**) We optimized three neural networks with VGG11 architecture: one specialized for facial expression, one specialized for facial identity, and one for both tasks jointly. (**B**) Decoding accuracy on the test set using activations extracted from the second to last layer of each model. Specialized face expression network outperforms face identity network for facial expression decoding, and specialized face identity network outperforms specialized face expression network in facial identity decoding. We generated 100 bootstraps on the test set to obtain error bars. (**C**) Joint network simultaneously trained on both tasks performs at least as well as the specialized network for both tasks. We generated 100 bootstraps on the test set to obtain error bars.

putational systems, learning task-specific features in deeper layers. However, training a network simultaneously on both facial identity and expression has not been attempted yet. This is a more challenging task, as the network receives the same type of input images (faces), and functional specialization is not necessarily expected, and harder to interpret.
Our study introduces a novel approach of simultaneously training a model on facial expression and identity, prioritizing interpretability of learned features to solve each task. Our contributions are: 1) Jointly trained ConvNets on facial expression and identity perform at least as well as specialized models trained on a single task, while specialized ones only performed well on their respective tasks, suggesting task-specific functional specialization of the joint model, 2) The joint model captures more efficient facial attributes specific to each task than specialized model for one task, 3) We offer an interpretable visual explanation of how the network processes and differentiates between facial identity and expression tasks, and 4) We propose using identified facial biomarkers to enhance facial processing skills in patient with facial recognition impairments.

2. Data and Method

2.1. Data

In our study, we aim to train neural networks on tasks related to facial expression and identity recognition. For this purpose, we have employed three diverse datasets comprising images of individuals displaying various facial expressions under different conditions:
The KDEF dataset (Lundqvist et al., 1998), with 4900 pictures of facial expressions from 70 individuals (35 women), showcases seven emotional expressions from five angles. Since its creation in 1987, this dataset has been widely used in neuroscience, psychology, and computer vision research.
The VoxCeleb dataset (Nagrani et al., 2017) a compilation of short video clips from interviews with 1251 individuals, represents a broad demographic range. We'll use this dataset to validate the findings in the identity task.
Lastly, the FER-2013 dataset, which includes 30,000 images of seven different facial expressions from various individuals, will be employed for validation in the expression task.

2.2. Model training and evaluation

To evaluate the performance of our joint model for facial identity and expression, we optimized three neural networks with VGG11 architecture (Figure 1A) with random weights initialization. VGG11 is a relatively straightforward CNN architecture comprising eight convolutional layers and three fully connected layers, making it well suited for our interpretability goals. The convolutional layers have a 3*3 kernel size and are followed by a rectified linear unit (ReLU) activation function and a max pooling layer with a 2*2 kernel size. We chose this well-known architecture for its performance in computer vision tasks, and its relatively shallow depth which aligns with the fast processing of the human visual system.
The KDEF dataset was used for training, with each set of 35 distinct identities split into two sets of 2450 frames each, with 80% of the frames used for training and 20% for testing, ensuring identity-wise splitting for expression. One set was

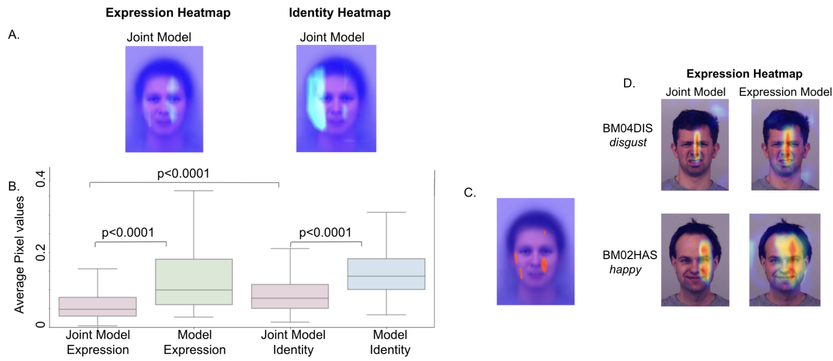

Figure 2. Class Activation Maps Average on the top predicted expression and identity for the three models. Each model generates an activation map for each image for each possible class. For each model, we then average the activation maps of the top predicted class for expression and identity. (**A**) Average Heatmap of the top predicted expression or identity class for each model. (**B**) Boxplot comparison of pixel values for expression and identity predictions among the three models. We conducted pairwise Mann-Whitney U tests to compare the pixel values. (**C**) Heatmap of intersection of joint model for expression and identity. We only activated pixel values above the 95th percentile. (**D**) Class activation visualization of the expression heatmap on two example input frames from the KDEF dataset.

used for training the identity network, the other for training the expression network, and both sets were used for training the joint model (multi-label network). All the models were trained for 250 epochs with identical parameters, using SGD with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.0001. The cross-entropy loss was used to update the weights of the model. To ensure generalization, we repeated the same procedure using the VoxCeleb dataset to train the identity network, the FER-2013 dataset to train the expression network, and both datasets to train the joint network.

To assess the transferability of representations learned from one task to the other, we extracted vector representations from the second-to-last layer of each train and test frame and trained a support vector machine on the extracted representations (Figure 1B). Since the classes are balanced, the performances of the models were measured by the classification accuracy of the top predicted class for each frame in the test set, with error bars obtained through bootstrapping on the predictions.

2.3. Interpretability methods

CNNs extract meaningful features from images through its convolutional filters. These filters, when visualized and interpreted, provide insights into how the CNN processes images, allowing us to identify the specific features the model leverages to predict a particular class.

Class Activation Maps (CAMs) (Selvaraju et al., 2017) is a technique that highlights the regions of an image that contribute most to the final class prediction of a model. This is done by calculating the weighted sum of the output feature map of the last convolutional layer, where the weights are determined by the gradient of the predicted class with respect to each channel. This produces an activation map that depicts the relative importance of each spatial location in the image for the target class. CAMs are highly interpretable, and allow identification of specific features and patterns the model utilizes for a particular class. This is particularly useful for understanding how the joint model's decision-making processes differ when predicting either identities or expressions. To identify general patterns for expressions and identities, we averaged the activation maps from the joint model of the top predicted class for expression and the top predicted class for identity. We displayed these activation maps on the average face background image using the straight images in the KDEF dataset, to ensure that these activation maps align visually with the specific facial attributes identified for each class.

Another way to understand the decision-making process of the model is to examine the features extracted by the model's filters (Yosinski et al., 2015). We visualized the preferred stimulus for each filter by initially presenting the network with a random noise input image, then modifying this input to maximize the activation of the filters. This was achieved using gradient ascent to adjust the values of the initial random noise input image and creating a loss function that maximizes the value of the filter. The iterative adjustment of the input image values using stochastic gradient ascent led to the maximization of the filter's activation. The resulting image is a visual representation of the filter's target features.

3. Results

3.1. Evidence of functional specialization in the joint model

We have trained 3 VGG11 networks: one for facial identity, one for facial expression, and one jointly on both tasks using the KDEF dataset as previously detailed. As anticipated, the specialized facial expression network effectively decoded untrained facial expressions, and the specialized facial identity network accurately decoded untrained facial identities. However, the specialized facial expression network demonstrated subpar performance in decoding facial identities ($p < 0.0001$, two-sided paired t test, Figure 1B), and the specialized facial identity network similarly underperformed for decoding facial expressions ($p < 0.0001$, two-sided paired t test, Figure 1B).

Interestingly, the network trained jointly on both tasks performed at least as well on each task compared to the specialized network trained solely on that task (Figure 1C). For further validation, we replicated these findings using the VoxCeleb and FER2013 datasets (S.Figure 1).

Despite using facial images as input, the representations learnt by the models trained on a single task did not seem to benefit greatly for the other task. In contrast, when trained simultaneously on both tasks, the model spontaneously segregated for both tasks, displaying a degree of functional specialization specific to each task.

3.2. Interpretability and visualization of what the models learn

To explore this functional specialization, we aimed to interpret what the models focus on for a given task. This is particularly interesting for the model trained jointly on expression and identity as, in the event of segregation, we would expect to find systematic differences in relevant facial attributes for each task, and potentially in the filters of the deeper layers utilized by the model to solve each task.

Firstly, the joint model learned a more effective representation characterized by smaller receptive fields compared to the specialized models ($p < 0.0001$, Figure 2).

Importantly, the joint model generated distinct heatmaps on opposite sides of the face, with the identity heatmap significantly larger than the expression heatmap ($p < 0.0001$, Figure 2A-B). For expression recognition, the joint model focused on a narrow vertical facial region extending from the upper eye down the mouth, a feature noted in human studies (Schyns et al., 2007). For identity recognition, the model focused on larger facial regions including the face shape and the eye. The joint model also learned a subset of features common to both tasks (Figure 2C).

We further analyzed individual facial frames for specific expressions and identities. The joint model accurately captured action units (Ekman & Friesen, 1978) for expression prediction. When predicting disgust for BM04DIS, the model created a narrow line extending from the nose to the bottom lips capturing the three action units characteristic of disgust (Nose Wrinkler, Lip Corner Depressor, Lower Lip Depressor) (Figure 2D). When predicting happiness for BM02HAS, the model created a narrow vertical line and considered the eyes, jaw and corner of the mouth, effectively capturing the happiness action units (Cheek Raiser and Lip Corner Puller).

In light of the previous analyses, we expected the network to segregate for both tasks by learning representations specific to each task. To understand the patterns that the convolutional layers seek in an image, we visualized the preferred stimulus of each filter in the model. The filters in the early layers showed similar features regardless of the task, encoding simple directional edges and colors (first convolutional layer), while the filters in higher layers encoded more complex combinations of edges and colors to represent facial features such as eyes (fourth and sixth convolutional layers). Units in the last convolutional layer encoded face appearances (Figure 3, S.Figure 2A). In the final layer, we observed an increasing number of blank filters, indicating that the features encoded by the filters were increasingly task-related rather than image-related. These results demonstrate that the development of distinctive features each task relies on becomes apparent in the later layers (S.Figure 2).

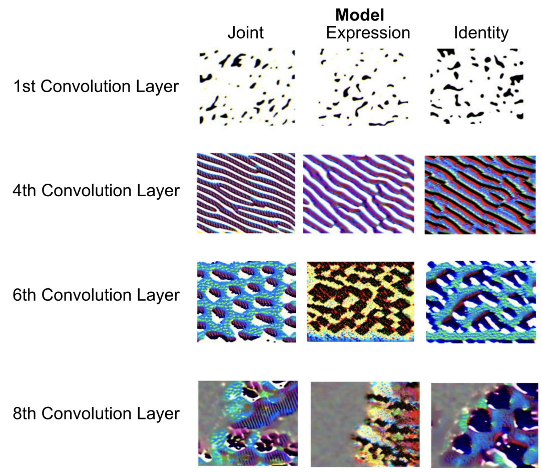

Figure 3. Exploring the representations learned by convolutional neural networks. Images optimized to drive responses in one example filter. We show images optimized to drive responses in one example filter for each of the three models in the first, fourth, sixth, and last (eighth) convolutional layers of VGG11.

4. Discussion and future work

Our research has shed light on the capacity of a single convolutional neural network, trained simultaneously on both facial expression and identity recognition, to exhibit functional specialization, thereby segregating distinct features specific to each task.

The interpretability methods provided meaningful insight into the decision-making process of the model. CAM and preferred stimulus visualization showed that the joint model focuses on different facial attributes for each task, and that task specific features only emerge in the deeper layers. These methods could be widely applicable to other studies aiming to understand the internal mechanisms of biological visual systems. Our study also identified task-specific facial biomarkers that could assist in training individuals with facial recognition impairments. Notably, by examining the images that maximally activate each unit in the model, we can potentially identify and tune disrupted neurons in the visual system to respond better to task-optimized images.

Looking ahead, there is a need to explore more biologically

plausible architectures that process facial information dynamically, enhancing our understanding of the mechanisms underlying facial recognition. Furthermore, additional research should focus on comparing these artificial neural networks with their biological counterparts in the brain, opening new avenues for neurobiological applications.
In conclusion, our work underscores the potential of CNNs as powerful tools in computational biology, generating interpretable models of complex biological processes like facial recognition.

5. Acknowledgements

We thank C.D. Gilbert, J.D. Victor, and A.J. Hudspeth for feedback, L.Yin for administrative assistance and the Rockefeller HPC cluster for the GPU support. This work was supported by Alexander von Humboldt-foundation (M.B); the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 (W.A.F); and by the Department of the Navy, Office of Naval Research under ONR award number N00014-20-1-2292 (W.A.F.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and the Office of Naval Research.

References

Adolphs, R. Perception and emotion: How we recognize facial expressions. *Curr. Dir. Psychol. Sci.*, 15(5):222–226, October 2006.

Andrew W. Young, Freda NewcombeEdward H. F. de Haan, M. S. and Hay, D. C. Face perception after brain injury. *Brain*, 1993.

Bar, M., Neta, M., and Linz, H. Very first impressions. *Emotion*, 6(2):269–278, May 2006.

Bruce, V. and Young, A. Understanding face recognition, 1986.

Dobs, K.; Schultz, J. B. I. G. J. Task-dependent enhancement of facial expression and identity representations in human cortex. *NeuroImage*, 2018.

Dobs, K., Martinez, J., Kell, A. J. E., and Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci Adv*, 8(11):eabl8913, March 2022.

Ekman, P. and Friesen, W. V. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

Frith, C. Role of facial expressions in social interactions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 364(1535):3453–3458, December 2009.

Haxby JV, Hoffman EA, G. M. The distributed human neural system for face perception. *Trends Cogn Sci*, 2000.

Hornak, J., R. E. T. . W. D. Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, 1996.

Katana, M., Röcke, C., Spain, S. M., and Allemand, M. Emotion regulation, subjective Well-Being, and perceived stress in daily life of geriatric nurses. *Front. Psychol.*, 10: 1097, May 2019.

Lundqvist, D., Flykt, A., and Öhman, A. Karolinska directed emotional faces, May 1998. Title of the publication associated with this dataset: PsycTESTS Dataset.

Mckone, E., Crookes, K., and Kanwisher, N. The cognitive and neural development of face recognition in humans. *The cognitive neurosciences., 4th ed.*, 4(2009):467–482, 2009.

Mellouk, W. and Handouzi, W. Facial emotion recognition using deep learning: review and insights. *Procedia Comput. Sci.*, 175:689–694, January 2020.

Nagrani, A., Chung, J. S., and Zisserman, A. VoxCeleb: a large-scale speaker identification dataset. June 2017.

Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.*, 89(4):344–350, 2001.

Schwartz, E., O'Nell, K., Saxe, R., and Anzellotti, S. Challenging the classical view: Recognition of identity and expression as integrated processes. *Brain Sciences*, 13, 2023.

Schyns, P. G., Petro, L. S., and Smith, M. L. Dynamics of visual information integration in the brain for categorizing facial expressions. *Current Biology*, 17(18):1580–1585, 2007. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2007.08.048. URL https://www.sciencedirect.com/science/article/pii/S0960982207019045.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Sergent, J., Ohta, S., Macdonald, B., and Zuck, E. Segregated processing of facial identity and emotion in the human brain: A pet study. *Visual Cognition*, 1(2-3):349–369, 1994. doi: 10.1080/13506289408402305.

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., and Anderson, A. K. Expressing fear enhances sensory acquisition. *Nat. Neurosci.*, 11(7):843–850, July 2008.

Tazi, Y., Berger, M., and Freiwald, W. A. Towards an objective characterization of an individual's facial movements using self-supervised person-specific-models, 2022.

Tranel, D., D. A. R. . D. H. Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, 1998.

Willis, J. and Todorov, A. First impressions, 2006.

Winston, J. S., Henson, R., Fine-Goulden, M. R., and Dolan, R. J. fmri-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, 92(3):1830–1839, 2004.

Yamins, D. L. K. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, March 2016.

Yang, Z. and Freiwald, W. A. Joint encoding of facial identity, orientation, gaze, and expression in the middle dorsal face area, 2021.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. Understanding neural networks through deep visualization, 2015.

Zadra, J. R. and Clore, G. L. Emotion and perception: the role of affective information. *Wiley Interdiscip. Rev. Cogn. Sci.*, 2(6):676–685, November 2011.

6. Supplementary Material

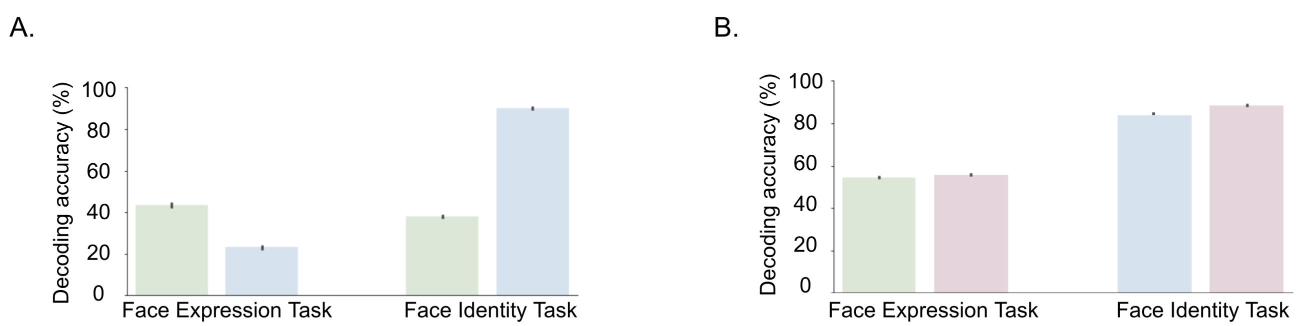

Figure S.1. Same as Figure 1B. and C. with two other datasets.

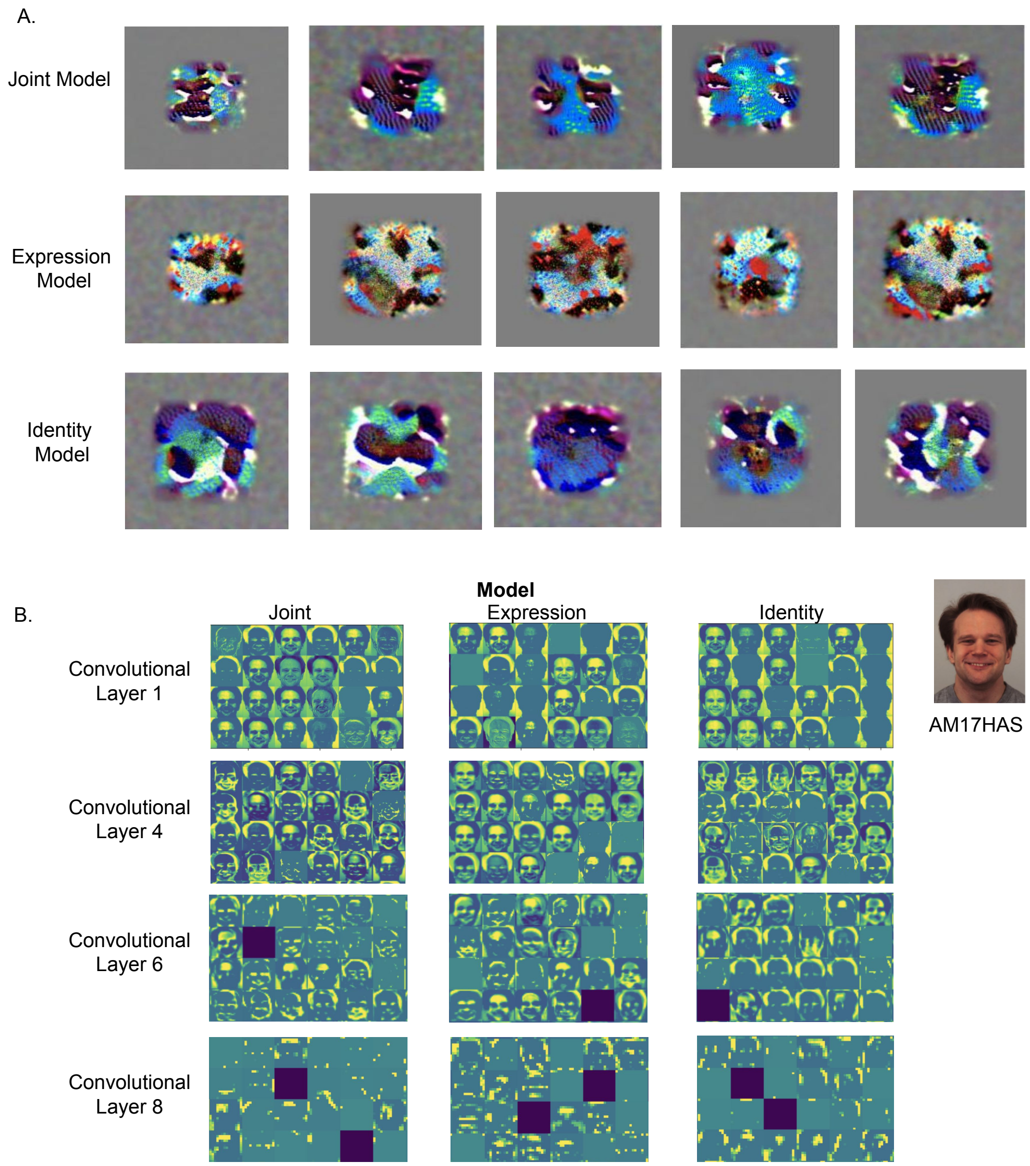

Figure S.2. Exploring representations learned by convolutional neural networks. (**A**) Images optimized to drive responses in five example units of the last convolutional layer for the three models. We show images optimized to drive responses in the top five units of the last convolutional layer for each of the three models. We focus on the last convolutional layer due to the small receptive fields in earlier layers. (**B**)Visualization of intermediate convnet outputs for the three models. We visualize intermediate Convnet outputs for each of the three models given an input image of AM17HAS from the KDEF dataset. This provides insight into the learned features and processing steps that occur in the Convnet.