Reevaluating the Ventral and Lateral Temporal Neural Pathways in Face Processing: Deep Learning Insights into Face Identity and Facial Expression Mechanisms

Emily Schwartz

A dissertation

submitted to the Faculty of

the department of Psychology and Neuroscience

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

June 2024

"*I know faces, because I look through the fabric my own eye weaves, and behold the reality beneath.*"

Kahlil Gibran

Dedicated to my brother Brian who without I would have never passed my first neural data analysis class.

Reevaluating the Ventral and Lateral Temporal Neural Pathways in Face Processing: Deep Learning Insights into Face Identity and Facial Expression Mechanisms

Emily Schwartz

Advisor: Dr. Stefano Anzellotti, Ph.D.

There has been much debate over how the functional organization of vision develops. Contemporary theories that are inspired by analyzing neural data with machine learning models have led to new insights in understanding brain organization. Given the evolutionary importance of face perception and the specialized mechanisms that have evolved to support evaluating it, examining faces offers a unique way to study a dedicated mechanism that shares much of its organization in ventral and lateral neural pathways with other social stimuli, and provide insight into a more general principle of the organization of social perception. According to a classical view of face perception (Bruce and Young, 1986; Haxby, Hoffman, and Gobbini, 2000), face identity and facial expression recognition are performed by separate neural substrates (ventral and lateral temporal face-selective regions, respectively). However, recent studies challenge this view, showing that expression valence can also be decoded from ventral regions (Skerry and Saxe, 2014; Li, Richardson, and Ghuman, 2019) and identity from lateral regions (Anzellotti and Caramazza, 2017). These recent findings have inspired the formulation of an alternative hypothesis. From a computational perspective, it may be possible to process face identity and facial expression jointly by disentangling information for the two properties. This hypothesis was tested using deep convolutional neural network (DCNN) models as a proof of principle. Subsequently, this is then followed by evaluating the representational content of static face stimuli within ventral and lateral temporal face-selective regions using intracranial electroencephalography (iEEG). This is then extended

to investigating the representation content of dynamic faces within these regions using functional magnetic resonance imaging (fMRI). The results reported here as well as the reviewed literature may help to support the reevaluation of the roles the ventral and lateral temporal neural pathways play in processing socially-relevant stimuli.

Contents

Acknowledgements

I would first and foremost like to thank my advisor, Dr. Stefano Anzellotti. None of this would have possible without your incredible mentorship. When I first met you, I immediately could see how passionate you were not only for your research, but also for shaping and nurturing young minds in this field. I have been exceptionally fortunate that my initial impression has held true over my past five years as your graduate student. Your commitment to your students, both graduate and undergraduate, truly shows your dedication to inspiring the next generation of scientists. I'm especially grateful for all the times you helped me track down the silliest errors in my code. I am very sad that I will not still be in the lab when I go on my two trips to Italy this September since I know you would have definitely encouraged me to stay the 3 weeks.

To my committee members, Drs. Liane Young, Maureen Ritchey, and Avniel Ghuman, I am so appreciative of the time you have taken to sit on my committee. It's also been a privilege to get to collaborate with Dr. Ghuman.

I also want to thank everyone I've worked with in the lab. A special shoutout to Aidas Aglinskas, my favorite postdoc, who has patiently answered my endless questions from day one. You've been so supportive, not only with helping me improve my technical skills, but also with life advice on figuring things out. To my current labmates, Hamed Karim and Yiyuan Zhang, your support over these past few months has been invaluable. I really appreciate your patience as I overthink and talk through everything in way too much detail. To the lab alumni, Tony Chen, Craig Poskanzer, and Mengting Fang, I'm grateful for the 2-3 years we spent as labmates. Even though most of that time was during Covid, you were always there

to help. A big thank you to Yiyuan Zhang, Jordan Wylie, and Michael Manalili for volunteering to run fMRI scans with me.

To my parents, I'm incredibly grateful every single day for your support and encouragement. Whether it was driving me all over the country for crew races in high school or backing my decision to return to school for another five years, you've taught me the value of hard work and kindness. To my mom, I am so fortunate that not only do I get to have you as my mom, but also as my best friend and role model. To my dad, I don't know anyone who works more than you do. You've shown me how to persevere and how to challenge myself. I will never be able to thank you and Mom enough for everything you have done for me. To my fiancé Rohan, thank you for supplying me when an endless amount of food and support through all of this. To my brother Matthew who had to live with his little sister for two years during covid and calmed me down after I noticed a very ridiculous typo in my NSF application. To my brother Brian, you and Matthew were the reason why I wanted to be good at math. Then, you were the reason why I thought coding would be cool. Now, you are the reason why I want to study the brain.

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
ATL	Anterior Temporal Lobe
aSTS	Anterior Superior Temporal Sulcus
BIC	Bayesian Information Criterion
CelebA	Large-Scale CelebFaces Attributes Database
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DISFA	Denver Intensity of Spontaneous Facial Action Database
DFEW	Dynamic Facial Expression in-the-Wild
EBA	Extrastriate Body Area
ERP	Event-Related Potential
FC	Fully-Connected
FER2013	Facial Expression Recognition 2013 Dataset
FFA	Fusiform Face Area
fMRI	Functional Magnetic Resonance Imaging
fs–pSTS	Face-selective Posterior Superior Temporal Sulcus
HAA500	Human-Centric Atomic Action Dataset with Curated Videos
iEEG	Intracranial Electroencephalography
IRIEH	Integrated Representation of Identity and Expression Hypothesis

IT	Inferior Temporal
KDEF	Karolinska Directed Emotional Faces Database
LGN	Lateral Geniculate Body
LOTC	Lateral Occipitotemporal Cortex
MD	Middle Dorsal
MT	Middle Temporal
MVPA	Multi-Voxel Pattern Analysis
OFA	Occipital Face Area
PCA	Principal Component Analysis
PFC	Prefrontal Cortex
PPA	Parahippocampal Place Area
pSTS	Posterior Superior Temporal Sulcus
RDM	Representational Dissimilarity Matrix
ROI	Region-of-Interest
RSA	Representational Similarity Analysis
TDANN	Topographic Deep Artificial Neural Network
TMS	Transcranial Magnetic Stimulation
stP	Single-trial potentials
STS	Superior Temporal Sulcus
VOTC	Ventral Occipitotemporal Cortex

Chapter 1

An Introduction to the Functional Organization of Higher-Order Vision and Faces in the Brain

1.1 Functional organization of vision in the brain

Humans, more so than most mammals, rely on vision over their other senses. One constantly needs to be able to understand their visual surroundings to know how to navigate through their environments both literally (i.e. spatial navigation) and figuratively (i.e. social navigation). Vision is a computationally demanding task, yet most of us recognize our surroundings effortlessly, hinting at the complex mechanisms and organization involved. This has driven many to study our visual system in order to better understand how we perceive the outside world.

The human visual system is not uniform and undifferentiated. Early visual regions (V1-V3) branch into three distinct streams: a ventral pathway, running along inferior temporal cortex; a lateral pathway, running along the superior temporal sulcus; and a dorsal pathway, cutting into the parietal lobe (Ungerleider and Mishkin 1983, Felleman and Van Essen 1991, Goodale and Milner 1992, Pitcher et al. 2021). In turn, these streams can be subdivided into distinct regions encoding topographic maps of the visual field (Silver and Kastner, 2009), and starting in the more anterior portions of the occipital cortex demonstrate an organization by object category (Sergent and Signoret, 1992; Allison et al., 1994; Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Chao, Haxby, and Martin, 1999) and real-world object size (Konkle and Oliva, 2012; Julian, Ryan, and Epstein, 2017). The organization into distinct streams is often discussed separately from category-selectivity and object size effects. However, the principles shaping these different aspects of the large-scale organization of visual cortex might be similar. Therefore, considering them jointly can help to paint a more comprehensive picture of the visual

system in which evidence concerning one aspect of organization could inspire insights into the others.

1.1.1 Cortical pathways in vision

Early anatomical and physiology studies first brought forth the idea of multiple pathways in the visual system (Minkowski, 1920; Livingstone and Hubel, 1988; Felleman and Van Essen, 1991). Researchers studying the primate visual system investigated the different cell types in the lateral geniculate body (LGN), a subcortical structure that receives input from the retina via the optic nerves. They found that the retina mapped onto the cells of the LGN differently depending on which layer of the LGN those cells were located within (Minkowski, 1920; Leventhal, Rodieck, and Dreher, 1981). Most notable was the distinction between mangocellular and parvocellular layers. Cells within these subdivisions had different anatomical properties, indicating they may be better suited for specific parts of visual processing (e.g., color versus acuity). A continuation of these subdivisions were found for V1 and V2, and seemed to become more pronounced further downstream (Livingstone and Hubel, 1988). The magnocellular and parvocellular cells in the LGN send inputs to V1 in primates. The cortex in V1 is made up of six different cellular layers, and these different cellular layers send inputs to different downstream visual regions. For instance, layers 2 and 4 in V1 which are involved in processing form and color have been mapped inputs into V2, but have not been shown to send inputs to area MT/V5. V4 additionally receives color-selective cell inputs from the parvo subdivisions in V1 (Livingstone and Hubel, 1988), and orientation-selective information related to shape in V1 as well (Mountcastle et al., 1987). Meanwhile, layer 4B in V1, which contains cells selective to

binocular disparity and directionality, as well as projections from magnocellular cells, is involved in processing motion and directs its outputs to both V2 and area MT (Figure 4 in Livingstone and Hubel, 1988). In line with this, areas V4 and V5, a region that directs its outputs to area MT, connect to separate subregions of V2 in macaques (Shipp and Zeki, 1985). Additionally, a third type of cellular layer has been found in the LGN called the koniocellular layer which plays a more specific role for representing colors along the yellow and blue spectrum (Hendry and Reid, 2000).

Given the high contrast sensitivity and low resolution of magnocellular cells, which are the primary projections to area MT, and the high-resolution capabilities of parvocellular cells for processing fine-grained details that project to area V4, it makes sense to further evaluate the distinct pathways that connect the areas more thoroughly. Felleman and Van Essen (1991) conducted a detailed review of the macaque visual system to establish the demarcations of visual regions and their connections (Figure 1.1, adapted from Van Essen et al. (2001)). This publication outlined concrete evidence for a hierarchical model of visual processing, evaluating connections between brain areas and demonstrating evidence for parallel, interconnected pathways. Their work extended the findings of Mishkin, Ungerleider, and Macko (1983), particularly concerning the branching patterns observed in the medial superior temporal region (MST).

Specifically, they identified distinct subdivisions within MST: MSTd, located dorsally, MSTl, positioned ventro-antero-laterally to MSTd, and FST, further ventro-antero-laterally. These subdivisions exhibited varying response properties, connectivity patterns, and different functional effects of eye gaze movement under

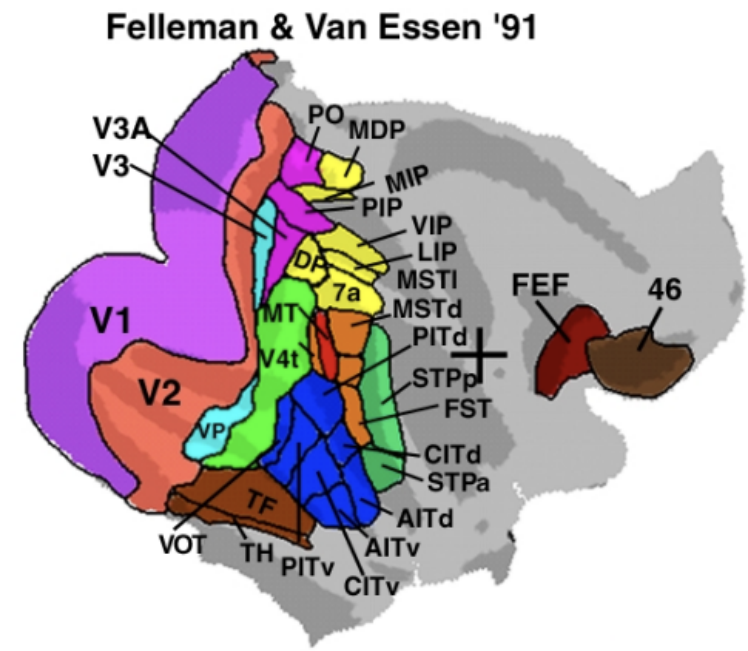

FIGURE 1.1: Partitioning schemes for macaque visual cortex adapted from Van Essen et al. (2001)

selective stimulation, which was particularly observed in MSTd and MSTl (Komatsu and Wurtz, 1989; Felleman and Van Essen, 1991). Macko et al. (1982) used 2-[14C]deoxyglucose to map the full extent of these visual pathways. The neurobehavioral studies closely aligned with the neuroanatomical mappings. The hypothesis that these distinct visual pathways may be extensions of the parvocellular and magnocellular systems (Livingstone and Hubel, 1988) align with the above findings related to the subdivisions of the higher-order visual regions, and help to affirm the distinct functional roles of the ventral, lateral, and dorsal pathways in visual processing that will be discussed later on.

In conjunction with this, further structures are observed within the streams. For example, different brain regions within the ventral stream respond selectively to distinct object categories (Kanwisher, McDermott, and Chun, 1997; Epstein and

Kanwisher, 1998; Downing et al., 2001). This is within a broader gradient of animacy in which inanimate objects drive responses in medial portions of ventral temporal cortex and animate objects drive responses in lateral portions, with similar patterns of response for both perceptual and semantic attributes of the object (Chao, Haxby, and Martin, 1999). Bracci and Beeck (2016) has also demonstrated that perceptual and semantic information coexist, yet have independent contributions, within ventral regions representing object animacy. Additionally, the ventral temporal cortex may be better explained by the object appearance rather than the object category. In a study where participants were shown lookalike objects and inanimate objects that look like animate objects (e.g., a mug that is made to look like a dog), lookalike objects were more similar to animate objects in ventral temporal regions (Bracci et al., 2019). Lesion studies have also shown that recognition of categories of living things can be damaged independently of the ability to recognize categories of nonliving things (Warrington and McCarthy, 1987; Caramazza and Shelton, 1998), and that these deficits can extend beyond visual recognition towards semantic attributes as well. An additional known constraint is superimposed on this organization – distinct regions show preferential responses to small versus large objects (Konkle and Oliva, 2012; Konkle and Caramazza, 2013; Julian, Ryan, and Epstein, 2017; Long, Yu, and Konkle, 2018). Smaller objects tend to activate neural regions that overlap with those activated by scenes, which makes sense from an organizational perspective considering that smaller objects often contribute to the representation of a scene. The large-scale organization of the visual system has been a very active topic of research in the past decades. Yet, the mechanisms that give rise to this organization and its functional significance remain debated in the field.

1.2 Theories of category-selectivity

Typically, in the literature on the large-scale organization of visual cortex, distinct theories account for the presence of specific processing streams and for the existence of category-selective and real-world-size-dependent effects. However, taking these aspects of the large scale organization of vision into account jointly might be key to reach a comprehensive understanding of the structure of the human visual system and of its functional relevance. In the following sections, I will discuss first the current theories of category-selective and size-dependent effects, starting from the rich neuropsychology literature on category specific deficits, and continuing with an overview of how functional imaging has shaped the understanding of these phenomena. Next, I will discuss the accounts of the distinction between different visual streams and of their functional roles.

1.2.1 Neuropsychology-based theories

Early work motivated by category-specific deficits in patients generated a rich space of hypotheses to account for impairments that selectively affected some types of objects but not all (Warrington and McCarthy, 1983; Caramazza and Shelton, 1998). These theories varied along several dimensions, with a key one being whether object representations are organized into separate brain regions specialized for different categories (the "neural structure principle"), or whether category structures emerge within a unitary region from the correlations between features within different types of objects (the "correlated structure principle", see Mahon and Caramazza, 2009). One influential proposal – the sensory-functional theory (Warrington and Shallice, 1984) – hypothesized that object representations were

organized into a system for the representation of sensory knowledge, and a system for the representation of functional knowledge. Impairments for sensory knowledge would result in deficits for the recognition of animate entities, and impairments for functional knowledge in deficits for the recognition of inanimate entities. For instance, recognition of living things and food would likely be spared together because they both rely on sensory inputs, while nonliving things and body parts could be found spared together due to the functional role of objects and the use of body parts to perform the function in conjunction with an object. An alternative proposal – the domain-specific hypothesis (Caramazza and Shelton, 1998) – argued that category-based organization results from specialization for the processing of evolutionarily-relevant object domains. This hypothesis predicts that a category-specific impairment would be associated with an impairment for both the perceptual and semantic information about the category. A third proposal – the organized unitary content hypothesis (OUCH, Caramazza et al. 1990), argued that there is a privileged relationship between some types of input representations (such as visual form) and certain types of output representations (such as knowledge of object manipulation). This would indicate that if categories consist of properties that are highly correlated, then they would be impaired or spared together. This is but a sampling of a broader theoretical landscape (see Mahon and Caramazza 2009 for an in-depth discussion).

These initial theories of category-selectivity were driven primarily by case studies of patients with selective recognition impairments. At both the perceptual and semantic level, individual studies in patients have provided crucial insight for the mechanisms involved in object and face recognition. For instance, a case study of one patient (Mr. W) demonstrates a selective impairment of object recognition

with seemingly intact face recognition abilities (Rumiati et al., 1994). D.F., a well known patient with visual form agnosia, has profound object recognition deficits (Goodale et al., 1994). D.F. does show some impairment for faces, however this seems to be only a partial effect due to D.F.'s ability to perform well on many face categorization tasks (Steeves et al., 2006).

1.2.2 Imaging-based theories

Prior to the invention of functional Magnetic Resonance Imaging (fMRI), work on the functional organization of the visual system primary came from behavioral studies and their neural implications, along with anatomical work and neuropsychological testing in humans, and work in animals. There was some electroencephalogram (EEG) research; however, EEG lacks the spatial resolution of fMRI. FMRI enabled researchers to identify which parts of the brain were activated with much higher resolution than previously possible, facilitating the mapping of a functional atlas of the brain. After the emergence of fMRI in the 1990s, many researchers delved into investigating how the brain perceives the outside world. This led to many conflicting theories on functional brain organization that are still being discussed today.

1.2.2.1 Functional module hypothesis

The domain-specificity hypothesis is driven by evolutionary pressures that lead to the existence of brain regions with specialized functions. The fact that brain regions are consistently involved in processing specific categories supports the

idea of evolutionarily-determined modules for brain functionality. These modules are thought of as specific machinery in the brain that are responsible for processing distinct inputs. This theory has been particularly influential in the study of higher-order vision and integrates well with the functional module hypothesis, suggesting that the brain contains dedicated areas for processing specific visual domains. Using controlled functional localizer tasks, Kanwisher, McDermott, and Chun (1997) employed fMRI to define specific regions in the brain that are more activated for faces compared to objects, houses, hands, and scrambled faces to control for low-level properties. They build upon work from Sergent and Signoret (1992) and McCarthy et al. (1997), and discovered a region that was consistent across subjects. This area became known as the fusiform face area (FFA), with the area responding selectively to faces. The pioneering work helped to support a foundational basis for the concept of mapping of functional modules in the brain. Similarly, Epstein and Kanwisher (1998) implemented the localizer method to constrain the parahippocampal place area (PPA), a brain region that responds selectively to visual scenes. In addition to this, Downing et al. (2001) found neural responses selective to the human body in the lateral occipital temporal cortex (LOTC), becoming known as the extrastriate body area (EBA).

Both face-selective, body-selective, and scene-selective neuronal patches have also been identified in monkeys. Tsao et al. (2006) uncovered face patches within the temporal lobe in macaques where 97% of the visually responsive cells in the region were activated by faces. This area is thought to be the primate equivalent of the FFA found in humans. In a similar manner, Popivanov et al. (2014) found a patch of neurons responding selectively to body parts in rhesus monkeys within the midSTS. Additionally, Kornblith et al. (2013) reported the discovery of a

scene-selective lateral place patch and a medial place patch in the parahippocampal gyrus of macaques. These patches may be homologous to the PPA found in humans. The detection of the selective patches in monkeys allowed researchers to develop a deeper understanding of the properties of single neurons through electrophysiology studies that cannot be studied in human subjects.

Event related-potential (ERP) studies have also been instrumental in demonstrating specialized processing for different categories. For example, ERP studies report an N170 component that has been repeatedly found in response to faces (Bentin et al., 1996; Eimer, 2000). At 170*ms* post face stimulus onset, electrodes located over the occipito-temporal area will show a negative voltage spike. This spike is comparatively larger than what is seen for objects and scenes. Scenes also modulate an ERP component known as P2, a positive component at about 220*ms*, showing amplitude sensitivity to global properties of scenes (Harel et al., 2016).

1.2.2.2 Expertise hypothesis

A different theory was also put forth known to explain the emergence of the face-specificity found in the FFA called the expertise hypothesis (Gauthier and Tarr, 1997; Gauthier et al., 1999). Gauthier and Tarr (1997) emphasized the view that faces are organized in similar configurations, and are recognized at an exemplar-specific level rather than a basic-level, which would predominately be the case for the recognition of most objects. In conjunction with this, the FFA may not actually be dedicated specifically to discriminate between faces, but rather may be dedicated to discriminating stimuli that the specific individual has expertise in. To test this, they created non-face stimuli called greebles. Subjects underwent training to become experts at recognizing greebles. In a followup study, Gauthier et al. (1999)

demonstrated that the FFA was activated to a greater extent for those who were greeble experts than those who were not greeble experts. However, this theory has been largely reconsidered in light of alternative explanations for the findings. Studies of individuals with acquired prosopagnosia who exhibit abnormalities in the FFA and severe deficits in face recognition, demonstrate normal performance in learning to recognize greebles (Rezlescu et al., 2014).

1.2.2.3 Distributed hypothesis

An initially alternative theory was proposed by Haxby et al. (2001) a few years after the discovery of the FFA (Kanwisher, McDermott, and Chun, 1997). At the time, most fMRI work implemented a univariate approach that investigated the mean activation of blood-oxygen-level-dependent (BOLD) response in a brain region of interest. Instead, Haxby et al. (2001) expanded on a multidimensional scaling technique by Edelman et al. (1998) of voxel space representations to analyze the pattern of responses to specific stimuli within the region of interest. They found that within the FFA they could identify distinct patterns of activation not only for faces, but for objects as well. At the same time, the activation in the FFA was still higher in response to faces compared to objects. However, this indicated that the FFA still carried information for objects, leading to the idea of distributed and overlapping neuronal regions for processing visual stimuli.

1.2.3 Contemporary theories

Improvements in neuroimaging scanner capabilities and analysis methods, as well as advances in computational models, in particular deep neural networks have influenced the theories of category-selectivity in neuroscience (e.g., Doshi and Konkle 2023, Dobs et al. 2022), shaping current hypotheses of the organization of object representations. Based on the literature, it could be argued that representations of objects are constrained on two ends: at the level of the inputs (due to the visual properties of objects in different categories), and at the level of the outputs (due to the optimization of performance for behaviorally-relevant tasks). Different accounts of category-specific organization vary in terms of the emphasis they place on constraints at these two levels. Additionally, a robust theory of functional organization should not only account for neural responses, but also explain how neurons are arranged in a manner that optimizes functionality (Margalit et al., 2024).

1.2.3.1 Visual inputs and statistical learning in shaping neural structures

The visual sensory inputs from the world that travel through the retina and into the brain are made up of consistent patterns and regularities. For instance, we typically focus our attention on faces, presenting them prominently in our visual field, and all human faces share a similar curvature. Functional cortical organization is likely partially constrained by the statistical properties of the visual inputs. Multiple topography-related theories have been proposed at the input level. Retinotopy involves the mapping of visual input from the retina onto corresponding neurons within visual brain regions. For example, visual stimuli that appear in the center of your field of view are captured by the fovea of the retina. These central visual inputs are then projected onto a specific, corresponding foveal region within

the visual cortex. This mapping ensures that the spatial relationships in the visual field are preserved in the brain's processing of visual information. Similarly, there are statistical relationships between the distance of one's center of gaze and a point in their visual field. This is referred to as eccentricity in vision. There are also topographic maps reflecting more complex properties of visual stimuli such as curvature. Neurons in V1 respond selectivity to edges and borders at specific orientations (Livingstone and Hubel, 1988), and higher order brain regions integrate this information into curvature patterns to recognize shapes and objects (Arcaro and Livingstone, 2017).

Statistical learning in vision identifies the structure of these visual inputs that frequently and systematically emerge, and certain visual features may be more informative for behavior. This raises the question: is there an a priori inherent mechanism that influences how inputs shape the functional organization? For instance, while evolution might have transformed functional organization to prioritize certain tasks, are there specific input-level mechanisms that have been favored because they are particularly advantageous for these tasks? For example, the ability to recognize faces is crucial for humans. Could the development of an eccentricity-based map partially be a result of our emphasis on performing face-related tasks? Conway (2018) suggests that regions within the inferior temporal (IT) cortex are predisposed to align with parts of the eccentricity template, driven by their relevance to specific goals. This underscores the interplay of both innate and environmental factors in shaping the functional organization of IT.

Several research groups in the field are now exploring the potential of unsupervised or self-organizing algorithms to shape functional cortical organization. These methods utilize types of learning that do not require explicit category-level

pressure to form useful representations of the data. Unsupervised learning models are algorithms that infer patterns and structures from unlabeled data. These models do not have to rely on predefined labels to learn the representations that are inherent in the dataset. Instead, they identify significant relationships, features, and distributions within the data based solely on the inputs, without using target outputs to help structure the learning process. A set of hypotheses have recently emerged within the cognitive neuroscience field that focuses predominantly on constraints at the level of the inputs. For example, the research by Arcaro and Livingstone (2017) demonstrates that the brain is segmented into various regions following retinotopic organization. This structure exists from birth as a protomap, albeit in an immature state, and subsequently matures through one's own experiential influences. Additional work from Arcaro et al. (2017) expands on this, demonstrating that looking behavior towards faces is essential to develop face domains, and that this looking behavior is not innate, but it becomes preferred due to learned reinforcement during development.

Another theory gaining traction proposes that the sensory organization of our brain is subject to spatial constraints, which in turn dictate the topographic organization of the cortex, a process that occurs regardless of the behavioral tasks the brain will eventually support (Doshi and Konkle, 2023; Finzi et al., 2023; Margalit et al., 2024). These groups of researchers were able to demonstrate that by putting certain constraints on the representational organization, done so at the input-level, several topographic features found in the brain emerged in their models without the need of any model supervision. For instance, Doshi and Konkle (2023) found a large-scale organization of animacy as well as object size using a self-organizing

data-driven approach that lacks explicit instruction, challenging the idea of specialized functional modules due to task-constraints. Instead, they propose this as a plausible computational theory where face- and scene- selectivity arise due to visuo-statistical differences. However, they do note some exceptions. For instance, they do not find body-selective units via their self-organizing model. They also find differences in the orthogonality between the animate-inanimate gradient and size features that are not found in the human brain.

Margalit et al. (2024) took a similar approach, constructing a more brain-like model by spatially constraining the organization and minimizing neuronal wiring length to enhance efficiency. Here, they use a different method than Doshi and Konkle (2023) for incorporating local spatial constraints. Instead of using an unsupervised model with self-organizing maps, Margalit et al. (2024) create a topographic deep artificial neural network (TDANN), a contrastive self-supervised network. The topographic portion is due to embedding the neuronal units of each convolutional layer of the model into a two-dimensional simulated cortical sheet where the unit positions are assigned retinotopically. Thus, units that respond to similar regions of an image are then nearby each other in the simulated cortical sheet. This is determined using a spatial loss function that encourages nearby pairs of units to have more correlated responses than pairs that are further away from one another. The model learns multiple signatures of brain functional organization. When compared to neural responses in macaques, TDANN shows corresponding V1 orientation tuning and is similar in its arrangement of orientation-selective neurons. It also predicts similar maps of spatial frequency and color preference in V1. Furthermore, TDANN seems to predict similar category-selectivity maps with face and body units more closely overlapping than face and place units.

The model also indirectly minimizes neuronal wiring length, a constraint that has been favored from an evolutionary standpoint for computational efficiency. Given that the need for a massive set of supervision labels seems unlikely in brain development, this work is compelling.

Additional work has been done using other forms of unsupervised contrastive learning to model the visual ventral stream (Zhuang et al., 2021). These types of model in vision exploit the visual statistics of the inputs by learning their latent representations without utilizing explicit labels or semantic content. Although Zhuang et al. (2021) found that these models do perform similarly to supervised models, they did speculate that unsupervised learning may serve as a proxy when supervised or semi-supervised learning is not possible. This is an intriguing idea since it is likely that the different computational mechanisms that support brain development do not need to be mutually exclusive. Although most of the groups do concede that the types of learning being implemented in the brain are not mutually exclusive, these views highlight the importance of domain-general representations in determining cortical organization.

1.2.3.2 Task-driven modulation in shaping representational structures

Research on the functional organization of the visual system has primarily concentrated on the level of visual inputs, rather than on how these inputs and their representations may be molded at the output level to promote computational efficiency for downstream tasks. However, there is significant evidence suggesting that 1) visual statistics are not enough to explain functional selectivity in the brain and 2) neuronal tuning of cells in the ventral pathway can be dependent on task optimization. There are three ways in which tasks can shape neural responses that

take place over different timescales. First, there are long-term effects from evolutionary selection that may innately specify cortical organization. In a seminal study by Kosakowski et al. (2022), infants as young as two months old underwent an MRI to look at functional responses to faces, scenes, and bodies. Using a carefully designed infant coil and a paradigm that controlled for the presence of protomap features, the researchers found face-, scene-, and body-selective regions in the same anatomical locations as adults. These responses could not be explained by visual features, and additionally they did not find selective responses in the OFA and occipital place area (OPA), challenging a strictly serial relationship of regions for bottom-up processing (Van Grootel et al., 2017). A second task-related mechanism that may shape brain representations is the role of long-term visual experience. Longitudinal studies in macaques demonstrate time-dependent responsiveness towards monkey faces that are present at 1 month of age, but becomes stronger in the first year of life, suggesting experience-related tuning (Livingstone et al., 2017). A third mechanism that affects the neural representations of tasks relates to short-term attention and its top-down role in modulating neural responses. Responses in IT often follow activation in the prefrontal cortex (PFC), which is crucial for planning and guiding behavior (Conway, 2018). This pattern, along with feedback connections from the PFC to the IT, indicates a top-down influence on the IT (Sheinberg and Logothetis, 1997; Sigala and Logothetis, 2002; Conway, 2018; Dobs et al., 2018). This discussion primarily focuses on the first point, attempting to understand the role of tasks shaped by evolutionary pressures.

What are the potential behaviorally-relevant mechanisms that may constrain

category-specificity? Goodale and Milner (1992), in their review exploring separate visual pathways for perception and action, suggested we have a "what" pathway and a "how" pathway, arising due to what is required of the outputs we need to produce for the tasks we are attempting to successfully complete. Studies in animals and humans have demonstrated that tasks do indeed shape neural representations (Yang and Maunsell, 2004). How the brain efficiently uses task-based information to shape neural responses is less clear. Drawing inspiration from the field of computer science, we can examine a single particular task within a broader spectrum of tasks. Zamir et al. (2018) does this by modeling the structure of the space of different visual tasks using transfer learning, a technique done to use a previously trained model to effectively perform a new task, and creating a taxonomy that identifies useful relationships among these tasks. This approach aims to develop an optimal and efficient model that capitalizes on redundant information across tasks, using it to benefit other tasks. Similarly, Wang, Tarr, and Wehbe (2019) adopt this theoretical perspective in the area of neural perception, positing that the brain utilizes these inter-task relations to optimize computational processes. This neural taskonomy proposal predicts that different types of information are processed within common brain regions when this organization will lead to computational benefits. Conversely, when separate computational mechanisms for different types of information should lead to improvements in performance, those types of information should be processed by separate neural substrates.

In line with this, multiple research groups have shown the benefits of using neural networks to understand functional specialization (Dobs et al., 2022; Schwartz et al., 2023b). For instance, Dobs et al. (2022) demonstrated that in deep convolutional neural networks (DCNN) that are dual-trained to recognize both faces and

objects, the DCNN branches into two processing streams – one specialized for faces and the other specialized for objects. Additionally, DCNNs have been used to understand task representations in the brain as well (Hong et al., 2016). In DCCNs trained to perform category recognition on ImageNet (Krizhevsky, Sutskever, and Hinton, 2012), the model suggested that not only identity-preserving transforms can be maintained, but also the ability to build various transform representations related to orientation and position as well (Hong et al., 2016). Their models were able to explain neural responses both in areas IT and V4 in macaques. However, the study did not test if the task itself was necessary to develop these representations within the models. Thus, it cannot be concluded if this is due to optimizing the model for the task.

How do tasks influence visual perception at the behavioral level as well? In a recent study, Dobs et al. (2023) shed light on key behavioral aspects of human face perception by examining DCNNs. Specifically, in models trained to recognize face identity, the researchers found a face inversion effect, mirroring the phenomenon observed in human face perception. Furthermore, aligning with human behavioral patterns, this effect was present exclusively in DCNNs trained for face identity recognition, as opposed to those optimized for face detection or general object recognition. Hong et al. (2016) also found that the decoding patterns of the neural populations of IT in macaques were consistent with human performance on behavioral tasks involving object properties.

Although unsupervised and semi-supervised models similarly predict neural responses, when compared to their unsupervised counterparts, supervised models

tend to surpass unsupervised and semi-supervised models in terms of their behavioral consistency with human observations (Zhuang et al., 2021). Improved unsupervised models, however, are beginning to outperform their supervised counterparts (Margalit et al., 2024). Furthermore, supervised models have some of the same problems as unsupervised models as well as additional limitations. Xu and Vaziri-Pashkam (2021) investigated the performance of 14 different convolutional neural networks (CNNs) and compared them to fMRI data using representational similarity analysis (RSA, Kriegeskorte and Kievit, 2013). The researchers found that although the models show significant correspondence to lower-level visual representations, they were unable to fully capture downstream representations from LOTC and VOTC. This emphasizes that although the models do share similarities with the brain, they do not have a perfect one-to-one correspondence.

1.3 The functional distinctions of the dorsal, ventral, and lateral pathway

As mentioned previously, the LGN and V1 are organized into distinct layers, consisting of magno cells, parvo cells, and konio cells. These three cell types make up three cellular pathways that differ in the types of input statistics they represent. Why are these cells and the input statistics they represent each organized into a different stream in this particular way? This might be driven by the need to support distinct sets of behavioral functions. The interplay between input statistics and task-based constraints may provide insights into how category-selectivity, size effects, and the organization of multiple higher-order pathways emerge.

1.3.1 Ventral versus dorsal distinctions

Mishkin, Ungerleider, and Macko (1983) traced through the visual system to describe the pathways along the inferior temporal areas (ventral pathway) and occipitoparietal areas (dorsal pathway). Initial investigations, predominantly conducted on monkeys, unveiled compelling evidence of lesion-related deficits. For example, when removing the bilateral area TE or the anterior inferior temporal cortex, monkeys were impaired when performing an object discrimination task. More specifically, in the experiment a monkey was familiarized with one object from a pair of objects beforehand, and then was rewarded for choosing the unfamiliar object from the pair after. Conversely, removal of the bilateral posterior parietal cortex disrupted landmark discrimination tasks, impeding the monkeys' ability to discern food walls in proximity to the landmark. Mishkin, Ungerleider, and Macko (1983) concluded that the ventral pathway is involved in processing object qualities, while the dorsal pathway processes the object's spatial location.

Goodale and Milner (1992) went on to expand and confirm this dissociation with human studies, demonstrating a distinction between perceptual identification of objects and visually guided actions directed towards objects. They revised the initial idea of a 'what' versus 'where' functional dichotomy to instead a distinction of 'what' versus 'how'. For example, patients with object ataxia (i.e. Balint's syndrome), a higher-order visual deficit related to misreaching when attempting to complete visual goals, cannot reach for an object in the correct direction. However, in addition to this, they also show deficits positioning their hands and fingers at the right orientation, and modifying their grasp to correctly fit the size of the object (Perenin and Vighetto, 1988; Goodale and Milner, 1992). One patient (D.F.) with visual form agnosia, the inability to recognize objects, and whose lesions most

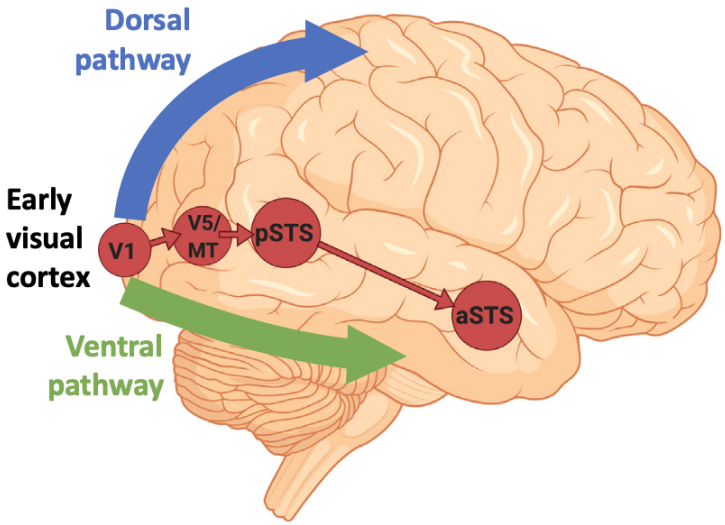

FIGURE 1.2: Depicts the three pathways of the visual system: dorsal, ventral, and lateral. Adapted from Pitcher and Ungerleider (2021)

likely are located in the ventral pathway, completed a series of tasks to evaluate the 'what' versus 'how' distinction. D.F. was unable to recognize the object along with various other features including size, shape, and orientation. When asked to demonstrate the width of the object with their hands, they said they were unable to do so. However, when asked to reach for the object, they were able to modify their grip and accurately locate the object to pick it up (Goodale et al., 1991). These patient deficits highlight the role that action-relevant information (e.g., 'how') plays in spatial vision.

1.3.2 Ventral versus lateral distinctions

More recently, Pitcher and Ungerleider detailed the argument for a three pathway model with distinct functions related to higher order vision. In the model, the dorsal pathway still corresponds to the 'how' pathway (e.g. processing location and actions related to objects) and the ventral pathway still corresponds to the

'what' pathway (e.g., processing the identity of visual objects). This third pathway (deemed the lateral pathway) includes the inferior occipital gyrus, area MT, and the STS (Figure 1.2 of the three pathway model, adapted from Pitcher and Ungerleider, 2021). The lateral pathway may be specialized to process dynamic aspects of social perception. Studies in macaques demonstrate a cortical pathway from V1 to area MT to the STS, bypassing the ventral pathway (Ungerleider and Desimone, 1986). These align with findings in humans that show a white matter pathway that projects to STS from area MT that is independent of the ventral pathway as well (Gschwind et al., 2012). The STS, recognized for its significance in motion-related processing (Grossman et al., 2000; Pitcher and Ungerleider, 2021), notably lacks the contralateral visual field biases observed in the ventral pathway (Hemond, Kanwisher, and Beeck, 2007). Findings from macaque studies indicate that both area MT and the STS boast larger receptive fields (Komatsu and Wurtz, 1989), encompassing broader regions of the visual field. This phenomenon aligns well with their roles in motion processing, underscoring the adaptive nature of their receptive field configurations and need to integrate across visual fields. Social perception often involves observing the complete visual field due to the many components typically involved in social interaction. This third pathway is also known for its role in multimodal integration (Anzellotti and Caramazza, 2017; Hasan et al., 2016). Given its association with dynamic information processing, the integration of various sensory modalities, such as visual and auditory, should be imperative for the effective analysis of social interactions.

While there has been widespread agreement regarding the existence of a lateral pathway in visual processing, early research predominantly emphasized the

differentiation between the ventral and dorsal pathways, leaving the role of the lateral pathway somewhat less clear. Multiple selective areas have been found in the STS in humans, and the macaque homolog (Tsao et al., 2006; Yang and Freiwald, 2021), during visual perception. These include selectivity for biological motion (Grossman et al., 2000), body parts including hands (Grosbras, Beaton, and Eickhoff, 2012), faces (Anzellotti and Caramazza, 2017) including eye gaze direction (Nummenmaa and Calder, 2009), and social interactions (Isik et al., 2017). FMRI studies using localizer tasks and point-light display stimuli of humans walking, dancing, and performing other typical movements identified the STS as selective for body movement (Grossman et al., 2000). These studies used randomly moving points while keeping other low-level features constant as a control, finding greater activation for the biological motion point-light displays in the pSTS (Grossman et al., 2000). Neighboring regions of the pSTS also respond to biological motion (Peelen, Wiggett, and Downing, 2006), and rTMS to the pSTS disrupts biological motion recognition (Grossman, Battelli, and Pascual-Leone, 2005).

Furthermore, the STS is thought to hold considerable importance in action recognition. Actions inherently include motion, and involve the movement of various body parts. Given its connections to area MT and its selectivity to biological motion, the STS emerges as a plausible hub for action perception. The LOTC which partially includes the pSTS is also part of the lateral pathway. While the ventral portion of the occipital temporal cortex may be predominately involved in the recognition of an object, the lateral portion may be predominately involved in action-related representations of the object (Wurm and Caramazza, 2022).

Finally, the lateral pathway has been thought to have very distinct roles in face perception. In the 1980s, Bruce and Young made a strong argument for a cognitive

model of face recognition, in which, after structural encoding of the face, the model branches into two discrete pathways. One pathway is specialized for identifying and generating the name of a face, while the other pathway is involved in facial expression, speech analysis, and gaze direction. After the development of fMRI in the 1990s, Haxby et al. (2001) came up with a neurocognitive model of face perception that dovetailed nicely with Bruce and Young's cognitive model. The Haxby, Hoffman, and Gobbini (2000) model states that early visual features are processed in the inferior occipital gyrus, where the OFA is located, which is followed by a branching into two different pathways. Invariant aspects of face processing like recognizing the identity of a face takes place in the ventral pathway which includes the FFA. This can then be extended to include anterior temporal regions like the ATL which is involved in person knowledge. Research using multi-voxel pattern analysis (MVPA) demonstrated that identity information can be decoded from neural responses in OFA and FFA (Natu et al., 2010; Nestor, Plaut, and Behrmann, 2011; Anzellotti, Fairhall, and Caramazza, 2013; Anzellotti and Caramazza, 2016; Dobs, Bülthoff, and Schultz, 2016), and fMRI adaptation studies find higher responses to different identities compared to the same identity (Winston et al., 2004). An fMRI study that used MVPA was also able to encode abstract information related to face identity in the ATL (Wang et al., 2017). However, the ATL can be difficult to evaluate for most studies due to signal dropout in this area. Changeable features of a face like facial expression, eye gaze, and lip movement are processed in the pSTS in the lateral pathway of the brain, and is independent from invariant recognition processing. Evidence shows the pSTS responds selectively to faces as well as point-light displays of facial motion (Andrews and Ewbank, 2004; Atkinson, Vuong, and Smithson, 2012). Moreover, the patterns of activity within this region

encode critical information regarding the emotional valence of facial expressions (Peelen, Atkinson, and Vuilleumier, 2010; Skerry and Saxe, 2014). Additionally, faces showing dynamic expressions do not evoke increased responses in OFA and FFA to the same degree as in pSTS (Pitcher et al., 2011). Taken together, this evidence supports the involvement of OFA and FFA in the recognition of identity, and the involvement of pSTS in the recognition of expressions. However, the evidence presented in support of the classical view of face perception does not rule out a role of the OFA and FFA in expression recognition, nor does it rule out a role of the pSTS in the recognition of identity. In sum, the evidence is not sufficient to demonstrate a separation between the mechanisms involved in the recognition of face identity and facial expressions.

Neuroimaging studies in the past decade began to reconsider the classical view of face perception. Identity information was found to be reliably decoded from the face-selective posterior superior temporal sulcus (fs–pSTS) in addition to the FFA (Anzellotti and Caramazza, 2017; Hasan et al., 2016; Dobs et al., 2018). In fact, one study demonstrated that identity could be decoded with greater accuracy from the pSTS than from both the OFA and FFA when the task at hand for the subject was to perform identity recognition (Dobs et al., 2018, Fig 6). Additionally, damage to the pSTS has been shown to lead to impairments in recognizing face identity across different facial expressions (Fox et al., 2011), underscoring a causal role for the pSTS in identity recognition. Conversely, in the study by Skerry and Saxe (2014) that decoded valence of facial expressions in pSTS, the authors were also able to decode the facial expression valence in the ventral pathway regions OFA and FFA. Furthermore, fMRI adaptation studies have demonstrated a release from adaptation for changes in facial expressions within the FFA as well

(Xu and Biederman, 2010). Even though deficits of identity recognition can spare expression recognition (Etcoff, 1984; Young et al., 1993), such double dissociations might occur at later stages of processing and do not exclude substantial integration between expression and identity recognition at earlier stages (Calder and Young, 2005). Coinciding with this, prosopagnosics often show some amount of impairment for expression recognition as well (Calder and Young, 2005).

These findings contradicted in part the previously indicated roles of the ventral and lateral pathways for face perception. Duchaine and Yovel (2015) proposed a revised model where the pathways interact, and the ventral pathway includes the OFA, the FFA, and the face-selective ATL, and the lateral pathway includes the face-selective pSTS, the face-selective aSTS, and the face-selective inferior frontal gyrus (IFG). The ventral pathway preferentially responds to form information and, thus, is important for recognizing invariant features where shape and form may have a greater weight. In line with this, regions in the ventral pathway respond to texforms, synthetic stimuli that contain the same mid-level texture and form information of an object while keeping the object unrecognizable (Long, Yu, and Konkle, 2018). The lateral pathway receives form information as well, but also receives motion information. This is important for dynamic features that may play a larger role in aspects like facial expression.

This evidence is part of a broader context of findings suggesting that dynamic information contributes to the recognition of person identity (Yovel and O'Toole, 2016). Dynamic information is used to recognize face identity (O'Toole, Roark, and Abdi, 2002). It is given more importance when the shape of the face is less reliable (Dobs, Ma, and Reddy, 2017) and when the face is familiar (Butcher and Lander, 2017). Person identity can also be recognized from gait presented with

point light displays (O'Toole et al., 2011). As mentioned above, dynamic faces also evoke stronger responses in the fs–pSTS (Pitcher et al., 2011). However, the view that the pSTS is specialized for the processing of dynamic stimuli (Bernstein and Yovel, 2015) does not fully account for the empirical data. Identity information can also be decoded from the fs–pSTS after presentation of static face images and voices (Anzellotti and Caramazza, 2017; Hasan et al., 2016). Rather than encoding exclusively dynamic information, pSTS appears to integrate form, motion, and sound (e.g., voices).

Collectively, these observations indicate that the evidence supporting separate processing systems for identity and expression is not strong. However, they do not directly refute the classical view that differentiates the ventral and lateral pathways, which are thought to process invariant and changeable features, respectively.

1.3.2.1 Utilizing the case of face perception

A set of predictions are tested and reviewed hereby following the trail of these surprising findings in face perception (Skerry and Saxe, 2014; Anzellotti and Caramazza, 2017; Dobs et al., 2018). A classical theory of face perception holds that information about identity and information about expressions is processed by different streams. In particular, representations of identity that are invariant to changes of expressions are computed by discarding expression information, and representations of expressions that are invariant to identity are computed by discarding identity information (Bruce and Young, 1986). In terms of neural implementation, in this view, information about identity is processed by the ventral pathway, and information about expression is processed by the lateral pathway (Haxby et al.,

2001). In contrast with this theory, recent evidence has shown that lateral regions (and in particular the fs–pSTS) also encode some information about face identity (Anzellotti and Caramazza, 2017; Dobs et al., 2018). Informed by the taskonomy proposal from Zamir et al. (2018) and its relationship to functional organization, it can be hypothesized that this is the result of constraints at the level of the outputs. In particular, an alternative account to the classical theory by Bruce and Young (1986) is presented. In this alternative account, information about identity and expressions is disentangled: separating out information related to identity also helps to separate out information related to expression, and vice versa.

1.4 Overview of Dissertations Chapters

The described alternate account results in a set of empirical predictions that can be evaluated. First, recognizing face identity should not require discarding information about expression, and vice versa. Instead, face identity recognition models might even learn spontaneously, to some extent, how to separate out information about expressions. This is tested in Chapter 2. Second, information about identity and expression should be represented within common brain regions for static face images. Rather than separate specialized mechanisms for identity and expression processing, information about identity and expression should be encoded to similar degrees in both ventral and lateral temporal brain regions. This is tested in Chapter 3. Third, this organization should not be limited to static stimuli (images), but it should also apply to the perception of dynamic stimuli (videos). This is then tested in Chapter 4. A set of studies are proposed that are designed to evaluate these predictions.

Chapter 2 presents computational findings published in Schwartz et al. (2023a), evaluating a proof-of-concept that may help to undermine the classical view of face perception. To determine if face identity and facial expression processing are separate mechanisms, it is first necessary to test whether discarding irrelevant task information is necessary for accurate face identity and facial expression recognition. If this is the case, expression information should decline when learning identity information and vice versa. Features from DCNNs trained to recognize expression were evaluated to determine if they could be used to recognize face identity, and vice versa, or if performance for the irrelevant task declined. Additionally, two other analyses were implemented: a network trained to recognize scenes and an untrained network to act as a control.

This computational study showed that integrated processing for identity and expression recognition is possible, and that discarding information for one task to complete the other is not a necessity. This could potentially indicate common mechanisms for identity and expression recognition. However, there is a weaker version of the classical theory of face perception that can still be in line with these computational findings. If this weaker version of the classical theory is correct, identity-specific regions should have representational dissimilarity matrices (RDMs) mostly driven by identity model features, and expression-specific regions should have RDMs mostly driven by the expression model features. As a next step, neural data from humans was used to test the relative amounts of identity and expression information within each region. Chapter 3 presents neural findings published in Schwartz et al. (2023b) that address the open question of fMRI decoding studies that find both face identity and facial expression information in shared brain regions. The models from Chapter 2 were used here to investigate intracranial

electrocorticography (iEEG) data and evaluated the presence of face identity and facial expression information in electrodes located in both the ventral and lateral temporal pathways.

The studies in Chapters 2 and 3 demonstrated the phenomenon of shared identity and expression representations and the process of disentanglement in the analysis of static face stimuli within DCNNs (Schwartz et al., 2023a), and importantly, both identity- and expression-trained image models showed similar correlations across ventral and lateral temporal regions (Schwartz et al., 2023b). However, it could be possible that static and dynamic properties are processed differently, such that even if face identity and facial expression information are processed jointly via static features, they may not be when using dynamic information. Thus, this hypothesis had to still be tested in dynamic stimuli by comparing both identity- and expression-trained DCNN models for video recognition to the ventral and lateral temporal pathways.

Using fMRI, neural responses to dynamic face stimuli varying in identity and expression were collected. Typically, DCNN architectures for video recognition employ two processing streams: one processing individual frames (spatial or static stream), and the other processing optic flow (temporal or dynamic stream). These models were trained and were used in conjunction with the fMRI data to test two predictions. It was first tested whether there are separate neural representations for face identity and facial expression information for dynamic face stimuli, which was anticipated to not be the case. As will be discussed in Chapter 4, RDMs from two-stream models trained on dynamic stimuli for identity recognition and two-stream DCNN models trained on dynamic stimuli for expression recognition similarly explained both ventral and lateral temporal brain regions. Why do we have

two pathways involved in processing face stimuli then? Do these pathways have distinct functional roles and what may these roles be? The rest of Chapter 4 continued to investigate functional distinctions between the ventral and lateral temporal pathways. Since the relative contribution of dynamic identity and expression features did not distinguish between the ventral and lateral regions, the ventral and lateral regions were tested to see if they may differ along a different dimension. The delineation between the ventral and lateral temporal pathways might be more accurately attributed to a differentiation between static and dynamic information. For example, static features may include texture and shape, while dynamic features may relate to motion (e.g., velocity and direction of motion).

The lateral pathway is thought to have a greater role in dynamics compared to the ventral pathway due to its role in social processing (Ungerleider and Desimone, 1986). If the two-stream models — which have a spatial component and a temporal component (that will be referred to as DCNN models themselves) - differed in their ability to explain neural response patterns within the ventral and lateral temporal regions, this would support a hypothesized static and dynamic distinction. This was evaluated by comparing neural RDM and model RDMs from two-stream DCNNs for video recognition. The results suggested no difference in the relative contribution for each model across the ventral and lateral temporal regions, supporting the conclusion that joint representations for identity and expression are found in both static and dynamic face stimuli, but that the pathways are not necessarily distinguished via static and dynamic feature distinction.

Overall, the findings presented in this thesis tested multiple hypotheses. The predictions tested are all aimed at understanding how face perception and specific visual properties essential for evaluating faces are organized in the brain. This

is done using a combination of behavioral, neuroimaging and machine learning methods.

Chapter 2

Challenging the Classical View: Recognition of Identity and Expression as Integrated Processes

The contents of this chapter have been published in the following research articles:

Challenging the Classical View: Recognition of Identity and Expression as Integrated Processes

Emily Schwartz [1,†], Kathryn O'Nell [2,†], Rebecca Saxe [3] and Stefano Anzellotti [1,]

[1] Department of Psychology, Boston College, Boston, MA, United States 02467

[2] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755

[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Brain Sci. 2023, 13(2), 296; https://doi.org/10.3390/brainsci13020296

2.1 Introduction

The human ability to recognize face identity and facial expression is used as a compass to navigate the social environment (Anzellotti and Young, 2020). Identity recognition enables us to acquire knowledge about specific individuals that we can retrieve in future encounters (Bruce and Young, 1986; Mende-Siedlecki, Cai, and Todorov, 2012). Expression recognition helps us to infer the emotional states of an individual (Wagner, MacDonald, and Manstead, 1986; Wu and Schulz, 2018; Saxe and Houlihan, 2017) and predict their future actions and reactions. However, face identity and facial expression coexist within a face image. Information about each property needs to be extracted without being confused with the other.

The classical view on the recognition of face identity and facial expression proposes that identity and expression are processed by distinct pathways (Bruce and Young, 1986; Haxby, Hoffman, and Gobbini, 2000). In this view, the pathway specialized for identity discards expression information, and the pathway specialized for expression discards identity information. With respect to the underlying neural mechanisms, it has been proposed (Haxby, Hoffman, and Gobbini, 2000) that face identity is recognized by a ventral temporal pathway, including the occipital face area (OFA) Gauthier et al., 2000 and the fusiform face area (FFA) Kanwisher, McDermott, and Chun, 1997. By contrast, facial expression is recognized by a lateral pathway (Haxby, Hoffman, and Gobbini, 2000), including the face-selective posterior superior temporal sulcus (fs–pSTS; Hoffman and Haxby, 2000).

In support of this view, several lines of evidence show that ventral occipitotemporal regions, such as OFA and FFA, play an important role in the recognition of face identity. Studies using fMRI adaptation show that changes in identity lead

to greater release from adaptation than changes in viewpoint (Xu and Biederman, 2010). Research using multi-voxel pattern analysis (MVPA) found that identity information can be decoded from responses in OFA and FFA (Natu et al., 2010; Nestor, Plaut, and Behrmann, 2011; Anzellotti, Fairhall, and Caramazza, 2013; Anzellotti and Caramazza, 2016; Dobs, Bülthoff, and Schultz, 2016). Structural connectivity measures reveal that congenital prosopagnosics (participants with congenital impairments for face recognition) present with reduced white matter tracts in the ventral occipitotemporal cortex (Thomas et al., 2009).

Other evidence indicates that fs–pSTS is a key region for the recognition of facial expression. The fs–pSTS responds selectively to faces (Andrews and Ewbank, 2004) and shows greater responses to moving faces than static faces (Pitcher et al., 2011). Furthermore, videos of dynamic facial expressions do not evoke increased responses in OFA and FFA to the same degree as in fs–pSTS (Pitcher et al., 2011). Additionally, the patterns of activity in this region encode information about the valence of facial expressions (Peelen, Atkinson, and Vuilleumier, 2010; Skerry and Saxe, 2014). Finally, patients with pSTS damage have deficits for facial expression recognition (Fox et al., 2011), providing causal evidence in support of the involvement of pSTS in facial expression recognition.

Nevertheless, there is also evidence that weighs against this view of separate representational streams. Previous work noted the lack of strong evidence in support of the classical view (Calder and Young, 2005). In particular, while findings that support the classical view indicate that the lateral pathway plays a role in expression recognition, they do not rule out the possibility that the ventral pathway might also play a role (Duchaine and Yovel, 2015). In the same manner, findings that suggest the involvement of the ventral pathway in identity recognition

do not rule out the possibility that the lateral pathway might contribute to identity recognition as well. Moreover, recent research directly shows that recognition of face identity and facial expression might be more integrated than previously thought. FMRI adaptation studies find release from adaptation for changes in facial expression in FFA (Xu and Biederman, 2010). Other work has shown that the valence of facial expression can be decoded from ventral temporal regions, including OFA and FFA (Skerry and Saxe, 2014; Kliemann et al., 2018). Duchaine and Yovel (2015) proposed a revised framework in which OFA and FFA are engaged in processing face shape, contributing to both face identity and facial expression recognition. At the same time, identity information can be decoded from fs–pSTS (Anzellotti and Caramazza, 2017; Hasan et al., 2016; Dobs et al., 2018). In fact, in one study, identity could be decoded with higher accuracy from fs–pSTS than from both OFA and FFA (Dobs et al., 2018, Figure 6), and two other studies demonstrated that identity could be decoded in fs–pSTS across faces and voices (Anzellotti and Caramazza, 2017; Hasan et al., 2016). Furthermore, pSTS damage leads to impairments for recognizing face identity across different facial expressions (Fox et al., 2011), suggesting that pSTS plays a causal role for identity recognition as well. Finally, animal studies recently identified the middle dorsal face area (MD) in macaque monkeys. Interestingly, this face-selective area was shown to encode information on both face identity and facial expression (Yang and Freiwald, 2021). Importantly, the area encodes identity robustly across changes in expression, and expression robustly across changes in identity (Yang and Freiwald, 2021), providing the strongest direct empirical challenge to the classical view.

The above evidence indicates that recognition of facial expression and face

identity are implemented by integrated mechanisms, and not by separate neural pathways. Here, we offer a computational hypothesis that can account for this phenomenon. Unlike the classical view, which suggests that information relevant to identity recognition should be shed as representations of facial expressions develop, we hypothesize that representations optimized for expression recognition contribute to identity recognition and vice versa. Moreover, this occurs because identity and expression are entangled sources of information in a face image, and disentangling one helps to disentangle the other (the "Integrated Representation of Identity and Expression Hypothesis"—IRIEH).

IRIEH leads to two non-trivial computational predictions. First, if recognition of face identity and facial expression are mutually beneficial, training an algorithm to recognize face identity might lead to the spontaneous formation of representations that encode facial expression information and, likewise, training a separate algorithm to recognize facial expression might lead to the spontaneous emergence of representations that encode face identity information. Second, if this phenomenon occurs because disentangling identity from expression helps to also achieve the reverse, then integrated representations would not arise because recognition of identity and expression rely on common features. On the contrary, features important for the recognition of face identity and features important for the recognition of facial expression should become increasingly disentangled and orthogonal along the processing stream.

In the present article, we tested 'in silico' these computational hypotheses inspired by the neuroscience literature. To do this, we analyzed representations of face identity and facial expression learned by deep convolutional neural networks (DCNNs). DCNNs achieve remarkable accuracy in image recognition tasks

(Krizhevsky, Sutskever, and Hinton, 2012; Parkhi, Vedaldi, Zisserman, et al., 2015), and features extracted from deep network layers have been successful at predicting responses to visual stimuli in the temporal cortex in humans (Khaligh-Razavi and Kriegeskorte, 2014) and in monkeys (Yamins et al., 2013; see Yamins and DiCarlo, 2016; Kietzmann, McClure, and Kriegeskorte, 2018 for reviews). Although artificially crafted stimuli ('metamers') have revealed differences between DCNNs and humans (Feather et al., 2019), DCNNs show similarities to human vision in terms of their robustness to image variation (Kheradpisheh et al., 2016). Recent work used DCNNs to test computational hypotheses of category-selectivity in the ventral temporal cortex (Dobs et al., 2022). In this article, we follow a similar approach and argue that a clearer understanding of representations of face identity and facial expression within DCNNs can serve as the foundation for future research on face representations in the brain.

To test our two predictions, we studied whether features from hidden layers of a DCNN trained to recognize face identity (from here onward the "identity network") could be used successfully to recognize facial expression (see Colón, Castillo, and O'Toole, 2021 for a related analysis). Symmetrically, we evaluated whether features from hidden layers of a DCNN trained to recognize facial expression (the "expression network") could be used to identify face identity. In line with our anticipated results, we found that in a DCNN trained to label one property (i.e., expression), the readout performance of the non-trained property (i.e., identity) was not just preserved, but improved, from layer to layer. This was in stark contrast with classical theories of abstraction in visual processing that suggest that information about task-orthogonal information is progressively discarded (Posner, 1970; Thornton, 1996; Kanwisher, Yin, and Wojciulik, 1999).

Finally, we investigated the relationship between features encoding information that distinguish between identities and expressions across different layers of the DCNNs. We demonstrated that identity-discriminating features and expression-discriminating features became increasingly orthogonal over the network layers.

2.2 Materials and Methods

2.2.1 Stimuli

The identity network was trained to label identities using face images from the Large-Scale CelebFaces Attributes (CelebA) dataset (Liu et al., 2015). CelebA is made up over 300,000 images. To match the dataset training size used for the expression network (see below), a subset of CelebA was used. The subset of the dataset contained 28,709 images for training and an additional 3589 images for testing (these latter images were used to test the performance of the network after training), and contained 1503 identities. These identities were randomly chosen, with at least 20 images per identity. All images were cropped to 178×178 pixels, resized to 48×48 pixels, and converted to grayscale by averaging pixel values of the red, green, and blue channels.

The expression network was trained to label facial expressions using the face images in the Facial Expression Recognition 2013 (FER2013) dataset (Goodfellow et al., 2013). The dataset contained 28,709 images for training and an additional 3589 images labeled as 'public test' (these latter images were used to test the performance of the network after training and to compare it to human performance). All images were originally sized 48×48 pixels and grayscale.

A network trained to recognize scenes was also implemented for comparison. The UC Merced Land Use dataset (Yang and Newsam, 2010), which consisted of 2100 images of 21 classes, was used to train the network to label land images. All images were resized to 48×48 pixels and converted to grayscale by averaging pixel values of the red, green, and blue channels.

The performance for each network was tested on stimuli from an independent dataset: the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist, Flykt, and Öhman, 1998). The KDEF dataset consisted of 4900 images depicting 70 individuals showing 7 different facial expressions from 5 different angles, each combination photographed twice. We used the frontal view images and those with views rotated by 45 degrees in both directions (left and right). Images were sized 562 (width) $\times$ 762 (height) and in color (RGB). For network transfer testing, in order to match the format of the training images, all KDEF images were converted to grayscale, cropped to squares, and downsampled to 48×48 pixels. The images were converted to grayscale by averaging pixel values of the red, green, and blue channels. As the positioning of the face within the image was consistent across KDEF images, the rectangular images were all cropped to the same 388×388 pixel region around the face. Example face images from the KDEF dataset, and example images similar (due to copyrights) to the CelebA and FER2013 datasets can be seen in Figure 2.1. Visual inspection confirmed that the face was visible in each KDEF image after cropping. Table 2.1 provides specific details about training and validation/testing set sizes.

2.2.2 Neural Network Architecture

Using Pytorch (Paszke et al., 2017), a densely-connected deep convolutional neural network (DenseNet) was implemented, consisting of 1 convolutional layer, 3

FIGURE 2.1: Face image examples. Top: naturalistic face images, similar to those from the CelebA and FER2013 datasets. Bottom: selected images from KDEF dataset (AF01AFHR, AF02SUHL, AF05AFS, AM01ANS, AM10HAHL, AM27NEHR).

TABLE 2.1: Dataset information.

Dataset	Training Set Size	Testing/Validation Set Size	Stimulus Type
CelebA (Liu et al., 2015) [1]	28,709	3589	Face
FER2013 (Goodfellow et al., 2013)	28,709	3589	Face
UC Merced Land Use (Yang and Newsam, 2010)	1890	210	Scene
KDEF (Lundqvist, Flykt, and Öhman, 1998)	2520–2646 [2]	294–420 [2]	Face

[1] Only a subset of the CelebA dataset was used to train and test the identity model. [2] Number of images used for training and held-out for testing depended on labeling task.

dense blocks, and 1 fully connected linear layer (Figure 2.2). A DenseNet architecture was selected since it has been shown to yield high performance on a variety of tasks (Huang et al., 2017), and because it features connections between non-adjacent layers, bearing a closer resemblance to the organization of the primate visual system (Van Essen, Anderson, and Felleman, 1992). The convolutional layer consisted of 64 channels of 2D convolutions using a 3×3 kernel and padding = 1. Each dense block consisted of 3 densely connected convolutional layers with kernel size = 3, stride = 1, and padding = 1. Each layer in the dense block produced 32 channels of output. Therefore, the number of input channels for the first layer in a dense block was equal to the number of output channels of the previous layer outside the dense block (i.e., for the first layer of the first dense block it was equal to 64: the number of output channels of the first convolutional layer). The number of input channels for each subsequent layer in each dense block increased by 32. This choice is widely used and featured on publicly available DenseNet implementations (i.e., `https://github.com/pytorch/vision/`

TABLE 2.2: Hyperparameters of the networks' layers.

Layer Name	Kernel Size	Input Channels	Output Channels
Conv1	3×3	1	64
Dense1-1	3×3	64	32
Dense1-2	3×3	96	32
Dense1-3	3×3	128	32
Transition1	2×2	160	80
Dense2-1	3×3	80	32
Dense2-2	3×3	112	32
Dense2-3	3×3	144	32
Transition2	2×2	176	88
Dense3-1	3×3	88	32
Dense3-2	3×3	120	32
Dense3-3	3×3	152	32
Avg pooling	8×8	152	32
FC	1×1	32	1

`blob/master/torchvision/models/densenet.py`, accessed on 1 November 2019).

Each dense block (except the last) was followed by a transition layer that received, as input, the outputs from all layers of the dense block plus the layer preceding the dense block, and produced an output with half the number of channels using a max pooling with a 2×2 kernel. The last dense block was followed by an average pooling with an 8×8 kernel and then by a fully connected linear layer. In sum, the number of input and output channels for the 13 layers of the network can be seen in Table 2.2.

All layers used rectified linear units (ReLU) as nonlinearity for an activation function. All layers in the dense blocks and all transition layers used 2D dropout with a dropout probability $p = 0.1$ (Dahl, Sainath, and Hinton, 2013). All convolutional layers were followed by batch normalization (Ioffe and Szegedy, 2015).

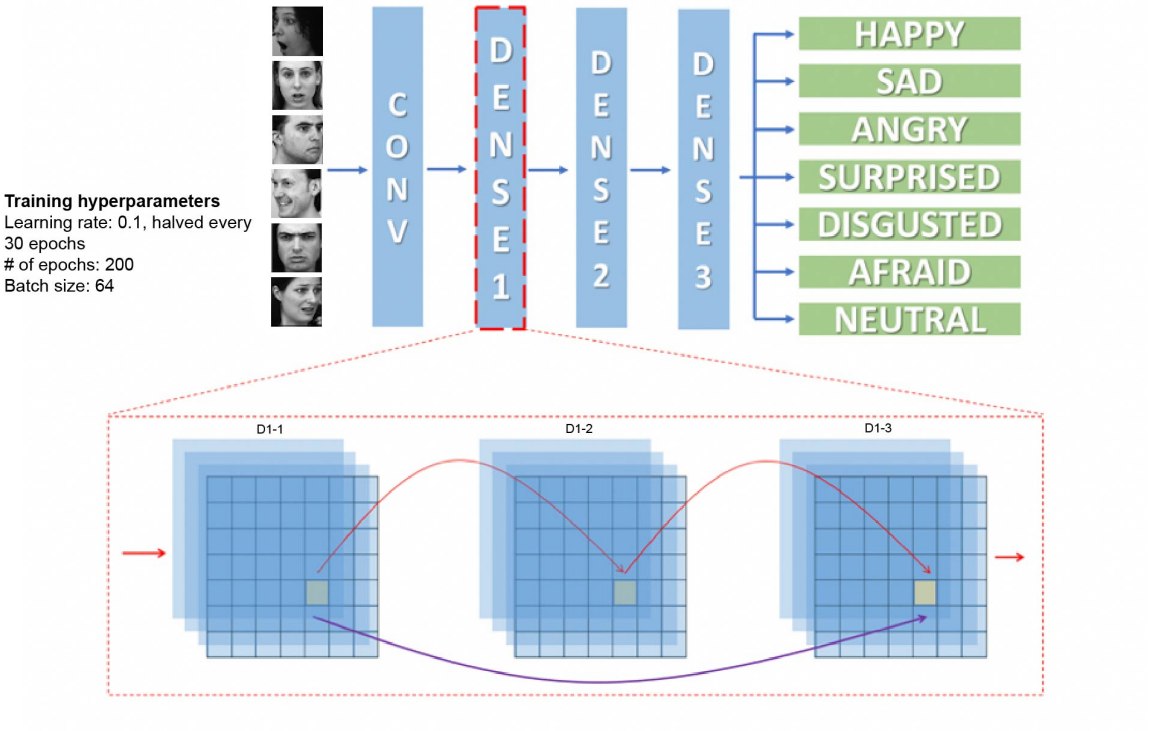

FIGURE 2.2: Neural network architecture. Top: Each network consists of a convolutional layer, three dense layers, and a fully-connected (FC) linear classifier. Expression classification is used as an example here. Bottom: Single dense block; red arrows represent connections that would exist in a typical convolutional neural network, the purple arrow represents connections that are unique to the densely-connected network. Selected images from KDEF dataset: AF01AFHR, AF02SUHL, AF05AFS, AM01ANS, AM10HAHL, AM27NEHR.

2.2.3 Training and Validation

We evaluated 4 sets of networks: identity-trained, expression-trained, scene-trained, and untrained (randomly initialized weights). Each network described was implemented 10 times with random weight initialization to test the consistency of the results. We report the average accuracy across the 10 initializations, including the standard error of the mean in the figures as error bars.

Given 48×48 grayscale images in the CelebA dataset, the identity network was trained to recognize 1503 face identities varying in pose and age. The network was trained to minimize the cross-entropy loss between the outputs and true labels using stochastic gradient descent. The learning rate began at 0.1 and halved every 30 epochs. The training was run for 200 epochs, and images were presented to the network in batches of 64. The performance of the trained network was validated using an independent subset of CelebA that was not used for any of the training. The identity network labeled face identity with an accuracy of 26.5% on the held-out 'test' images (chance performance at 0.06%). The CelebA database did not include viewpoint labels, so we were unable to test cross-viewpoint validation performance.

The expression network produced an output of 7 values, one for each expression label in the dataset (surprised, angry, fearful, disgusted, sad, neutral, and happy). The network was trained to minimize the cross-entropy loss between the output and the true labels using stochastic gradient descent, with a learning rate starting at 0.1 and halved every 30 epochs. The training was run for 200 epochs, and images were presented to the network in batches of 64. After training, the accuracy of the expression network was validated using an independent subset of the FER2013 dataset that was not used for training (the images marked as 'PublicTest').

The network achieved an accuracy of 63.5% (chance performance at 14.2%), closely matching the reported human accuracy on the FER2013 stimuli (65%; Goodfellow et al., 2013). The FER2013 database did not include viewpoint labels, so we were unable to test cross-viewpoint validation performance.

The scene network was trained to recognize various land images. This network matched the architecture used for the identity and expression networks, and followed the same training and validation protocols. The trained network was able to label the validation set with an accuracy of 80.95%. The untrained network (with randomly initialized weights) used the same architecture as all other networks, but it did not undergo any training.

2.2.4 Transferring to KDEF

After training with each dataset was completed, the weights of each network were fixed ('frozen') to prevent further learning. Henceforth, we refer to a network that has the weights fixed after the initial training as a 'pre-trained network'. To test identity and expression labeling, we used a new dataset of images: the KDEF dataset (Lundqvist, Flykt, and Öhman, 1998), in which each image has both an identity and an expression label.

2.2.4.1 Labeling Identity across Expression and Expression across Identity

To evaluate whether the identity network could successfully perform the task it was trained for, we tested whether it could accurately label identity in the KDEF dataset. Then, we tested the identity network's performance at labeling expression. To assess the transformation of representations across different stages of the neural network, we evaluated the readout accuracy of identity and expression

for features extracted from different layers. For each of the 10 identity networks trained with the CelebA dataset, accuracy was evaluated for features extracted from the first convolutional layer, and for features extracted from the last layer in each dense block, after they had been summed with the inputs of the block. The outputs that the networks needed to produce for identity labeling and for expression labeling were different. For instance, the number of identity labels was different than the number of expression labels (70 v 7). To accommodate for this, we extracted the corresponding layer feature representations by running an image through the pre-trained model (up until the specified layer). We then ran the image's feature representation through batch normalization, ReLU, and an average pooling with an 8×8 kernel, followed by a fully connected linear layer that produced, as output, the identity or expression labels (referred to as the 'readout layer' from here on). Critically, these added fully connected readout layers achieved very different performances depending on the layer of the network that they were attached to (that is, depending on the nonlinear features that they received as an input). Readout performance was then tested on the held-out portion of the KDEF data. The performance of a linear layer trained directly on pixel values was used as a control.

We followed an analogous procedure for the expression network. First, we tested the expression network to ensure that it could accurately perform the expression recognition task on the KDEF dataset. Next, for each of the 10 expression networks, we used the same readout procedure as above to probe the accuracy of expression and identity labeling. To assess the transformation of representations across different stages of the neural network, accuracy was evaluated for features extracted from the first convolutional layer, and the last layer in each of the 3 dense

blocks, after they had been summed with the inputs of the block. As in the case of the identity network, the performance of a linear layer trained directly on pixel values was used as a control.

Due to the ability of these models to rely on low-level features, we partitioned the KDEF dataset into training and testing sets, and tested the models across different viewpoints. To look at cross-viewpoint generalization, the identity and expression networks' performances were tested with a readout layer trained using all but one of the viewpoints (frontal, 45 degree left, or 45 degree right), and accuracy was tested using the held-out viewpoint (as in Anzellotti, Fairhall, and Caramazza, 2013). Accuracy values for both identity and expression labeling were then averaged across the three conditions. This choice was made to provide a more stringent test of identity and expression recognition, as rotation in depth alters all parts of the face.

The added readout layers' performances were heavily dependent on the non-linear features received as inputs. If the added readout layers trained with a subset of the KDEF images could achieve high accuracy without needing the features from a pre-trained network, this should have been evident when they were attached to early layers of that pre-trained network (or when attached to layers of the untrained network, see below). When using features from late layers as compared to features from early layers of the pre-trained networks, accuracy improvements could not be due to the attached readout layer that was trained with a subset of KDEF images because the same readout layer was used for both early and late layers.

2.2.4.2 Labeling Identity and Expression Using Untrained and Scene Network Features

The procedure described above was enacted to evaluate the performance of identity and expression labeling on KDEF images using the following: (1) randomly initialized, untrained neural network weights and (2) scene-optimized neural network weights. KDEF images were run through the various networks and their feature representations were extracted at multiple layers. The same readout procedure was used to learn the identity and expression labels for the KDEF images. After training the readout layer only, identity and expression labeling performances on the various KDEF feature representations were obtained.

2.2.5 Overlap between Identity and Expression Features

If, as we predicted, information about the non-trained feature (i.e., identity for the expression network and expression for the identity network) was not discarded during training, there were two potential explanations. First, it could be that the same image features were important for classifying both identity and expression. Alternately, it could be that distinct image features were important for classifying identity and expression, and both were retained within the network. In this case, the presence of features that contributed to labeling the irrelevant task indicated that the abstraction-based model of feature representations in the brain was not supported by the kind of representations that were learned spontaneously by the deep convolutional neural networks. In order to dissociate these outcomes, we tested the congruence of the spaces spanned by the opposing identity and expression features in all 3 of the trained (identity, expression, and scene) networks.

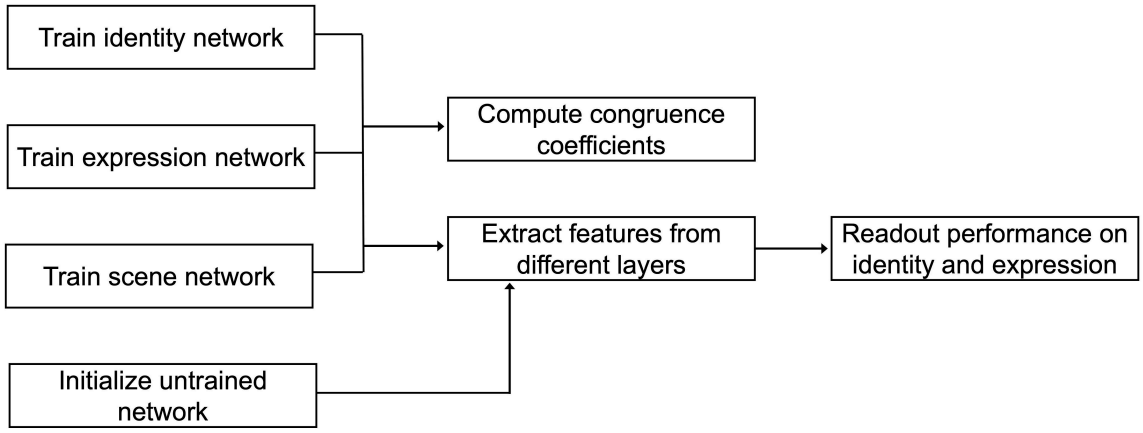

FIGURE 2.3: Analysis flowchart. An overview of the analysis steps performed.

To do this, we averaged a layer's responses across different expressions, obtaining an average response pattern across the layer features for each identity. Next, we used principal component analysis (PCA) to extract the 5 dimensions that explained most of the variation across identities. The same procedure was repeated by averaging layer responses across identities, obtaining an average response pattern for each expression, and ultimately 5 dimensions that explained most of the variation across expressions.

Finally, we used a congruence coefficient (introduced in Krzanowski, 1979) to evaluate the similarity between the spaces spanned by the features. Considering the matrix L_e of the loadings of principal components for expression on the layer features and the matrix L_i of the loadings of principal components for identity, we obtained the matrix $S = L_i L_e' L_e L_i'$ and measured overlap as the sum of the eigenvalues of S, which was equal to the sum of the squares of the cosines of the angles between all pairs of principal components where one component in the pair was for expression and the other was for identity (Krzanowski, 1979).

An overview representing the research procedure can be seen in Figure 2.3.

2.3 Results

2.3.1 Validation Performances of Trained Neural Networks

A densely-connected deep convolutional neural network (DenseNet, Huang et al., 2017, Figure 2.2) was trained to recognize face identity using a subset of the CelebA dataset. The network was able to label face identity with an accuracy of 26.5% on the held-out 'test' images (chance performance at 0.06%). A confusion matrix can be found in Appendix A (Figure A.1 A).

A DenseNet (Huang et al., 2017, Figure 2.2) was trained to recognize facial expressions (surprised, angry, fearful, disgusted, sad, neutral, happy) using over 28,000 facial expression images (FER2013). The network was able to label facial expression on the held-out 'test' images with an accuracy of 63.5% (chance performance at 14.2%). A confusion matrix can be found in Appendix A (Figure A.1 B).

A third DenseNet (Huang et al., 2017, Figure 2.2) was trained to label land images. The network was able to label the different scene categories on the held-out 'test' images with an accuracy of 80.95% (chance performance at 4.76%). A confusion matrix can be found in Appendix A (Figure A.1 C).

2.3.2 Neural Networks Trained to Recognize Identity Develop Expression Representations

Recognition of face identity across changes in viewpoint is notoriously difficult (Poggio and Edelman, 1990; Anzellotti, Fairhall, and Caramazza, 2013). Thus, we aimed to investigate the invariance of the identity network's face representations

across image transformations. To do this, we used images from the KDEF dataset that included frontal views, as well as 45 degree views (left and right) of the faces. We explored, across different viewpoints, whether the identity network could label both face identity and facial expression after the newly attached readout layer was trained using two of the three views, and then, tested with the held-out view.

The identity network generalized to the KDEF dataset for identity recognition. The network achieved an accuracy of 53.82% (chance performance at 1.42%) when testing on held-out viewpoints (Figure 2.4A, bottom left). The readout layers that received the identity network's extracted features as inputs achieved a higher accuracy for identity recognition when testing on a held-out viewpoint, compared to a fully connected linear layer that received pixel values of the KDEF images as inputs. Specifically, the linear layer that received the pixel values as inputs achieved an accuracy of 6.31%. By contrast, readout layers applied to the features from the convolutional layer, first, and second dense blocks yielded accuracy values of 9.61%, 11.91%, and 22.65% respectively (Figure 2.4 A, bottom left). Thus, accuracy increased from layer to layer.

Having established that the identity network successfully generalized to the KDEF dataset for the task it was trained to perform (identity recognition), we next studied whether the identity network developed features that could yield accurate expression recognition when testing on the held-out viewpoint. As detailed in the Methods section, in order to generate the 7 facial expressions as output (instead of the 70 face identity labels), a readout layer was attached to the outputs of a hidden layer of the pre-trained identity network, and then trained with KDEF images consisting of two viewpoints to label expression. Critically, the identity network

weights were fixed at this stage, and only the weights of the newly attached readout layer would be able to change.

When using identity-trained weights, expression classification of images from the KDEF dataset across different viewpoints (44.37%, Figure 2.4A, bottom right) was greater than chance. By contrast, a linear layer that received pixels as inputs achieved an accuracy of 20.40%. Importantly, as in the case of identity classification, the accuracy of the network increased from early layers to late layers. Readouts of features extracted from the initial convolutional layer, and first and second dense blocks of the identity network yielded accuracy values of 17.61%, 16.67%, and 23.02%, respectively, when labeling expression, finally reaching 44.37% in the third dense block, as mentioned previously (Figure 2.4A, bottom right). A large increase in accuracy was observed in the second and third dense blocks, paralleling the increase in accuracy observed for identity labeling at the same processing stages. This indicated that in the network trained to label identity and then tested on expression recognition, the findings deviated from the predictions of the classical view (Figure 2.4A, top right).

2.3.3 Neural Networks Trained to Recognize Expression Develop Identity Representations

In parallel to the identity network analysis, we investigated the invariance of the expression network's face representations across image transformations. The expression network was not trained to recognize identity across different viewpoints, but it was trained to label facial expression across viewpoints. Could the features it developed for labeling facial expression be used to support the demanding task of view-invariant identity recognition? To address this question, we again used

images from the KDEF dataset showing a frontal view as well as 45 degree views (left and right) of the faces. We investigated whether the expression network could label facial expressions and identities when the newly attached readout layer was trained with two of the three views, and then tested with the held-out view.

The final accuracy at cross-viewpoint expression labeling on the KDEF images was high (53.43%, Figure 2.4B, bottom left), showing that the expression network generalized successfully to the new dataset. As expected, labeling accuracy increased from layer to layer of the expression network. A readout layer applied directly to the pixels of the KDEF images obtained an accuracy of 20.40% for expression classification, but subsequent layers were necessary to reach the final accuracy of 53.43%. Features extracted from the initial convolutional layer, and first and second dense blocks of the expression network yielded accuracy values of 17.22%, 17.31%, and 24.93%, respectively, when labeling expression (Figure 2.4B, bottom left). Similar to the patterns in accuracy that were found when using the identity network, a large increase in accuracy was observed in the third dense block with a final accuracy of 53.43% (Figure 2.4B, bottom left).

Next, the expression network weights were used to label identity. In order to generate the 70 identities as output (instead of the 7 facial expression labels), a readout layer was attached to the outputs of a hidden layer of the expression network pre-trained with the FER2013 dataset, and trained with images consisting of 2 viewpoints to label identity. The expression network weights were fixed at this stage, and only the weights of the newly attached readout layer could change.

Final identity classification of images from the KDEF dataset (20.2%, Figure 2.4B, bottom right) was greater than chance. By contrast, linear classification using

the pixels as input achieved an accuracy of only 6.31%. Importantly, readout accuracy increased from early to late layers in the network. Features extracted from the initial convolutional layer, and first and second dense blocks of the expression network, yielded accuracy values of 9.56%, 6.32%, and 14.81%, respectively, when labeling identity, reaching a final accuracy of 20.20% in the third dense block (Figure 2.4B, bottom right). An increase in accuracy was observed in the second and third dense blocks. Although to a smaller degree, this paralleled the increases in accuracy observed for expression labeling at the same processing stages. This finding was in contrast with the decrease in identity information that would have been expected in the classical view (Figure 2.4B, top right).

2.3.4 Recognition of Identity and Expression Using Features from an Untrained Neural Network

We next aimed to investigate an untrained network's face representations across image transformations. Like before, we used images from the KDEF dataset showing a frontal view as well as 45 degree views (left and right) of the faces. We explored whether the randomly initialized, untrained network could label facial expressions and face identities when the newly attached readout layer was trained with two of the three views, and then tested with the held-out view.

For expression labeling, features extracted from the initial convolutional layer, and the first, second, and third dense blocks of the untrained network yielded accuracy values of 16.54%, 16.22%, 15.51%, and 16.51%, respectively (Figure 2.5A, top right). The untrained network performed similarly for all layers of the network, with each layer performing close to chance level.

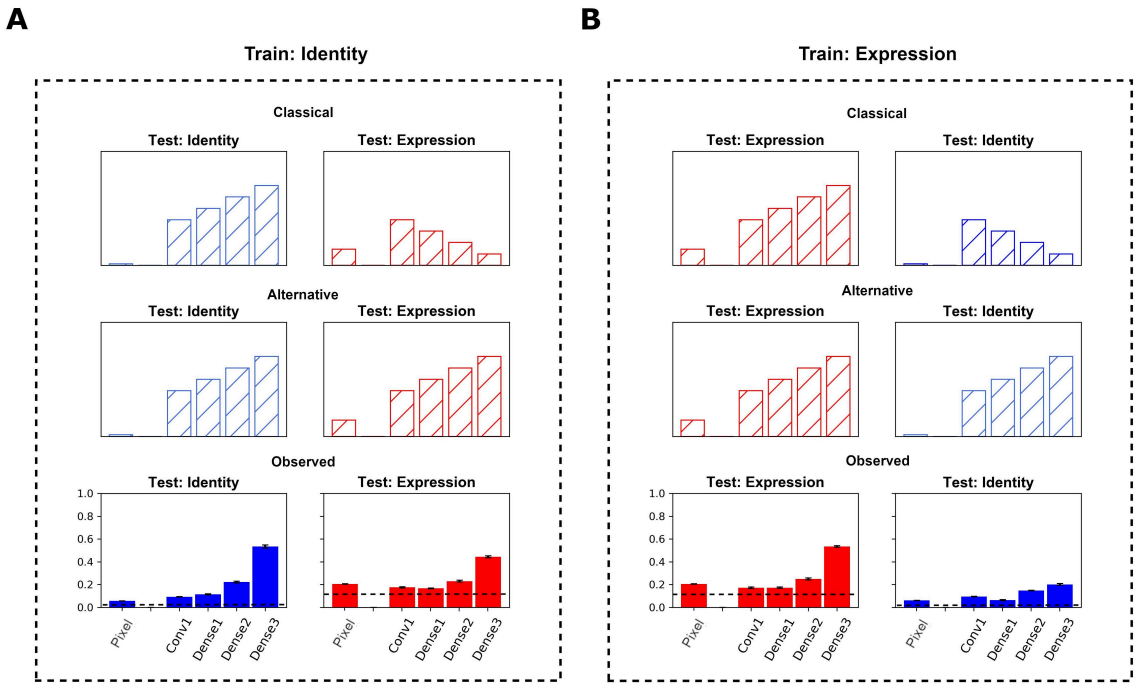

FIGURE 2.4: Identity and Expression Networks. (**A**) Identity Network. (Top row) Expected pattern of results following a classical view of abstraction. (Middle row) Expected pattern of results following an alternative view of abstraction. (Bottom row) Observed Results. Classification accuracy for identity (left) and expression (right) for a readout layer attached to successive sections of the pre-trained identity network. Dotted line represents performance at chance. Leftmost bar represents performance of the unattached linear classifier. (**B**) Expression Network. (Top row) Expected pattern of results following a classical view of abstraction. (Middle row) Expected pattern of results following an alternative view of abstraction. (Bottom row) Observed Results. Classification accuracy for expression (left) and identity (right) for a readout layer attached to successive sections of the pre-trained expression network. Dotted line represents performance at chance. Leftmost bar in each plot represents performance of the unattached linear classifier. Error bars denote the SEM of the performance of each network instance.

For identity labeling, features extracted from the initial convolutional layer, and the first, second, and third dense blocks of the untrained network yielded accuracy values of 7.90%, 7.13%, 13.62%, and 6.10%, respectively (Figure 2.5A, bottom right). The untrained network decreased in classification performance overall.

Figure 2.5B shows the accuracy differences for expression and identity labeling when subtracting the untrained network performance from the trained network performance of the transferred task. Overall, the difference between the transferred task performance and the untrained performance increased from layer to layer, showing the relative advantage of the trained network.

2.3.5 Recognition of Identity and Expression Using Features from a Neural Network Trained to Recognize Scenes

To test the transfer performance of a network trained to recognize an unrelated category, we explored the ability of a network trained for scene recognition to label facial expression and face identity across image transformations. Unlike facial expression and face identity recognition tasks, which both involve face images as inputs, scene recognition does not involve faces. We examined whether a scene network (that received no face input during training) could label facial expression and face identity after the newly attached readout layer was trained using two of the three views, and was then tested with the held-out view from the KDEF dataset.

When labeling expression, features extracted from the initial convolutional layer and first, second, and third dense blocks of the scene network yielded accuracy values of 15.9%, 16.0%, 23.5%, and 33.0%, respectively (Figure 2.6A, top right). Although the scene network increased from layer to layer, it did not perform as well

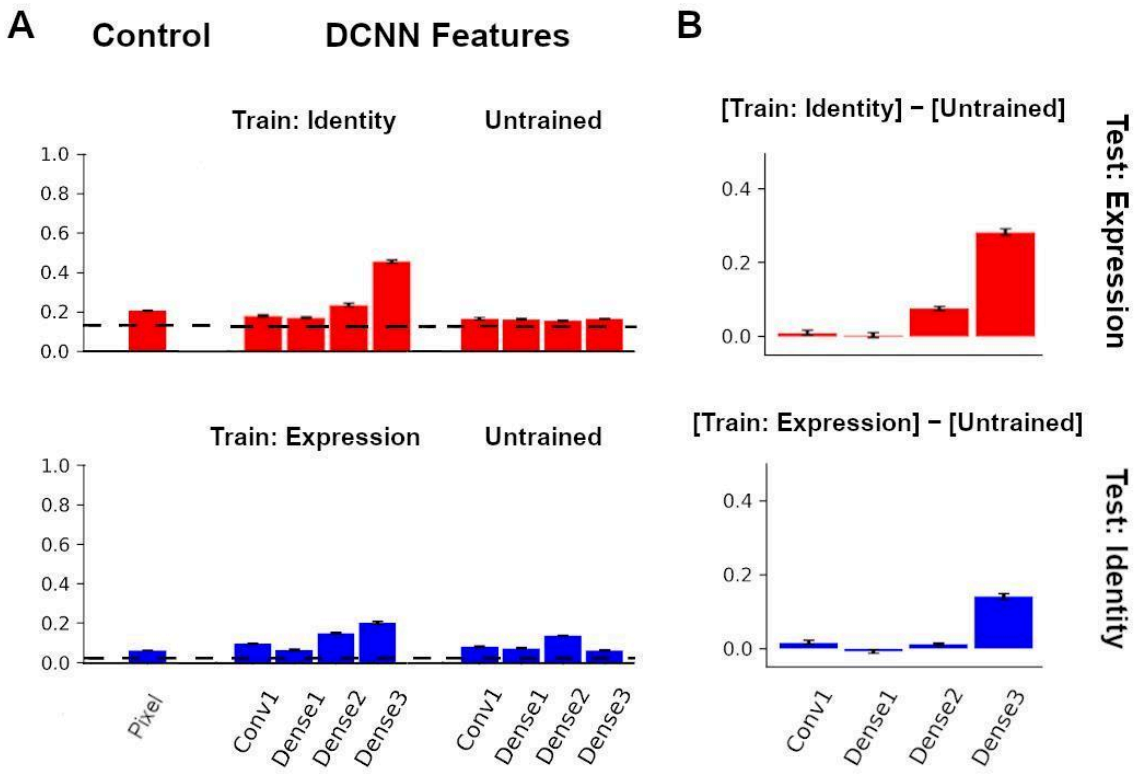

FIGURE 2.5: Comparisons with the Untrained Network. (**A**) Classification performance using identity features and untrained features for expression labeling (top) and expression features and untrained features for identity labeling (bottom). (**B**) Difference in expression classification between identity network and untrained network (top). Difference in identity classification between expression network and untrained network (bottom). Error bars in plots denote the SEM of the performance of network instances.

as the expression and identity networks for expression classification. The differences in accuracy between the identity and scene network for expression labeling can be seen in Figure 2.6B (top).

When labeling identity, features extracted from the initial convolutional layer, and the first, second, and third dense blocks of the scene network yielded accuracy values of 9.5%, 7.8%, 17.3%, and 29.6%, respectively (Figure 2.6A, bottom right). Although the scene network increased from layer to layer, it did not perform as well as the identity network. However, interestingly, the scene network was more accurate at identity labeling than the expression network. This can be seen in Figure 2.6B (bottom).

2.3.6 Overlap between Identity and Expression Features May Decline across Layers

Different hypotheses could account for the observed increase in accuracy for identity labeling in correspondence with the increase in accuracy for expression labeling. According to one hypothesis, recognition of face identity and facial expression might rely on similar features. Therefore, the features learned by the network trained to recognize expression would also yield good accuracy when labeling face identity. Instead, according to a different hypothesis, recognizing identity and expression would require disentangling two generative sources that jointly contribute to the same image. In this case, separating what aspects of the image were due to identity could prevent a neural network from erroneously attributing those aspects to expression. For this reason, a neural network trained to label identity or expression might develop representations of expression and identity, respectively. The representations could then be used to disentangle identity and expression,

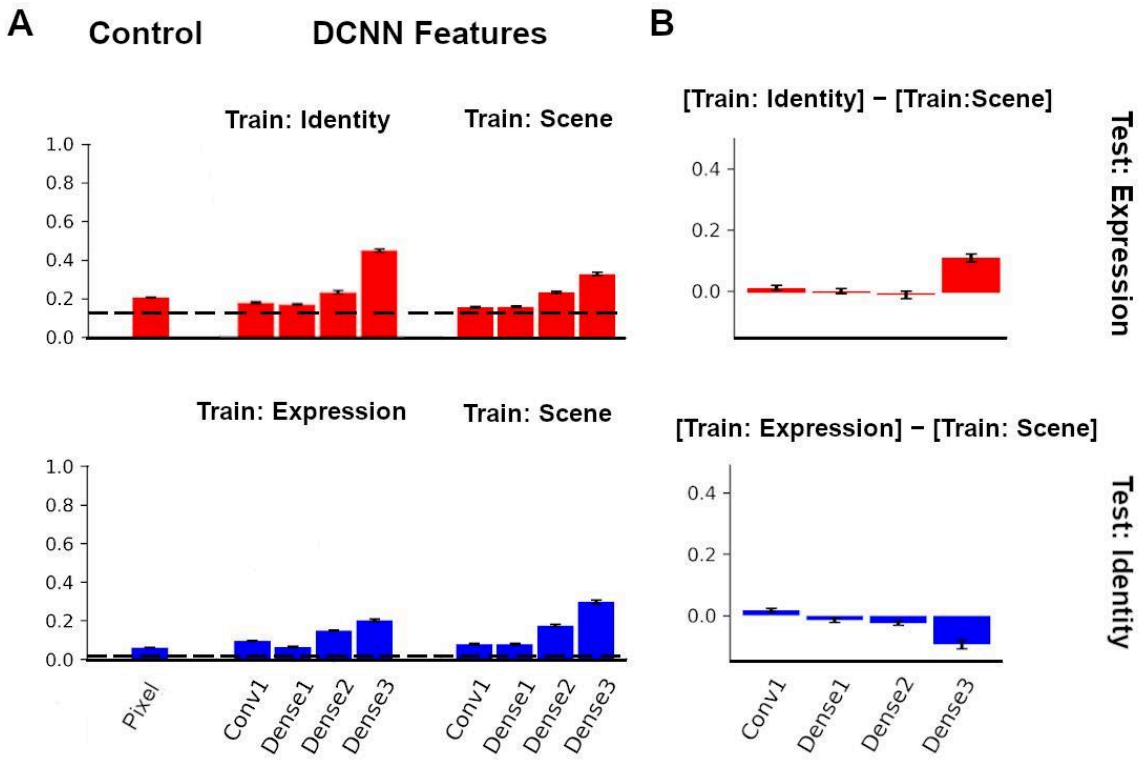

FIGURE 2.6: Comparisons with the Scene Network. (**A**) Classification performance using identity features and scene features for expression labeling (top) and expression features and scene features for identity labeling (bottom). (**B**) Difference in expression classification between identity network and scene network (top). Difference in identity classification between expression network and scene network (bottom). Error bars in plots denote the SEM of the performance of network instances.

even when recognition of identity did not rely on the same features as expression recognition.

If the features that were most useful for labeling identity and expression were similar, the dimensions that best discriminated between identities and those that best discriminated between expressions should also be similar. Thus, the angles between identity dimensions and expression dimensions should be small and congruence should be high. If, on the other hand, features needed to recognize identity and expression were disentangled by the net, the angles between identity dimensions and expression dimensions should become increasingly larger from layer to layer. Furthermore, if training with identity or with expression induced disentanglement between identity and expression features, training the network with scene images should yield comparatively higher congruence between identity and expression features compared to training with identity or expression.

We differentiated between these predictions by calculating a congruence coefficient between the first five principal components (PCs) for expression and the first five PCs for identity for each layer of each trained neural network. A larger congruence coefficient would signify that the identity and expression dimensions were more similar to one another, and a smaller congruence coefficient would indicate they were less similar. In both the network trained to label identities and the network trained to label expressions, the PCs for identity and expression exhibited higher congruence values in the earliest layer. For both the identity and expression networks, congruence decreased from layer to layer (Figure 2.7A). The scene network's congruence values followed the same decreasing pattern. However, the congruence coefficients between identity and expression were larger compared to the other networks, indicating that the identity and expression features were less

disentangled in the scene network.

For visualization purposes, the activation patterns across network features in response to different face images were projected onto the top two identity and expression PCs for each layer within a network (see Figure 2.7B–E). In each case, the relevant aspect (expression or identity) visibly clustered in deeper layers of the net, while the other aspect did not, further showing that discrimination of expression and identity relied on co-existing but different features.

2.4 Discussion

Recent studies revealed the presence of information about face identity and facial expression within common brain regions (Anzellotti and Caramazza, 2017; Dobs et al., 2018), challenging the view that recognition of face identity and facial expression are implemented by separate neural mechanisms, and supporting alternative theoretical proposals (i.e., Duchaine and Yovel, 2015; Pitcher and Ungerleider, 2020). In the present study, we proposed the Integrated Representation of Identity and Expression Hypothesis (IRIEH), according to which recognition of face identity and facial expression are 'complementary' tasks, such that representations optimized to recognize face identity also contribute to the recognition of facial expression, and vice versa. This would account for the observation that both identity and expression information coexist within common brain regions, including the face-selective pSTS (Anzellotti and Caramazza, 2017; Dobs et al., 2018). Based on IRIEH, we predicted that features from artificial deep networks trained to recognize face identity would be able to support accurate recognition of facial expression, and reciprocally so too would features from deep networks trained to recognize facial expression be able to support accurate recognition of face identity.

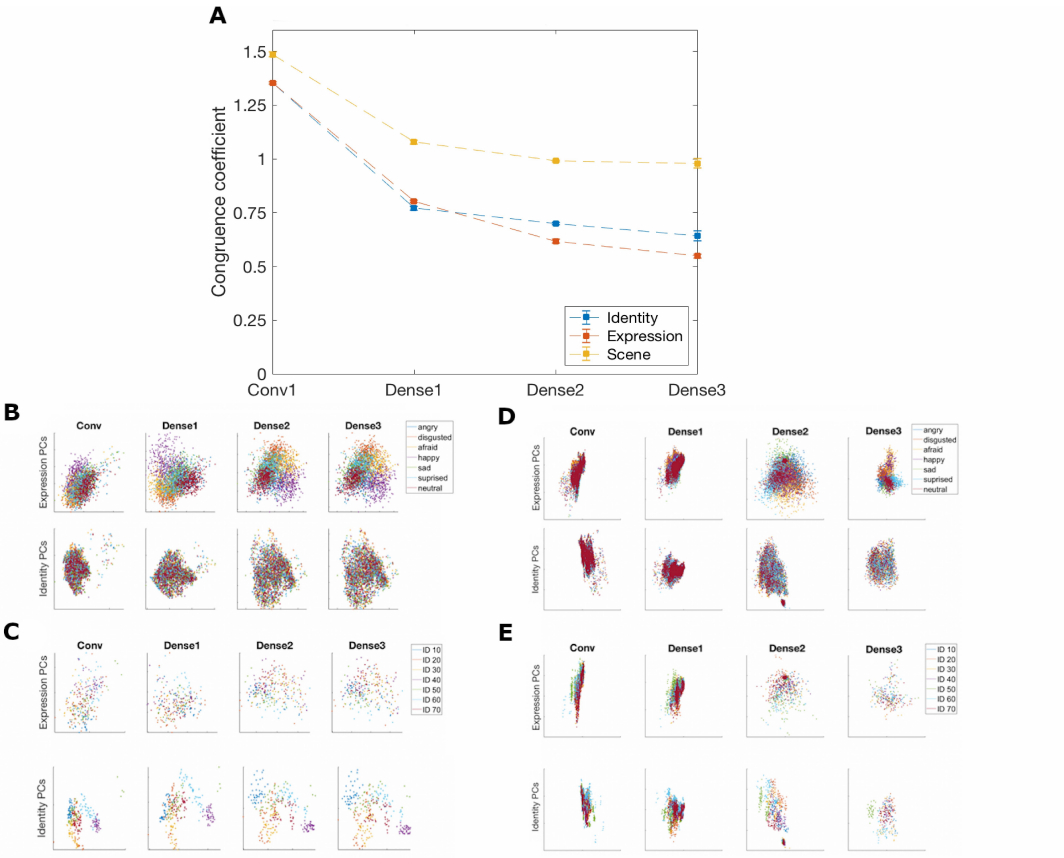

FIGURE 2.7: Trained neural networks and principal components. (**A**) Identity, expression, and scene network congruence coefficients between principal components derived from activations averaged over expression and identity. (**B**) Face activations labeled by expression projected into expression and identity principal component spaces for each layer of the identity network. (**C**) Face activations labeled by identity (only 7 of 70 identities are displayed for clarity) projected into expression and identity principal component spaces for each layer of the identity network. (**D**) Face activations labeled by expression projected into expression and identity principal component spaces for each layer of the expression network. (**E**) Face activations labeled by identity (only 7 of 70 identities are displayed for clarity) projected into expression and identity principal component spaces for each layer of the expression network.

To evaluate this hypothesis, we trained a deep convolutional neural network (DCNN) to label face identity, and found that, as the labeling of identity increased in accuracy from layer to layer, the labeling of expression also correspondingly improved, despite the fact that the features of the identity network were never explicitly trained for expression recognition. We also demonstrated that this phenomenon was symmetrical. The same DCNN architecture trained to label expression learned features that contributed to labeling identity, even though the features of the expression network were never explicitly trained for identity recognition. Additionally, in the models that we tested, features from a network trained to categorize scenes also supported identity and expression recognition, indicating that this phenomenon might not be restricted to within domain-tasks.

Our findings could serve as proof, that in order to perform identity recognition, expression information does not necessarily need to be discarded (and vice versa). In fact, within the set of models that we tested in this article, networks trained to perform one task did not just retain information that could be used to solve the other task, but rather, they enhanced it. The accuracy for labeling expression achieved with features from intermediate layers of the network was higher than the accuracy achieved with features from early layers. Likewise, the accuracy of labeling identity using features trained for expression recognition improved over layer progression. These same patterns held for the identity network, in that accuracy improved over the layers when labeling identity and expression.

In seeming contrast with our results, a previous study Yosinski et al., 2014 found that features became increasingly specialized for the trained task in the later layers of the network. In the present article, despite features encoding expression and identity becoming increasingly orthogonal from early to late layers, accuracy

at labeling progressively increased for the tasks. A fundamental difference that sets apart the study by Yosinski and colleagues (Yosinski et al., 2014) from the present study is that we attached a read-out layer directly to the frozen hidden layer, rather than continuing to train the rest of the model. When retraining multiple layers, starting from an early pre-trained layer yields better accuracy (Yosinski et al., 2014). However, our results indicated that, at least in the case of identity and expression, when using a simple readout, features from later layers yielded better accuracy than features from earlier layers.

Lastly, one could conclude that the increase in performance seen in late layers was not due to common features found between tasks. Our factor congruence analysis comparing identity and expression spaces suggested that the similarity between the dimensions that best distinguished between identities and the dimensions that best distinguished between expressions decreased from layer to layer in both the identity and expression networks (and this was true for the identity and expression dimensions from the scene network as well). Since a small amount of congruence remained, it was not possible to rule out some overlap. However, the representations of identity and expression became increasingly orthogonal from layer to layer. Our findings dovetailed with previous work that proposed that object recognition was a process of untangling object manifolds (DiCarlo and Cox, 2007; DiCarlo, Zoccolan, and Rust, 2012). Each image of an object can be thought of as a point in a high-dimensional feature space, and an object manifold is the collection of the points corresponding to all possible images of an object. Using pixels as the features, object manifolds are not linearly separable. Object recognition maps images onto new features that make the object manifolds linearly separable (DiCarlo and Cox, 2007). In the case of face perception, we can think of face identity

manifolds (the points corresponding to all possible images of a given face identity), and facial expression manifolds (the points for all images of a given expression). By interpreting the identity and expression results from this perspective, face perception is not only limited to untangling identity manifolds, but also to untangling expression manifolds. In other words, the process of untangling one set of manifolds naturally untangles the other to some extent, similar to pulling two ends of yarn to unravel a knot.

There are several aspects that need to be taken into consideration when interpreting our findings. First, while our results do provide a proof of principle that identity representations arise naturally in simple, feedforward architectures trained to achieve near-human accuracy at expression recognition and vice versa, this does not guarantee that all neural network architectures show the same effect. Nevertheless, in support of the view that recognition of identity and expressions might be more integrated than previously thought, some recent studies tested one direction of this classification (training on identity and testing on expression) for the top layers of a ResNet-101 (Colón, Castillo, and O'Toole, 2021) model and a VGG-16 (Zhou, Meng, and Zhou, 2021) model, providing some converging evidence that this phenomenon is not restricted to the one specific neural network architecture.

Secondly, although DCNNs share similarities with brain processing, findings from DCNN models cannot be directly used to reach conclusions about the human brain (Xu and Vaziri-Pashkam, 2021). Nonetheless, DCNNs are a useful tool for proof of principle tests of computational hypotheses (see Saxe, McClelland, and Ganguli, 2019 for an elegant example) and can inspire us to generate hypotheses that we can then test with neural data.

Finally, we found that while untrained DCNNs did not lead to increasing accuracy for identity and expression recognition from layer to layer, transfer from DCNNs trained for scene recognition to face tasks (identity and expression) performed similarly to transfer from DCNNs trained for one of the face tasks (e.g., identity) to the other face task (e.g., expression). Thus, our findings cannot be interpreted as supporting the possibility that face-selectivity in the brain might be the result of greater transfer accuracy for tasks within a same category (e.g., faces) than across categories. Note that each network was retrained ten times to account for random variation in weight initialization, indicating that these results were consistent across multiple choices of the networks' initial weights.

Given the scene network's transferring ability, an open question that remains is why a model that was trained to recognize scenes was able to label identity and expression with increasing performance. Substantial evidence indicates that face and scene processing are specialized tasks and do not take place within the same brain regions (Haxby, Hoffman, and Gobbini, 2000; Epstein, 2008). If the DCNN models show that shared representations for scenes and faces are possible, then why does this not occur in the brain? One can speculate that there may be other mechanisms that may constrain category-specificity (Dobs et al., 2022). For instance, one can envision this using different types of neural network modeling, such as models that leverage multi-task learning. If one were to train a multi-task neural network to perform identity and expression recognition together and a different multi-task neural network to perform identity and scene recognition simultaneously, the former may perform significantly better than the latter. Taken together, it is likely that different sets of algorithmic learning principles determine the constraints of category-specificity.

Chapter 3

Intracranial Electroencephalography and Deep Neural Networks Reveal Shared Substrates for Representations of Face identity and Expressions

The contents of this chapter have been published in the following research articles:

Intracranial Electroencephalography and Deep Neural Networks Reveal Shared Substrates for Representations of Face Identity and Expressions

Emily Schwartz[1], Arish Alreja[2,3,4,5], R. Mark Richardson[6,7], Avniel Ghuman[2,5,8], and Stefano Anzellotti[1]

[1] Department of Psychology, Boston College, Boston, MA, United States 02467

[2] Center for the Neural Basis of Cognition, Carnegie Mellon University and University of Pittsburgh, Pittsburgh, PA 15213

[3] Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213

[4] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

[5] Department of Neurological Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA 15213

[6] Department of Neurosurgery, Massachusetts General Hospital, Boston, MA 02144

[7] Harvard Medical School, Boston, MA 02115

[8] Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA 15213

Journal of Neuroscience 7 June 2023, 43 (23) 4291-4303; DOI: 10.1523/JNEUROSCI.1277-22.2023

3.1 Introduction

Humans are exposed to a multitude of faces everyday; each face provides rich information about an individual's identity and emotion. The social importance of faces makes it critical that we understand how we recognize others and their facial expressions.

According to an established hypothesis (henceforth the 'classical' view), face identity and facial expression are processed by distinct, specialized pathways (Bruce and Young, 1986; Haxby, Hoffman, and Gobbini, 2000). In this view, face-selective regions in ventral temporal cortex ('ventral stream') are specialized for identity recognition, while face-selective regions in lateral temporal cortex ('lateral stream') are specialized for expression recognition (Haxby, Hoffman, and Gobbini, 2000). Indeed, previous studies indicate that the ventral stream plays a key role in face identity recognition. Response patterns in the ventral stream can be used to decode face identity (Nestor, Plaut, and Behrmann, 2011; Ghuman et al., 2014; Anzellotti, Fairhall, and Caramazza, 2014; Axelrod and Yovel, 2015; Dobs et al., 2018; Li, Richardson, and Ghuman, 2019; Boring et al., 2021), and participants with face recognition deficits have reduced structural connectivity in ventral regions (Thomas et al., 2009). In parallel, other studies indicate that the lateral stream plays a role in expression recognition. Facial expression valence can be decoded from a region in lateral temporal cortex: the face-selective posterior superior temporal sulcus (pSTS) (Peelen, Atkinson, and Vuilleumier, 2010; Skerry and Saxe, 2014). Additionally, patients with pSTS damage experience expression recognition deficits (Fox et al., 2011), suggesting a causal role of the lateral stream in expression recognition.

While these findings support the lateral stream's involvement in expression recognition, they do not rule out that the ventral stream might also play a role. Similarly, results suggesting ventral stream involvement in identity recognition do not rule out that the lateral stream might contribute to identity recognition. Considering this, an alternative hypothesis suggests that identity and expression are not necessarily independent neural mechanisms (Duchaine and Yovel, 2015). The ventral and lateral streams, instead, might differ in whether they represent form or motion (Duchaine and Yovel, 2015; Pitcher and Ungerleider, 2021). Consistent with this alternative, facial expression can be decoded in ventral face-selective regions (Skerry and Saxe, 2014; Li, Richardson, and Ghuman, 2019), and face identity can be decoded in lateral regions (face-selective pSTS; Anzellotti and Caramazza, 2017 Hasan et al., 2016, Dobs et al., 2018). Furthermore, behavioral studies find correlations between expression and identity recognition abilities (Connolly, Young, and Lewis, 2019).

Even considering this evidence, it is still possible that ventral and lateral streams might be specialized for identity and expression recognition, respectively. Behavioral correlations between recognition abilities might result from differences in upstream regions, before face processing separates into ventral and lateral streams. Furthermore, ventral representations specialized for identity might contain a small amount of expression information that would support fMRI decoding, and vice versa. Compatible with this possibility, computational studies using deep convolutional neural networks (DCNNs) found that identity-trained networks encode some expression information (Colón, Castillo, and O'Toole, 2021), and vice versa (Schwartz et al., 2023a). In fact, one study found that in contrast to untrained

DCNNs and DCNNs trained to recognize non-face objects, DCNNs trained to recognize face identity have expression-selective units that share similarities with human expression recognition, making similar errors (Zhou et al., 2022). Together with our results, this suggests that identity and expression recognition might share common mechanisms both in the brain and in DCNNs.

While DCNNs trained to recognize identity encode some expression information (and vice versa; Colón, Castillo, and O'Toole, 2021, Schwartz et al., 2023), DCNNs trained to recognize identity and DCNNs trained to recognize expression still have distinct representations (Figure 3.1, see Methods). If the classical view is correct, representational dissimilarity matrices (RDMs) from identity-trained DCNNs should correlate with RDMs from ventral regions, and symmetrically, RDMs from expression-trained DCNNs should correlate with RDMs from lateral regions. Critically, there would need to be an interaction between DCNN type (identity- or expression-trained) and brain region. By contrast, if ventral and lateral regions contribute to both identity and expression recognition then one would anticipate that the DCNNs should correlate with both ventral and lateral regions, and that there would not necessarily be an interaction between DCNN type and brain region. Furthermore, these conclusions hold if the models either equally correlate with the regions or if one model outperforms the other for both sets of regions. We test this directly by analyzing neural responses measured with intracranial electroencephalography (iEEG) to faces varying in identity and expression. Comparing the representational geometry of neural responses in ventral and lateral regions to the representational geometry in DCNNs trained to recognize identity and expression, we examine whether RDMs extracted from these DCNNs correlate differentially with RDMs based on responses in face-selective electrodes in ventral

and lateral regions.

3.2 Methods

3.2.1 Participants

The experimental protocols were approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from all participants. Participants were a subset of patients selected a-priori from Li, Richardson, and Ghuman, 2019 and Boring et al., 2021 who performed two variations of the face individuation task. Eleven human patients (7 females; mean age 31.8 years, $SD = 9.89$) underwent surgical placement of electrocorticographic (surface and depth) electrodes for seizure onset localization. One subject was initially excluded due to noisy data (as determined with a reliability analysis described in the Temporal Localizer section below). None of the subjects showed evidence of epileptic activity on electrodes located in the ventral and lateral temporal lobes.

3.2.2 Experimental Design and Statistical Analysis

3.2.2.1 Stimuli

Subjects viewed face images from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist, Flykt, and Öhman, 1998). The KDEF dataset consists of 4900 images depicting 70 individuals (50% female) showing 7 different expressions from 5 different angles. The following expression categories were included in the experiment: happy, sad, afraid, angry, and neutral. Each combination of a face identity and a facial expression was shown in different viewpoints, including 0 degrees

(frontal view), 45 degrees left and right views, and 90 degrees (profile) left and right views.

3.2.2.2 Experimental paradigm

Prior to completing the main task, participants completed a functional localizer task (Li, Richardson, and Ghuman, 2019; Boring et al., 2021). Subjects were shown images of faces, houses, bodies, words, hammers, and phase scrambled faces. More details about the design of the functional localizer can be found in Li, Richardson, and Ghuman, 2019; Boring et al., 2021. The data from the functional localizer was used to identify electrodes that respond selectively to faces. An electrode was deemed face-selective using the criteria described in the "Electrode localization" section below.

Two different sets of participants completed two different versions of the experiment (Li, Richardson, and Ghuman, 2019; Boring et al., 2021), which we will refer to as A and B. In both experiments, each trial began with a face image presented for 1000*ms*. This was followed by a 500*ms* inter-trial interval, during which a fixation cross was presented at the center of the screen. Subjects were instructed to press a button to identify if the presented face was male or female. Subjects were asked to respond as quickly and as accurately as possible. A set of 10 practice trials was executed before the start of the experiment.

In experiment A, each subject performed one session containing 600 trials. Subjects viewed a set of stimuli that contained 8 identities, 5 expressions, and 5 viewpoint angles (left/right profile, left/right 45 degree, and frontal). Each stimulus was presented three times within a session. In experiment B, subjects performed at least two sessions, and viewed a different subset of KDEF stimuli. Subjects viewed

a set of stimuli that contained 40 identities, 5 expressions, and 3 viewpoint angles (profile, 45 degree, and frontal). Each stimulus was shown only once per session.

3.2.2.3 Data preprocessing

Data was preprocessed at the University of Pittsburgh. Further details can be found in Li, Richardson, and Ghuman, 2019 and Boring et al., 2021. The data analyzed here contains 14 depth electrodes and 11 surface electrodes. Depth electrodes and surface electrodes were used to record local field potentials at 1000 Hz. Reference and ground electrodes were distantly placed from the recording electrodes subdurally and having contacts oriented towards the dura. Surface area of the recording site was similar across grid and strip electrode contacts. Here, "electrode contacts" will be referred to as "electrodes" in this manuscript. There were no consistent differences in neural responses observed between the grid and depth electrodes. To extract single-trial potential signals, the raw data was band-pass filtered preserving the frequencies from 0.2 Hz to 115 Hz. This step was implemented using a fourth order Butterworth filter. After removing slow and linear drift as well as high-frequency noise, a 60 Hz line noise was also removed with 55-65 Hz as the stop-band. Single-trial potentials (stP) were time-locked to the stimulus onset for the trial with the signal sampled at 1000 Hz.

Raw data was also inspected to identify and reduce artifacts. There were no ictal events detected. The mean maximum amplitude across all trials was computed and any trials with a maximum amplitude 5 standard deviations above the mean were discarded. Trials that had a difference greater than or equal to 25 μV between back-to-back sampling instances were discarded as well. This resulted in fewer than 1% of trials removed.

3.2.2.4 Electrode localization

The location of the electrodes (Figure 3.2A) was determined using an automated method that was used to coregister grid electrodes and electrode strips (Hermes et al., 2010). Patient high-resolution post-operative CT scans were coregistered with anatomical MRI scans to section electrode contacts before patients underwent surgery and implantation of the electrodes. Pre- and post-operative imaging scans were also used to localize SEEG electrodes.

Face-selective electrodes were identified by analyzing data from a functional localizer, during which participants were shown images of faces, bodies, words, hammers, houses, and scrambled faces. An electrode was defined as face-selective if its temporal response patterns could be used to decode faces from other object categories significantly above chance (see Li, Richardson, and Ghuman, 2019 and Boring et al., 2021 for details).

3.2.2.5 Deep convolutional neural network models

Deep convolutional neural networks (DCNNs) were implemented to model the neural data. Each network was trained to perform one task only, either identity recognition or expression recognition. Therefore, identity-trained models will be referred to as identity DCNNs and the expression-trained as expression DCNNs. For both the identity and expression DCNNs, we used a densely connected architecture (DenseNet, Huang et al., 2017; see Figure 3.1A), as well as a residual neural network (ResNet-18) architecture.

The identity DCNNs were trained to label identities using the CelebA dataset (Liu et al., 2018). CelebA consists of over 300,000 images. To match the size of the dataset used for the two networks, a subset of CelebA was used. The subset

contained 28,709 images for training and an additional 3,589 images labeled for testing, containing a total of 1,503 identities. These identities were randomly chosen, ensuring that at least 20 images were available for each identity. All images were sized 48×48 pixels and grayscale.

The expression DCNNs were trained to label expressions using face images from the facial expression recognition 2013 (FER2013) dataset (Goodfellow et al., 2013). The dataset is split to contain 28,709 images specified for training and 3,589 images labeled as 'public test' for validation. All images were sized 48×48 pixels and grayscale.

Once trained, the DCNNs were tested on their ability to perform identity and expression recognition using the KDEF dataset (Lundqvist, Flykt, and Öhman, 1998). This was done by freezing the DCNNs' weights, and extracting the activations of units in the last convolutional layer of each network for each of the images. Activations for the different images were then used as the inputs to a simple readout layer. To test for identity labeling, the readout layer was trained on all KDEF images except from one expression category (85.7% train, 14.3% test). The left-out expression category was then used to test the network's ability to label identity. Cross-validation was performed so that each expression category could be left-out for training (7 testing sets) and performances were averaged. To test for expression labeling, images from 7 identities were held out of the training set for the readout layer (90% train, 10% test). Cross-validation was performed so that each set of 7 identities could be left-out for training (10 testing sets) and performances were averaged.

The DenseNet trained to recognize identity achieved an accuracy of 26.5% on a left-out subset of CelebA, and the DenseNet trained to recognize expression

achieved an accuracy of 63.5% on a left-out subset of FER2013 (Schwartz et al., 2023a) Using the DenseNet, each network was able to transfer to KDEF for the task it was trained to perform (identity DenseNet on identity recognition: accuracy = 95.2%, chance level = 1.42%; expression DenseNet on expression recognition: accuracy = 81.9% , chance level = 14.2%). The identity DenseNet was able to label facial expression on the KDEF dataset with an accuracy of 77.7%. The expression DenseNet was able to label face identity on the KDEF dataset with an accuracy of 89.7%.

To facilitate comparison with previous studies, we additionally trained an identity and an expression DCNN based on the ResNet architecture (ResNet-18, He et al.). The ResNet-18 networks were trained using the same datasets that were used for the DenseNets. The ResNet-18 trained to recognize identity achieved an accuracy of 28.0% on a left-out subset of CelebA, and 91.5% on KDEF (chance level = 1.42%). The ResNet-18 trained to recognize expressions achieved an accuracy of 61.3% on a left-out subset of FER2013, and 66.4% on KDEF (chance level = 14.2%). When transferring to the different tasks, the identity ResNet-18 labeled facial expression and the expression ResNet-18 labeled identity with accuracies of 55.7% and 80.1% on KDEF, respectively. Therefore, both DCNNs performed better than chance on left-out images from the datasets that they were trained on, as well as on images from the KDEF dataset. However, they did not transfer to KDEF as well as the DenseNets.

A ResNet-18 pre-trained on ImageNet to perform object recognition (henceforth referred to as the object ResNet-18) was implemented as an additional model

comparison. Details on the training can be found in He et al., 2016. The object ResNet-18 was trained using images in RGB. Since the identity and expression DCNNs were trained using grayscale images, we modified the weights of the conv1 layer here by summing over the dimension of the input channels. The object ResNet-18 was able to label identity and expression on KDEF images with accuracies of 96.2% (chance level = 1.42%) and 61.4% (chance level = 14.2%), respectively. A randomly initialized DenseNet and Resnet-18 (same architectures as trained DCNNs) were also used as additional control analyses.

3.2.2.6 Training and testing datasets comparisons

Since we could not access a sufficiently large dataset including both identity and expression labels, the identity DCNNs and expression DCNNs were trained using two different datasets. It is possible that the testing dataset (KDEF) might be more similar to one of the two training datasets (either CelebA or FER2013). If this is the case, the networks for which the training and testing datasets are more similar might perform better. In order to test this possibility, both training datasets were compared to the testing dataset by evaluating the similarity between image representations using features from the object ResNet-18 (see "Deep neural network models"). The object-trained Resnet-18 was used to extract feature representations from different layers for images in the identity and expression training datasets, and for images in the testing dataset. For each layer, Pearson *r* correlation coefficients were computed between the features of image pairs where one image is taken from the testing dataset (KDEF) and one from either the identity or expression training dataset (this analysis was performed separately for each training dataset). Correlations were computed for one channel at a time and averaged

across channels. This was done for 100 different randomly chosen image pairs, and the correlations were averaged across the pairs. This procedure yielded for each layer a measure of the similarity between the training and testing datasets based on the features in that particular layer. In order to estimate the robustness of the results, the analysis was conducted 10 times, each time selecting a different randomly chosen set of image pairs.

When comparing images from CelebA and KDEF, this analysis yielded mean values of 0.1254, 0.3606, 0.1894, 0.2430, and 0.1244 for conv1, and hidden layers 1-4 respectively. When comparing images from FER2013 and KDEF, this analysis yielded mean values of 0.1708, 0.4263, 0.2141, 0.2739, and 0.0848 for conv1, and hidden layers 1-4 respectively.

The similarity between the training datasets and the testing dataset is comparable. In addition, neither of the training datasets is more similar to the testing dataset for all layers of the object ResNet-18. If anything, the FER2013 dataset shows greater similarity to KDEF for most layers. Therefore, if the CelebA-trained networks were to better account for neural responses, it would be unlikely that this is due to CelebA being more similar to the testing dataset (KDEF).

3.2.2.7 Representational similarity analysis: comparison between DCNNs

Before comparing the representations in DCNNs to neural responses, we sought to quantify how different are the representations learned by the identity DCNNs and by the expression DCNNs. Transfer-learning tests conducted in a previous study demonstrate that these DCNNs learn representations that can be used to perform the other task with above-chance accuracy (Schwartz et al., 2023a). For example, representations in layers of the expression DenseNet could be used to read out the

identity of faces (Schwartz et al., 2023a). However, this does not imply that the identity and expression DenseNets have the same representations.

To test the similarity of the representations in the two DCNNs, we used representational similarity analysis (RSA). We analyzed the representations in multiple hidden layers of the neural networks. Specifically, features from either four or five hidden layers were extracted: the first convolutional layer, and the last layer in each of the three dense blocks (after shrinkage) or the last layer in each of the four residual blocks. For each of these layers, we calculated representational dissimilarity matrices (RDMs) using a three-step procedure. First, we extracted feature vectors for all KDEF images used in the experiment. Next, we mean-centered the feature vectors by calculating and subtracting the mean feature vector across all KDEF images. Finally, for all pairs of images, we calculated the correlation distance between their mean-centered feature vectors (correlation distance is $1 - r$ where r is Pearson's correlation). In experiment B, information about viewpoint only included the viewpoint angle, without distinguishing between left and right viewpoints, therefore the feature vectors for the left and right viewpoints were averaged (i.e., left and right profile views averaged, left and right half views averaged).

This procedure produced RDMs of size 200×200 for experiment A, and RDMs of size 600×600 for experiment B (see Figure 3.1 B and C). Note that, as described in the *Experimental paradigm* section, the sizes of the RDMs are different in the two experiments because different subsets of the KDEF images were used in experiment A and experiment B. In the end, Kendall τ_B was used to compute the similarity between the RDMs from different layers in the two different DCNNs. A 4×4 cross-network similarity matrix for the trained DenseNets is shown in Figure 3.1D.

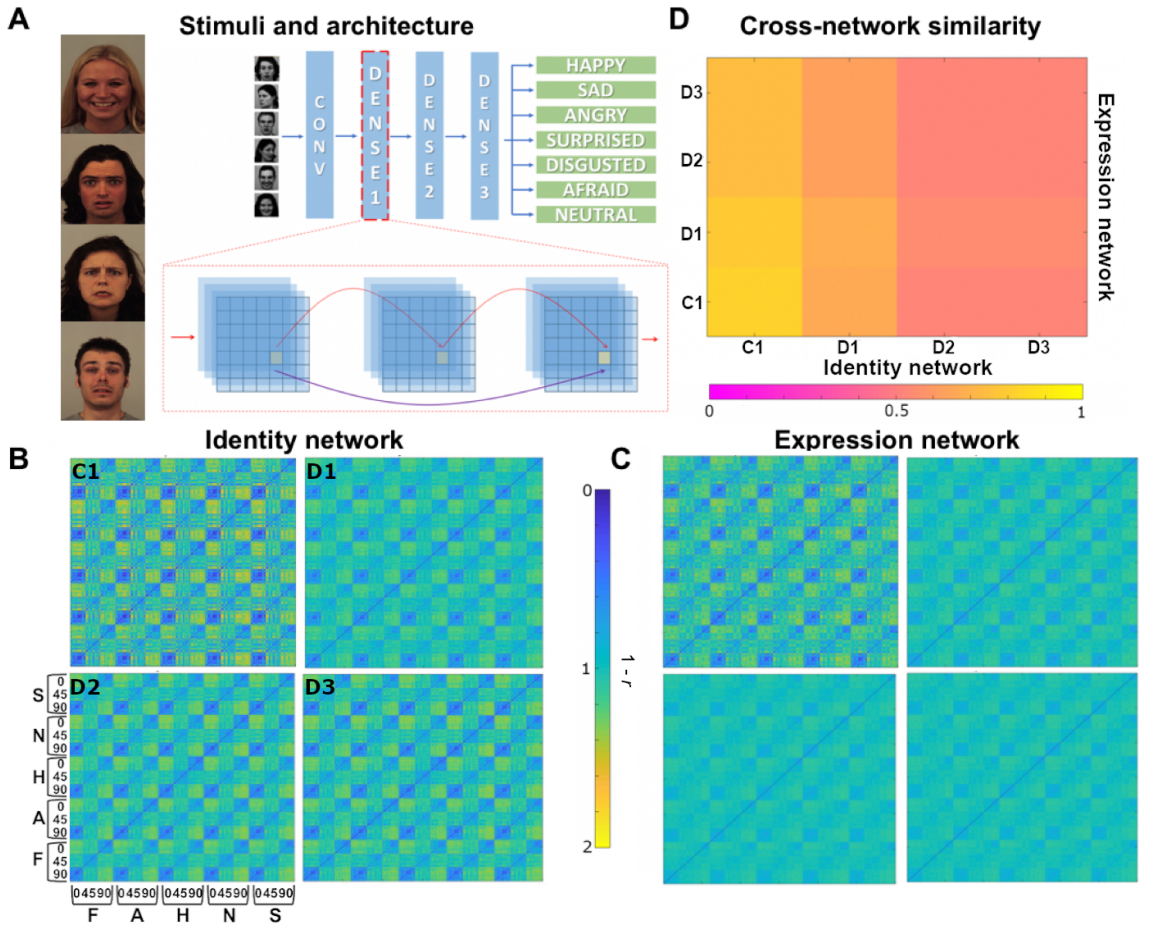

FIGURE 3.1: Face representations in a DenseNet trained to recognize identity or expression. A: KDEF stimuli (AF27HAS, AM01AFS, AF06ANS, AM29AFS) and neural network architecture examples. B: RDMs of the identity DenseNet features from KDEF images used in version A of the experiment. C: RDMs of the expression DenseNet features from KDEF images used in version A of the experiment. D: Kendall tau values between identity DenseNet RDMs and expression DenseNet RDMs. Each tick on the horizontal axis represents an identity DenseNet RDM and each tick on the vertical axis represents an expression DenseNet RDM. C1, conv 1; D1, dense block 1; D2, dense block 2; D3, dense block 3.

3.2.2.8 Representational similarity analysis of neural data

To retain as much data as possible, we initially performed an analysis on all of the face-selective electrodes, including those from participants who were shown each stimulus once. In this analysis, we computed separate RDMs for each of three temporal windows (125*ms* - 175*ms*, 175*ms* - 225*ms*, 225*ms* - 275*ms*). This specific temporal range was chosen based on previous studies on the temporal dynamics of visual face perception (Barbeau et al., 2008). As discussed in more detail later, it remains possible that some face information might be encoded in later time windows as well (Ghuman et al., 2014, Li, Richardson, and Ghuman, 2019, Boring et al., 2021; see also Temporal Localizer section below). For each temporal window per each electrode, we extracted a 50-dimensional vector, such that the value for each dimension reflects the amount of measured response in the corresponding millisecond of the 50*ms* window. RDMs were obtained by following the same procedure used for the DCNN RDMs, using correlation distance to determine the dissimilarity between the response patterns for each pair of stimuli. As in the RSA analysis for the DCNNs, the average response over all stimuli was subtracted from each stimulus response to remove any baseline that is stimulus-independent.

In addition to this, we performed an RSA analysis restricted to highly reliable responses from electrodes located in the fusiform gyrus (n=7). Highly reliable electrodes and time windows were identified following the procedure described below. We then extracted patterns of response from each reliable electrode and time window and performed the RSA analysis following the same approach described in the previous paragraph, comparing neural RDMs to RDMs extracted from the DenseNet models.

3.2.2.9 Temporal localizer

We sought to identify time windows during which face-selective electrodes show the most reliable responses. The data time-series was segmented using disjoint, successive time windows of 50ms. The first window was centered at 0 post stimulus onset, and the last at 500. Therefore, the windows included time points starting from 25ms before stimulus onset to 525ms post-onset. Disjoint windows were used to reduce the number of multiple comparisons. To identify which of the time windows contained relatively less noise compared to the amount of base signal, all presentations of a stimulus' neural response were correlated within a specific time window. An average correlation over all time windows for that stimulus was obtained as well. The average correlations across all time windows were then subtracted from the time window-specific correlations. A paired t-test was performed between the correlation of the stimulus responses for a given time window and the average correlation of the mean response averaged over all time windows for that stimulus ($p < 0.05$). This determined which of the time windows contained response patterns whose test-retest reliability was significantly higher than average. To correct for multiple comparisons, all t-tests were Bonferroni corrected. One electrode was excluded from the RDM analysis due to not containing any time windows with reliable responses ($p > 0.05$).

3.2.2.10 Representational similarity analysis: comparison between neural activity and DCNNs

We next aimed to compare the neural representations in specific time windows to the representations in the DCNNs. In particular, we evaluated the extent to which

RDMs computed using the identity DCNNs and using the expression DCNNs correlate with RDMs based on the iEEG measurements (Kriegeskorte and Kievit, 2013; Khaligh-Razavi and Kriegeskorte, 2014). To calculate the concordance between the DCNN's RDMs and the neural RDMs, we performed two types of analysis. In the first type of analysis, we compared the RDMs extracted from neural data to RDMs extracted from individual layers of the identity DCNNs and of the expression DCNNs, calculating Kendall Tau's rank correlation coefficient, τ_B. Since negative variance explained values are uninterpretable, any negative τ_B correlations were set to 0 (Fang, Poskanzer, and Anzellotti, 2022). The smallest of negative values was -0.003. This affected a total of 37 out of 360 tau values. This procedure was repeated using RDMs from the object ResNet-18 and untrained DCNNs as well. However, since this analysis compares neural representations to the representations in one DCNN layer at a time, one limitation of this analysis is that it does not capture the overall correspondence between neural data and the representations across all layers of a DCNN jointly.

Comparing neural representations to DCNN representations one layer at a time does not reveal to which extent different layers of the DCNN encode redundant information or unique information. To address this question, we introduced a new type of analysis, using semi-partial Kendall Tau's rank correlation (Kim, 2015) to evaluate the overall correspondence between the RDMs extracted from the neural data and each of the identity and expression DCNNs when considering jointly the representations in all layers of the DCNNs.

Semi-partial correlations measure the strength of the relationship between two variables (i.e., between the neural RDM and the first hidden block RDM) while controlling for the effects of other variables (i.e., the initial convolutional RDM).

Within each DCNN model, the semi-partial τ_B was calculated for each layer, controlling for the effect of the previous layers. Then, the semi-partial τ_B values were summed to obtain a cumulative τ_B value. This allows one to control for redundancy between the layers, evaluating the overall similarity between the models and the data without inflating the τ_B values.

After calculating the semi-partial τ_B values between the face-selective electrodes and identity and expression DCNNs, we performed model comparison using Bayes Factor to potentially establish evidence for the absence of differences between the DCNN's ability to account for neural responses (Keysers, Gazzola, and Wagenmakers, 2020). This was done to evaluate the statistical evidence for the possibility that there is no difference between the identity DCNN's representational similarity to the neural representations and the expression DCNN's representational similarity to the neural representations (and more precisely, that they come from a same distribution). The analysis with Bayes Factor was performed using the set of all face-selective electrodes to maximize statistical power.

3.2.2.11 Relative contribution of identity and expression

Next, we set out to test if different sets of electrodes were more strongly correlated with one DCNN over the other. The dataset included electrodes located in the ventral stream as well as electrodes located in lateral temporal regions. If ventral regions are specialized for identity recognition, and lateral regions are specialized for expression recognition, ventral electrodes might have a greater cumulative τ_B with the identity DCNN, while lateral electrodes might have a greater cumulative τ_B with the expression DCNN. Alternatively, electrodes in ventral and lateral regions might be similar in terms of their relative correspondence to the identity

DCNN and to the expression DCNN.

To compare the relative similarity of neural RDMs in individual electrodes to the RDMs of the identity and expression DCNNs, each electrode at each time window was plotted as a point in a 2D space, where the coordinate along the x-axis was determined by the cumulative Kendall τ_B between the electrode's RDM and the identity DCNN RDM, and the coordinate along the y-axis was determined by the cumulative Kendall τ_B between the electrode's RDM and the expression DCNN RDM. If ventral electrodes have comparatively higher Kendall τ_Bs with the identity DCNN, and lateral electrodes have comparatively higher Kendall τ_Bs with the expression DCNN, the two sets of electrodes should fall on lines with different slopes, where the slopes correspond to the ratio between the cumulative τ_B for the identity model and the cumulative τ_B for the expression model. In particular, electrodes in the ventral stream that are comparatively better explained by the identity DCNN should fall on a line which is closer to the identity axis, while electrodes in the lateral stream should fall on a line which is closer to the expression axis (despite electrodes varying in how well they are explained overall). This would demonstrate the presence of an interaction between DCNN model type and brain region (in line with the classical view). By contrast, if all the electrodes fall on the same line, it means that the relative performance of the identity and expression DCNN models at explaining neural responses is similar for the two streams (in contrast with the classical view), demonstrating the absence of an interaction between DCNN model type and brain region.

Frequentist tests are designed to test for the presence of significant interactions, but a lack of significant effects does not demonstrate no interaction. This makes it challenging to test for the absence of an interaction. However, Bayesian tests are

built in such a way that they can evaluate the strength of evidence for the absence of an effect. Thus, a Bayesian approach is implemented to evaluate the relative support for a model in which all the electrodes fall on the same line compared to a model in which the electrodes can fall on two separate lines, one for each stream.

To statistically test if ventral and lateral electrodes fall on lines with different slopes, we fit the data with two competing linear regression models: one model with two separate slopes for the ventral and lateral electrodes, and one model with a single slope. We then performed model selection with the Bayesian Information Criterion (BIC) to determine which linear regression model provides a better account for the data. A lower BIC score signifies the better model. The difference between BIC scores, $\delta = BIC_{separate} - BIC_{combined}$, determines the size of the effect: a difference greater than 10 denotes strong evidence for the better model (Raftery, 1995).

To further examine the ratio between identity and expression model performance for the DenseNet models, we calculated an index ranging from $-\infty$ to ∞, where negative values indicate that the neural representations correlate more with representations in the expression DCNN, and positive values indicate that they correlate more with the identity DCNN. To accomplish this, for each electrode and time window, we calculated the index $LR = log(\tau_{id}/\tau_{exp})$. This was then plotted as a histogram where log ratios between $-\infty$ and 0 represent expression-preferring electrode/time-window combinations and log ratios between 0 and $+\infty$ represent identity-preferring electrode/time-window combinations. When conducting comparisons with the DenseNets, three electrodes had cumulative τ_B values smaller or equal to zero in one time window, therefore the log-ratio could not be calculated and they were not included in the log-ratio histogram (Figure 3.3B).

3.3 Results

3.3.1 Representations in deep networks trained for identity and expression recognition

We compared the representations in two deep convolutional neural networks (DCNNs) with the same DenseNet architecture (Figure 3.1A) where one network was trained to recognize identity and the other was trained to recognize expression. For each network, we calculated representational dissimilarity matrices (RDMs) using activations in the first convolutional layer, and in the last layer of each dense block (Figure 3.1B and 3.1C). To compare the feature representations across the two DCNNs, the similarity between the RDMs was computed using Kendall's τ_B (Figure 3.1D). Early layers were more similar to one another compared to late layers. The τ_B values between the DCNNs steadily decreased from layer to layer, indicating that the representations in the two DCNNs become increasingly different in later layers. A similar pattern was found when comparing the identity and expression ResNet-18 representations.

3.3.2 Localization of face-selective electrodes

After probing the representations of faces in the DCNNs, we localized face-selective electrodes to analyze the neural representations of the same set of face stimuli. Out of the 1,079 total electrodes across 11 participants, 25 were found to be face-selective (2.3%). Of these 25 electrodes, 12 were located in the ventral stream (defined as the ventral portion of the temporal cortex and of the occipital cortex anterior to area V2) with 10 of them being located in the fusiform (as determined with

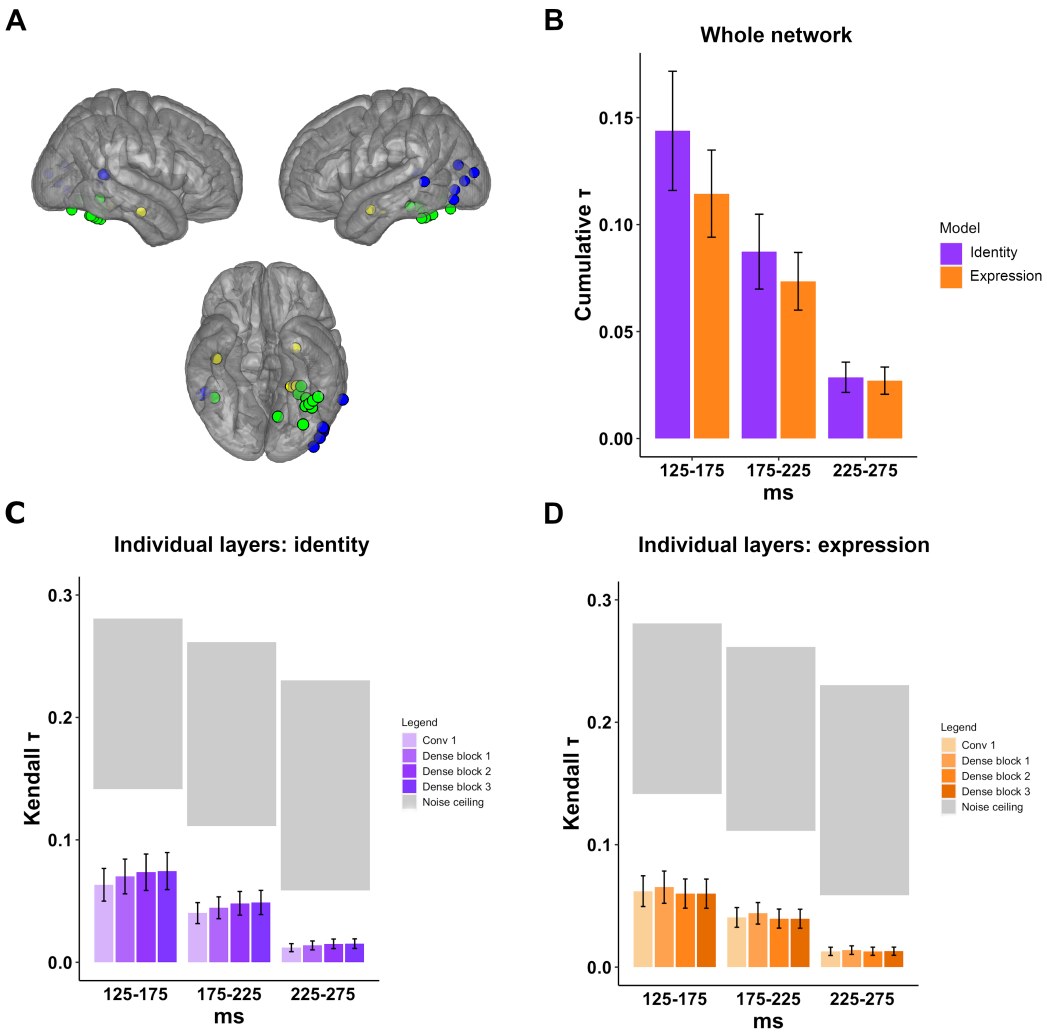

FIGURE 3.2: Face-selective electrodes and Kendall τ_B correlations between their representational similarity and the representational similarity in DenseNet layers. A: Face-selective electrode locations (n=24). B: Semi-partial τ_B values were computed to examine contribution across layers. This is plotted as a cumulative value obtained from each model and averaged over electrodes. SEM bars are depicted. C: Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the identity DenseNet averaged over electrodes (n=24). SEM bars are depicted. D: Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the expression DenseNet averaged over electrodes (n=24). SEM bars are depicted.

Neurosynth, Yarkoni et al. 2011, we additionally used Marsbar, Brett et al. 2002, to confirm that these electrodes were located within Brodmann area 37). However, one of these face-selective electrodes did not surpass our reliability analysis and was removed from further analyses. The remaining 24 electrodes are shown in Figure 3.2A. Eight of the face-selective electrodes were located in the lateral stream (defined as the lateral temporal cortex and lateral occipital cortex anterior to area V2 - including V3d, V5, and the superior temporal sulcus), and four of the electrodes were located in regions outside the ventral and lateral streams, and thus were labeled as "other".

Comparison between face-selective neural responses and deep networks

Having identified the face-selective electrodes, we next sought to compare representations in these electrodes to representations in the trained DenseNet models. To this end, for each electrode and time window, we computed neural RDMs, and we compared them to the RDMs extracted from the DenseNets using Kendall's τ_B. This analysis revealed that representational similarity between the model RDMs and the neural RDMs decreased from the 125-175ms time window to the 225-275 ms time window (Figure 3.2C, D), for both the identity and expression models (this might be due to a decline in the reliability of the signal, see Discussion). However, within each time window, Kendall τ_B values were comparable for both DenseNets (Figure 3.2C, D).

To probe more rigorously the representational similarity between neural responses and the identity and expression DCNNs overall, we used a novel approach, that consists in calculating a cumulative Kendall τ_B value between neural responses and multiple layers of a DCNN combined (see Methods for details). While the cumulative Kendall τ_B between the identity DenseNet and neural responses was numerically higher than the expression DenseNet (Figure 3.2B), the difference showed weak evidence for one model over the other (Bayes Factor: 0.412-0.441).

To evaluate the robustness of our results, we then repeated our analysis using ResNet-18 for our model. Following the same approach as the DenseNet analysis, RDMs were extracted from the ResNet-18 and compared to each neural RDM. Similarly to the DenseNet results, representational similarity between the ResNet-18 RDMs and the neural RDMs decreased from the 125-175*ms* time window to the 225-275 *ms* time window (Figure 3.4A, B), for both the identity and expression models. For almost all time windows, the identity ResNet-18 outperformed the expression ResNet-18 (Figure 3.4A, B). Bayes Factor was performed on the cumulative Kendall τ_B values. This again found weak evidence for one model over the other (Bayes Factor: 0.444-0.503).

Previous work (Storrs et al. 2020) found similar amounts of correspondence between trained and untrained neural network models and neural RDMs (unless tuning was used). Consistent with this, untrained DCNNs show similar correspondence with neural responses in this study. While identity and expression DCNNs yielded different representations (Figure 3.1 B-D), these differences did not capture corresponding differences between the neural responses in ventral and lateral regions. The untrained DenseNet layer correlations to the neural data had

values ranging between 0.0567-0.0623, 0.0355-0.0416, and 0.0105-0.0124 for time windows 125-175*ms*, 175-225*ms*, and 225-275*ms*, respectively. Neural responses showed a lower correspondence but similar pattern with the untrained ResNet-18 (Table 3.1). The ResNet-18 model that was pre-trained on object recognition (Table 3.1) also performed comparably to the identity DCNNs. Overall however, the identity ResNet-18 outperformed the pre-trained object network.

TABLE 3.1: ResNet-18 and neural responses

	Conv 1	Hidden layer 1	Hidden layer 2	Hidden layer 3	Hidden layer 4
125-175 ms					
Identity ResNet-18	0.0532	0.0616	0.0702	0.0861	0.1067
Expression ResNet-18	0.0530	0.0552	0.0687	0.0694	0.0456
Object ResNet-18	0.0523	0.0695	0.0772	0.0728	0.1014
Untrained ResNet-18	0.0540	0.0525	0.0478	0.0422	0.0372
175-225 ms					
Identity ResNet-18	0.0333	0.0399	0.0438	0.0534	0.0688
Expression ResNet-18	0.0334	0.0357	0.0435	0.0435	0.0279
Object ResNet-18	0.0329	0.0433	0.0468	0.0449	0.0664
Untrained ResNet-18	0.0336	0.0331	0.0306	0.0272	0.0233
225-275 ms					
Identity ResNet-18	0.0101	0.0127	0.0141	0.0153	0.0161
Expression ResNet-18	0.0103	0.0121	0.0136	0.0139	0.0092
Object ResNet-18	0.0100	0.0134	0.0145	0.0140	0.0185
Untrained ResNet-18	0.0101	0.0099	0.0093	0.0083	0.0071

3.3.3 Examining relative similarity in individual electrodes

The pattern of results observed across all face-selective electrodes might arise from averaging electrodes with distinct properties: ventral temporal electrodes in regions specialized for identity recognition, with greater representational similarity to the identity DCNN, and lateral temporal electrodes in regions specialized for expression recognition, with greater representational similarity to the expression

DCNN. Alternatively, the representations measured by electrodes in both ventral and lateral temporal regions might be similar in terms of the extent to which they correlate with activations in the identity and expression DCNNs, respectively.

We investigated this question with two converging analyses. First, for each electrode we evaluated how similar the electrode is to the identity DenseNet RDM and then how similar it is to the expression DenseNet RDM. We used these two values as coordinates for a scatter plot (Figure 3.3A). If ventral electrodes have comparatively higher cumulative Kendall τ_B with the identity DenseNet, and lateral electrodes have comparatively higher cumulative Kendall τ_B with the expression DenseNet, the two sets of electrodes should fall on two different lines with different slopes. Instead, all observed electrodes were located along one line - showing a similar ratio of expression τ_B to identity τ_B (Figure 3.3A). To quantify this, we used the Bayesian Information Criterion (BIC, lower values indicate a better model) to compare a model with separate slopes for the ventral and lateral electrodes separately ($BIC_{separate} = -198.13$) to a model with a single slope for both the ventral and lateral electrodes ($BIC_{combined} = -603.84$, $BIC_{separate} - BIC_{combined} = 405.71$). Differences greater than 10 in BIC values are interpreted as providing strong evidence in favor of the model with lower BIC (i.e., the combined model; Raftery 1995). All electrodes (surface and depth) fall on the same line (Figure 3.3A), suggesting that they have a similar ratio of match to the identity and expression DenseNet models.

The same procedure was repeated using the ResNet-18 model. In accordance with the DenseNet results, the face-selective electrodes were located along one line - showing a similar ratio of expression τ_B to identity τ_B (Figure 3.4C).The Bayesian Information Criterion analysis confirmed this: a model with separate slopes for the

ventral and lateral electrodes separately had a smaller BIC ($BIC_{separate} = -191.91$) compared to a model with a single slope for both the ventral and lateral electrodes ($BIC_{combined} = -478.17$, $BIC_{separate} - BIC_{combined} = 286.26$). Again, this suggests there is strong evidence in favor of a model where ventral and lateral face-selective electrodes are modeled with a single slope.

Next, we computed an index capturing the relative contribution of RDMs from the expression DenseNet and RDMs from the identity DenseNet to account for neural RDMs. The index ranges from $-\infty$ to ∞: negative values indicate a greater contribution of the expression DCNN while positive values indicate a greater contribution of the identity DCNN (see Methods for details). The distribution of index values is shown in Figure 3.3B.

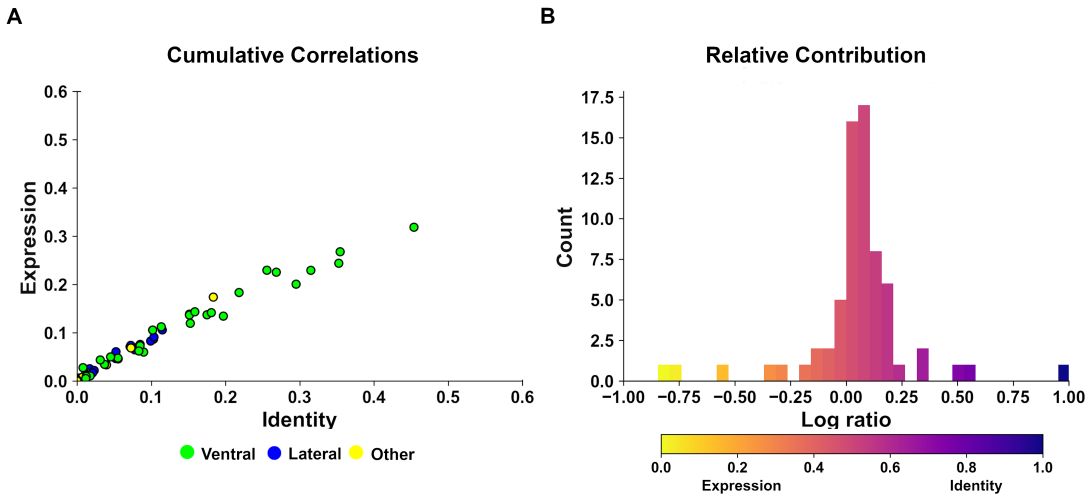

FIGURE 3.3: Variation across individual electrodes. A: Scatter plot comparing τ_B values from identity and expression DenseNet models matched on electrode (n=24) and time window. Each electrode's neural response was segmented into 3 time periods, generating 72 data points. B: Histogram showing relative contribution of identity and expression DenseNet models (69 datapoints, 3 electrodes had one time window dropped). Expression-preferring electrodes have a log ratio from $-\infty$ to 0 while identity-preferring electrodes have a log ratio from 0 to $+\infty$.

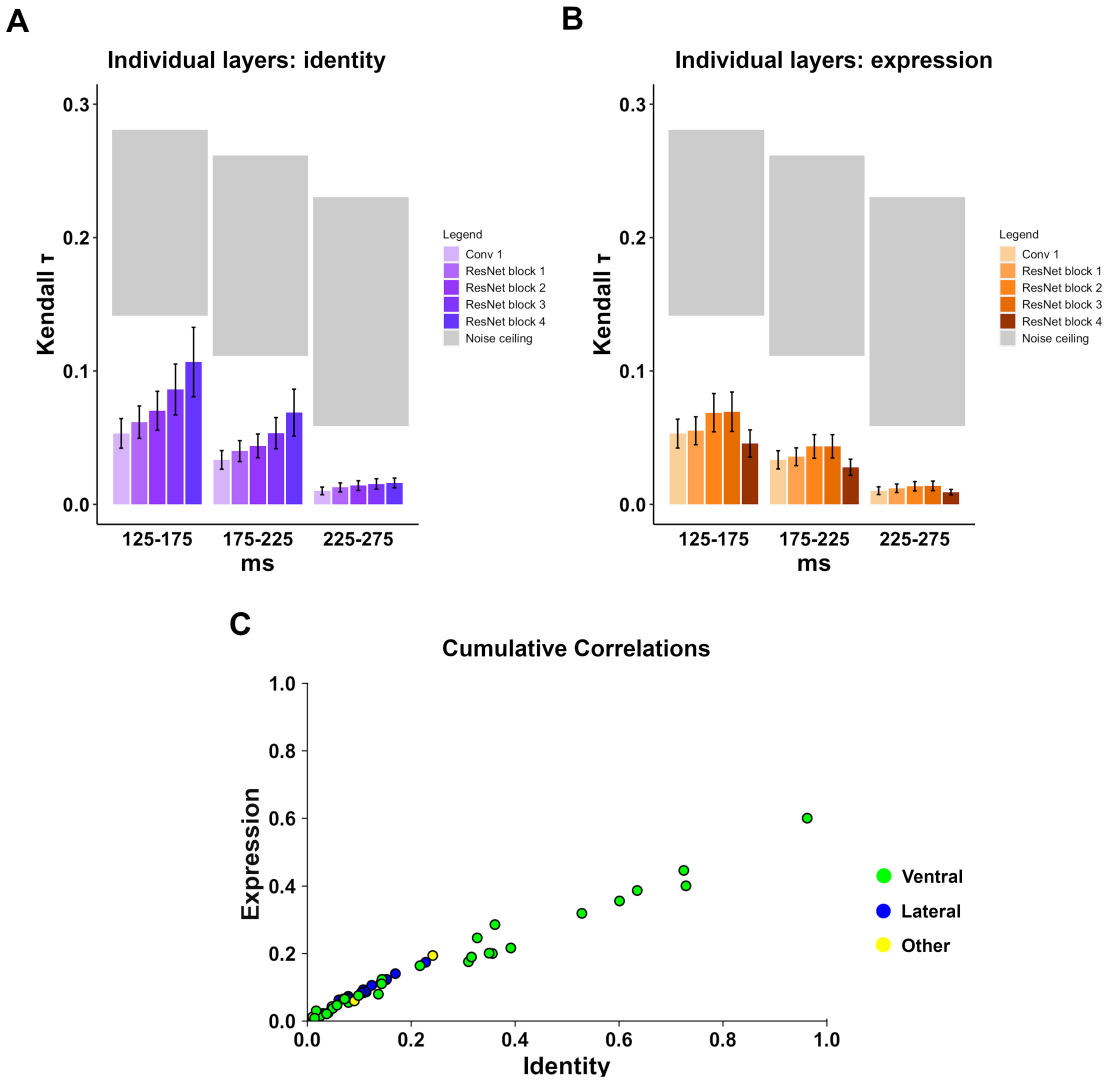

FIGURE 3.4: Face-selective electrodes and Kendall τ_B correlations between their representational similarity and the representational similarity in ResNet-18 layers. A: Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the identity ResNet-18 averaged over electrodes (n=24). SEM bars are depicted. B: Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the expression ResNet-18 averaged over electrodes (n=24). SEM bars are depicted. C: Scatter plot comparing τ_B values from identity and expression ResNet-18 models matched on electrodes (n=24) and time window. Each electrode's neural response was segmented into 3 time periods, generating 72 data points.

3.3.4 Comparison between fusiform neural responses and deep networks

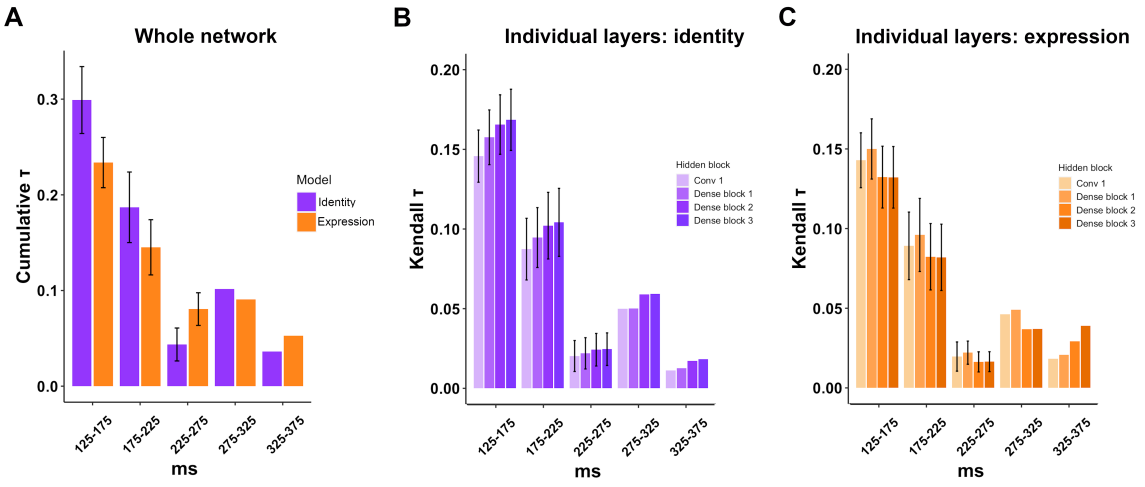

FIGURE 3.5: Representational similarity Kendall τ_B correlations between fusiform electrode responses and DenseNet deep networks' layers. A: Semi-partial τ_B values were computed to examine contribution across layers for fusiform electrodes (n=7) in time windows showing high reliability (see Methods: Temporal localizer). This is plotted as a cumulative value obtained from each model and averaged over electrodes. SEM bars are depicted for time windows with more than one electrode. B: Kendall τ_B values between fusiform iEEG RDMs and layer feature RDMs from the identity DenseNet averaged over electrodes (n=7). SEM bars are depicted for time windows with more than one electrode. C: Kendall τ_B values between fusiform iEEG RDMs and layer feature RDMs from the expression DenseNet averaged over electrodes (n=7). SEM bars are depicted for time windows with more than one electrode.

As a final step, we performed an additional analysis restricted to highly reliable responses in face-selective electrodes located in the fusiform, a region known to play a key role in face perception (Kanwisher, McDermott, and Chun, 1997). Several electrodes in this region (n=7) had highly reliable responses across multiple time windows. The τ_B values for fusiform-located electrodes comparisons

were averaged across electrodes for each time window. Figure 3.5A shows results examining the contribution across layers of each DenseNet model. The τ_B values are higher as compared to the average of all face selective electrodes shown in Figure 3.2 (panels B, C, and D), but follow a similar pattern. Within most time windows, the identity DenseNet model displayed a numerically larger cumulative semi-partial τ_B compared to the expression DenseNet model when examining fusiform electrodes (as was the case in the analysis with all face-selective electrodes as well). This difference between the DenseNet models was greatest in the 125*ms*-175*ms* range.

Figure 3.5B shows fusiform responses correlated with individual layers of the identity DenseNet and Figure 3.5C shows fusiform responses correlated with individual layers of the expression DenseNet. Similar to the face-selective pattern mentioned above, both identity and expression DenseNet models were best able to explain neural responses in the 125*ms*-175*ms* range, followed by 175*ms*-225*ms*, and then 225*ms*-275*ms* when averaging data over multiple electrodes. The 275*ms*-325*ms* and 325*ms*-375*ms* windows are included for single electrodes that showed reliable responses during one of the two periods. Similar to the face-selective electrodes again, within each time window, later layers of the identity DenseNet model outperformed earlier layers. This was not the case for the expression DenseNet model.

3.4 Discussion

According to a classical view in the field, face identity and facial expression are processed by separate mechanisms (Bruce and Young, 1986): identity is processed by regions in ventral temporal cortex, while expression is processed by regions

in lateral temporal cortex (Haxby, Hoffman, and Gobbini, 2000). If this is the case, features optimized to recognize facial expression should better capture the similarity between neural responses in lateral regions, and features optimized to recognize face identity should better capture responses in ventral regions. Thus, the classical view would predict that RDMs from identity-trained DCNNs should correlate more with RDMs from the ventral regions, and RDMs from expression-trained DCNNs should correlate more with RDMs from the lateral regions. However, this was not what we found: both identity and expression DCNNs were able to explain neural responses in ventral and lateral regions. The identity DCNNs outperformed the expression DCNNs in both sets of regions (even though this difference was not found to be significant).

These results cannot be dismissed as being due to noise. First, if the data were too noisy, we would have encountered poor correlations between the DCNN models and neural responses. However, Kendall τ_B values in this study were comparable to other studies (Higgins et al., 2021). It should also be noted that the Kendall τ_B for both the identity and expression DCNN models were close to zero in later time windows, indicating that values found in earlier time windows were not just due to the method used. Second, statistical analysis using Bayesian Information Criterion (BIC) revealed that the results provide strong support for the hypothesis that the relative contribution of identity and expression DCNN models is similar for ventral and lateral electrodes (Figure 3.3). Finally, when restricting our analysis to electrodes and time windows with very reliable responses, the pattern of results was unchanged (Figure 3.5).

Successful transfer learning can be difficult due to potential differences in the data distribution between source and target datasets (Madan et al., 2022). Thus, it

is important to determine whether the neural networks trained for their respective tasks can successfully generalize to the KDEF dataset. Both the identity and the expression DCNNs yielded high accuracies on the KDEF dataset. Despite that the expression DCNN labeled expressions with a lower accuracy than the identity DCNN labeled identity, its accuracy was well within the human range (from 72%, Goeleven et al., 2008 - to 89.2%, Calvo and Lundqvist, 2008) for the DenseNet. For this reason, while it is difficult to rule out domain shift problems entirely, it is unlikely that the accuracy difference between the two DCNNs is due to a failure of transfer.

Instead, the difference might be driven by task difficulty. Some facial expressions can be ambiguous even for human observers (Aviezer, Trope, and Todorov, 2012; Guo, 2012; Tarnowski et al., 2017). While expression recognition performance on KDEF ranges from 72% (Goeleven et al., 2008) to 89.2% (Calvo and Lundqvist, 2008), human observers are very accurate (above 90%) at recognizing identity (Bruce, 1982; Burton et al., 1999), even in the presence of changes in viewpoint.

The gender recognition task that the participants performed might have affected their neural responses, and in turn, their correspondence with the DCNN models. Previous work demonstrated that attention selectively enhances face representations (Dobs et al., 2018). It could be argued that the gender detection task is more similar to an identity task. However, gender provides only limited information about identity, and gender information can be decoded from neural responses earlier than identity information (Dobs et al., 2019). Despite this, we cannot entirely rule out that the gender recognition task might have differentially engaged

identity recognition mechanisms, potentially enhancing the amount of identity information in face-selective regions.

Nonetheless, our findings are still difficult to reconcile with the classical view. If ventral regions are specialized for identity and lateral regions are specialized for expressions, we would expect that a gender task would enhance ventral responses, and leave lateral responses unaffected (or suppressed). Instead, we find that lateral responses show robust correlations with the identity DCNN. A gender recognition task can only enhance identity representations in lateral regions if there can be identity representations in those regions to begin with. Therefore, the correspondence between the identity DCNN and lateral regions challenges the view that representations of identity and expressions are separate.

If ventral and lateral regions are not specialized respectively for the recognition of identity and expression, do they serve the same functional role? If not, what are their functional differences? Studies using combined transcranial magnetic stimulation (TMS) and fMRI (Pitcher, Duchaine, and Walsh, 2014) suggest that the posterior superior temporal sulcus (pSTS) might receive inputs from both regions responding to motion and regions encoding shape information. In addition, there is evidence for pSTS involvement in audiovisual integration (Nath and Beauchamp, 2012; Anzellotti and Caramazza, 2017; Rennig and Beauchamp, 2021). Considering this evidence, we speculate that lateral temporal regions along the superior temporal sulcus might host the convergence of static visual information, dynamic visual information, and auditory information.

In seeming contrast to the proposal that recognition of face identity and facial expression share common neural mechanisms, some previous studies reported patients with dissociations between these two abilities. For example, Hornak, Rolls,

and Wade (1996) reported a case of a patient with impaired recognition of expressions but spared recognition of identity. However, the patient had damage in ventral frontal cortex, not in lateral temporal cortex. As proposed in Calder (2011), processing of identity and expressions might diverge at later stages, but they might still rely on common regions in posterior temporal cortex. In a more recent study (Jansari et al., 2015), one patient (DY) with acquired prosopagnosia showed identity recognition deficits, but relatively intact expression recognition as tested with FEEST (Young et al., 2002). However, DY did have difficulty recognizing anger (Jansari et al., 2015), indicating some impairment for expression recognition. In addition, while DY was impaired relative to controls at recognizing the identity of upright faces, his performance was similar to that of controls when distinguishing inverted faces and fractured faces, suggesting that he might rely on featural information (Jansari et al., 2015). Such featural information might have also been sufficient to distinguish between the different emotions in FEEST. This possibility is consistent with the previously reported finding that anger recognition is particularly affected by face inversion (Bombari et al., 2013, Figure 2): in DY, an impairment for configural face processing might have led to the observed difficulties for recognizing the identity of upright faces and also to his disproportionate difficulty for recognizing anger.

The present findings are part of broader research efforts indicating that information about object category and other object properties coexist in common regions within temporal cortex (Hong et al., 2016). A relevant study reported that speaker identity and speech content can be decoded in the superior temporal cortex (Formisano et al., 2008; Bonte et al., 2014). Together, these studies reveal that some sets of tasks rely on shared brain regions, while others are implemented by

distinct neural substrates. Recent work is beginning to investigate what are the optimal ways of structuring and sharing representations across multiple different tasks (Zamir et al., 2018; Schwartz et al., 2022).

It is important to note that this study is affected by some limitations. First, the DCNNs were trained using two different datasets. It would be preferable to use a training dataset that included both identity and expression labels, but we were unable to find one such dataset with a sufficient number of images. To mitigate this concern, the training datasets we used (FER2013 and CelebA) are similar in that they include images with a broad range of variation in viewpoint and pose. The DCNNs trained with the two datasets both achieved high performance on the KDEF dataset. It is worth mentioning that even if we had used a single dataset with labels for both identity and expression, the same dataset could include very different expressions but similar identities (or vice versa). Therefore, ensuring that the DCNN's transfer accuracy is high is essential to determine whether the training procedure was successful for both identity and the expression tasks.

The Bayes Factor analysis only showed weak evidence for the identity DCNN's abilities to explain the neural responses compared to the expression DCNNs when evaluating Kendall τ_B values for all of the face-selective electrodes together (DenseNet in 3.2B). It is possible that there would be stronger evidence if more data could have been included in the analysis, but given the number of data points available, the evidence for this difference is only weak. However, even if the difference between the two models were strong, this would not alter the conclusion that the results challenge the classical view: Figure 3.2B includes electrodes from both the ventral and lateral streams, and the BIC scores strongly favored a single-line fit for both streams (Figures 3.3A, 3.4C).

We found that neither the identity nor the expression DCNN models accounted for a large proportion of the variance in later time windows (Figures 3.2, 3.4A, 3.4B), suggesting that the DCNN models we used do not fully capture the structure of face representations. This conclusion is in line with work showing that feedforward DCNNs do not offer a complete account of representational similarity between different images of objects (Xu and Vaziri-Pashkam, 2021). Models that incorporate recurrence are promising candidates to improve the concordance with neural representations (Kar et al., 2019; Kietzmann et al., 2019). Additional studies are needed to test whether they provide a better characterization of neural responses to faces in later time windows.

Recent findings have suggested that object-trained DCNN models can explain similar or greater variance in neural responses to face stimuli compared to face-trained DCNN models (Grossman et al., 2019; Chang et al., 2021; Ratan Murty et al., 2021). Some research groups have interpreted this to mean that face-selective cells are not entirely domain-specific (the "domain-general view"; Vinken, Konkle, and Livingstone, 2022). Alternatively, it has also been proposed that face-selective cells may have a generalist-like function (the "generalist view"; Chang et al., 2021), in the sense that these cells might support multiple face perception tasks (e.g. recognition of expressions, age, et cetera). If this is the case, DCNNs that encode features that can support several different face perception tasks would show more similarity to neural representations of face images. In turn, DCNNs trained to perform object recognition might encode such a variety of features because they are trained with many different object classes that vary widely in shape, color, and texture. This could explain why object recognition models show more similarity of neural responses to face images.

In our study, the ResNet-18 trained to perform face identity recognition and the ResNet-18 trained to perform object recognition performed similarly in terms of their correlation with the neural data. It is possible that this could be due to face-selective regions encoding domain-general features. Alternatively, if a ResNet-18 model was trained to perform multiple face tasks, rather than just a single task, it is possible this face-specific model would significantly outperform the object-trained ResNet-18. This would suggest that face-selective regions do encode domain-specific features that support multiple different face tasks. Our study is not designed to discriminate between the domain-general view and the generalist view. However, our results are at least consistent with the generalist view, suggesting that face-selective regions contribute to both identity and expression recognition. Future studies will need to be implemented to distinguish between these two alternatives.

Even though we did not observe differences between the ventral and lateral streams in terms of their correlations with identity and expression DCNNs, comparing the representations learned by these DCNNs in more detail remains an interesting question for future research. Methods that localize the regions of an image that are important for a given classification (Selvaraju et al., 2017) might offer cues about features that are key for both identity and expression recognition, and features that might be uniquely relevant for one of the two tasks.

Lastly, iEEG is a correlational method. Therefore, we are unable to demonstrate that representations recorded by lateral electrodes causally contribute to identity recognition, nor that representations recorded by ventral electrodes causally contribute to expression recognition. Studies using causal methods (such as TMS;

Pitcher et al., 2007) will be needed to establish the causal involvement of these representations for face perception. Even considering these limitations, the findings challenge the view for which lateral regions are specialized for expression recognition while ventral regions are specialized for identity, and converge with recent evidence to suggest that face identity and facial expressions share common neural substrates.

Chapter 4

Investigating the Representation of Static and Dynamic Face Features Using fMRI and Deep Learning Models for Video Recognition

4.1 Introduction

Vision plays a crucial role for our understanding of social interactions. In social situations, one must quickly identify other individuals, their potential mental states, and analyze the surrounding environment to decide how to act appropriately. Both static and dynamic features can carry important information related to someone's visual appearance (O'Toole et al., 2011; Dobs, Ma, and Reddy, 2017). Substantial research has been conducted to investigate higher-order visual regions in the human brain to understand how these areas contribute to social perception, particularly with faces (Isik et al., 2017; Pitcher and Ungerleider, 2021).

The classical view of face perception hypothesizes that the recognition of face identity is performed by the ventral pathway of the brain, and the recognition of facial expression by the lateral temporal pathway of the brain (Haxby, Hoffman, and Gobbini, 2000). Recent work has challenged the classical view of face perception. As highlighted in the previous chapters, information about both face identity and facial expressions can be found in both the ventral and lateral temporal pathways (Skerry and Saxe, 2014; Hasan et al., 2016; Anzellotti and Caramazza, 2017; Schwartz et al., 2023b). In light of this evidence it has been proposed that identity and expression rely on shared mechanisms (Duchaine and Yovel, 2015). Work from Schwartz et al. (2023b) supports this possibility, but only tested representations of static face stimuli (images). It is important for one not to assume that any results obtained using static stimuli will hold true for dynamic stimuli as well without proper investigation (Dobs, Bülthoff, and Schultz, 2018). However, there is reason to hypothesize shared face identity and facial expression mechanisms in the brain for dynamic faces. Thus, this study aims to evaluate if this phenomenon extends

to dynamic face stimuli as well.

As discussed in Chapter 1, the STS receives direct inputs from area MT/V5, which is known for its role in visual motion perception (Komatsu and Wurtz, 1989). This makes the pSTS a highly plausible candidate for processing representations of motion information. Additionally, many studies have suggested the role of the lateral temporal pathway in social processing (Pitcher and Ungerleider, 2021). The STS is thought to be a hub for social perception (Deen et al., 2015). It plays a role in perceiving faces, voices, biological motion, as well as understanding mental states and the actions of others (Grossman et al., 2000; Shultz et al., 2011; Anzellotti and Caramazza, 2017). It is also a prominent location for audiovisual integration (Beauchamp, Nath, and Pasalar, 2010; Nath and Beauchamp, 2012), an important component for understanding social interactions.

Several studies have also reported higher neural responses to dynamic neutral faces in the fs–pSTS compared to static faces (Pitcher et al., 2011). Conversely, neural responses in the FFA do not differ for neutral dynamic faces compared to static (Pitcher et al., 2011). Emotional faces have also been used to test differences between dynamic and spatial stimuli-evoked fMRI, finding no difference in responses to static and dynamic faces in the FFA as well (Furl et al., 2013).

Here, fMRI was used to investigate the neural responses involved in processing dynamic faces. In order to study the representation of static and dynamic information, we used models that separated these different kinds of information into distinct processing streams: two-stream models for video recognition. By comparing these neural patterns with those of the models — also trained on dynamic faces — it aimed to establish that the observed phenomena in both Chapters 2 and 3 extend beyond static facial representations. Furthermore, this approach may offer insights

into the role of the two pathways for processing static and dynamic information.

Two-stream neural network models for video recognition contain a spatial stream to process individual frames, and a temporal stream to process motion extracted from sets of frames. Here, hidden two-stream neural network models (Zhu et al., 2019), a type of DCNN, were trained on face videos. Leveraging these models, it was tested whether 1) identity and expression DCNN models trained on static face frames do not differ in terms of their relative contribution in ventral and lateral regions when using fMRI, 2) identity and expression DCNN models trained on dynamic faces do not differ in terms of their relative contribution in ventral and lateral regions when using fMRI, and 3) the ventral and lateral temporal regions vary in terms of whether they represent static information, dynamic information, or a combination of both. First, we would expect both ventral and lateral temporal regions to have a similar correspondence with identity and expression static DCNNs. Second, we would expect both ventral and lateral temporal regions to have a similar correspondence with identity and expression dynamic DCNNs. Third, based on the current literature, we would anticipate that the temporal stream (interchangeable with the term dynamic stream) would exhibit weaker similarity with the ventral pathway (OFA and FFA), and that the temporal stream would exhibit greater similarity with the lateral temporal pathway (pSTS) than it would with ventral regions. Additionally, the spatial stream (interchangeable with the term static stream) could still correlate with the ventral pathway, but to a lesser degree compared to the lateral pathway. The pSTS would show greater correspondence when using a combination of the spatial and temporal stream features, while accounting for a potential benefit of additional redundant information. Thus, a combination of the spatial and temporal stream should correlate more

strongly with the lateral temporal pathway than either stream alone, specifically the face-selective pSTS. This account is illustrated in Figure 4.1. Following a similar approach to the one used in Chapter 3, this was tested directly by analyzing fMRI responses using RSA using videos of faces varying in both face identity and facial expression. Comparing the representational geometry of neural responses in ventral and lateral temporal regions to the representational geometry in DCNNs optimized for either 1) identity or expression information, and 2) static or dynamic information, we examined whether RDMs extracted from these DCNN models correlate differently with RDMs based on responses in face-selective ROIs in ventral and lateral temporal regions.

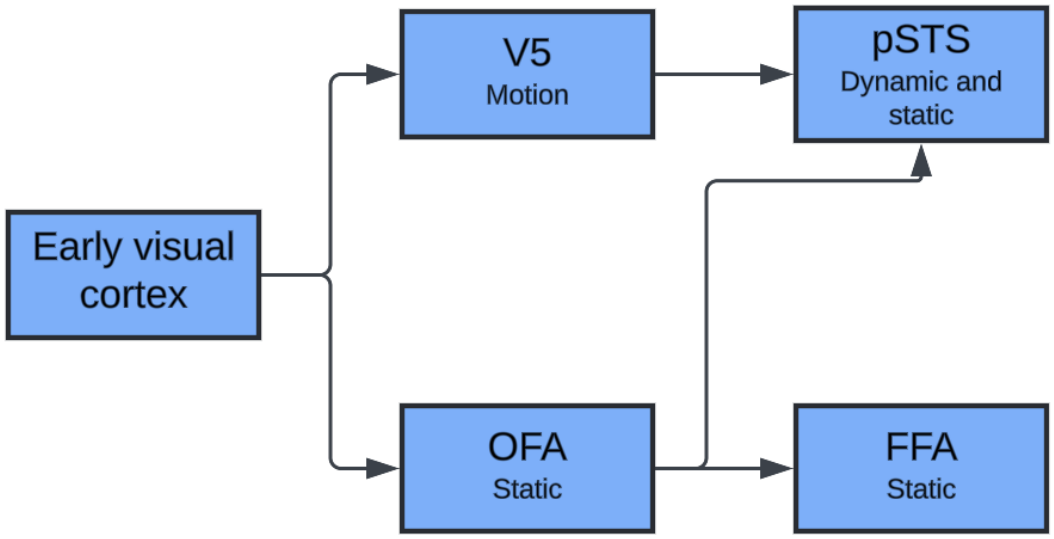

FIGURE 4.1: The alternative account of face perception processing. It can be hypothesized that the OFA and FFA are involved in processing static features of faces while the pSTS is involved in processing dynamic features of faces.

4.2 Methods

4.2.1 Participants

Flyers were used to recruit study participants. Data was collected for 20 subjects between the ages of 18-29 years(10 females; mean age 22.05 years, $SD = 2.74$). Scanning was done at Athinoula A. Martinos Imaging Center - MIT McGovern Institute. Each participant underwent an MRI safety pre-screening before taking part in the experiment. One subject was excluded due to a coil issue. This study was approved by the Boston College Institutional Review Board.

4.2.2 Experiment Design

In this study, each subject participated in two fMRI experiments. Experiment 1 used a functional localizer task and Experiment 2 used a dynamic face perception task. Both experiments had the subject use a button-press controller during the tasks.

4.2.3 Stimuli

In experiment 1, subjects were shown 240 images or videos. These images and videos fell into 6 different categories of stimuli: static images of faces, static images of bodies, videos of faces, body parts performing actions, artifacts and scenes. Face images were obtained from the KDEF database (Lundqvist, Flykt, and Öhman, 1998), while all other stimuli categories were from the Kanwisher lab (Julian et al., 2012). Each of the six categories included 36 unique stimuli. For the KDEF images, the faces were extracted and each placed on a black background.

Most previous studies of emotion perception have relied on posed facial expressions. However, posed expressions display stereotypical facial movements that can be different from those produced in naturalistic settings. To mitigate this issue, for experiment 2, we chose to use the Denver Intensity of Spontaneous Facial Action Database (DISFA, Mavadati et al., 2013) since these stimuli display spontaneous facial expressions. DISFA consists of four minute videos evoking facial expressions from 32 different identities. Subjects were recorded while watching emotion-eliciting videos. Emotions elicited include happiness, sadness, surprise, disgust, and fear. Since some of the spontaneous expressions were very subtle, we selected a subset of the expressions so that they displayed an adequate amount of motion. Videos were watched in full, and each reaction was coded for expressiveness in order to determine if the stimulus would be a good fit for the experiment. After viewing all 32 videos, seven identities were selected to be included in the experiment due to having highly-rated reactions via recognizability (by two different raters) for all expression categories. The surprise videos depicted situations that could elicit multiple emotions (e.g., an alligator suddenly snapping). These videos could have elicited both surprise and fear, and in fact the resulting facial expressions were similar. Therefore, fear and surprise were combined into one category ("fear/surprise").

After selecting the 7 identities, a research assistant cropped 2 second segments of the original videos that best captured the emotional expression. Each consisted of a person starting in a neutral expression and then moving their face into a strong emotional expression. Once the clips were selected, low-level perceptual features of the videos were adjusted to control for brain regions that are sensitive to low-level features. Stimulus transformations included RGB normalization, contrast,

and brightness, as well as various croppings of the frames. More specifically, stimuli were resized at 4 different scales by adding different amounts of padding and then resampling the frames back to 512x512.

4.2.4 Paradigms

Experiment 1 consisted of a single run. The localizer was used to identify the regions of the brain that respond preferentially to particular stimuli. During the run, participants were shown the 6 types of stimuli: static images of faces, static images of bodies, videos of faces, body parts performing actions, artifacts and scenes (Figure 4.2). For each stimulus type, 4 blocks with 20 seconds duration were shown, separated by 6 seconds of fixation, leading to a total duration of approximately 12 minutes. Participants performed a 1-back task, and in 10% of the trials, two identical stimuli were shown in a row.

Experiment 2 consisted of 4 runs. The experimental paradigm made it possible to capture the neural responses for a stimulus set consisting entirely of controlled, but spontaneous facial expressions over various identities. During each of the 4 runs, participants were shown 2 second long videos of facial expressions: 10 for each emotion label (sadness, fear/surprise, disgust, happiness, neutral). Videos were presented in randomized order. Participants were asked to press a button using a button box after a neutral facial expression video was shown. This task was chosen to limit button press movements by removing the use of multiple buttons. Seven face identities were shown, each contributing two videos for each emotion. Halfway through the run, there was a 30 second break for participants (which they are made aware of beforehand). Each video was followed by a jittered intertrial interval of 4-8 seconds extracted from a uniform distribution, leading to a total run

duration of approximately 9.7 minutes. Accuracy performance on the task was given after the completion of each run to help motivate participants to pay attention. The same stimuli in randomized order were shown in each run, but the size of each stimulus video varied between runs. The size changes were done to increase the likelihood of capturing representations of identity and expression that would be robust to changes in image transformations such as size variation, rather than capturing representations of low-level visual features. Prior to scanning, participants underwent a practice session outside of the scanner where they completed a shortened version of the controlled, dynamic face task. Figure 4.2 (right) depicts the task paradigm.

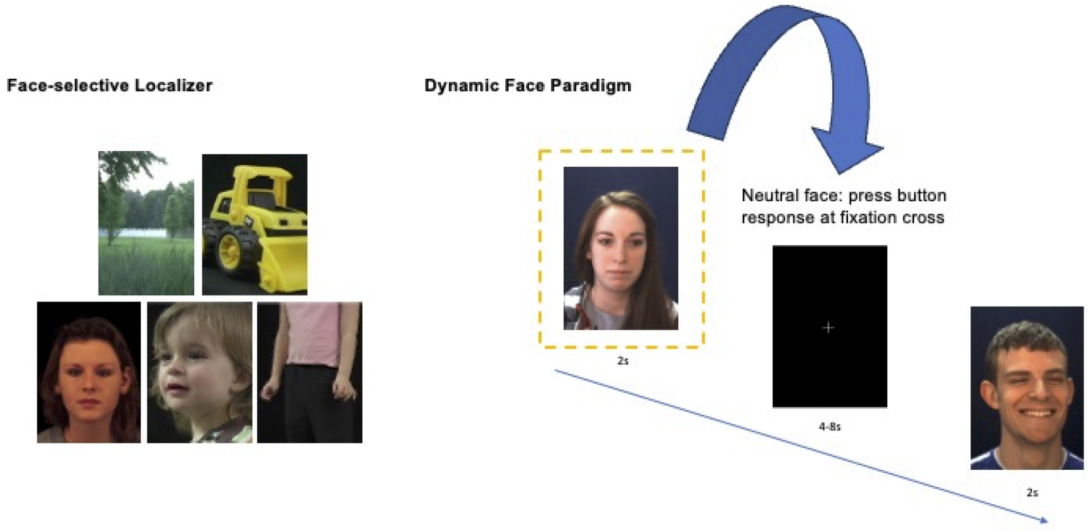

FIGURE 4.2: Face-selective Localizer. Subjects viewed images and videos of faces, body parts, and videos of objects and scenes. Subjects pressed a button when they saw a stimulus repeated twice in a row (N-1 back task). Dynamic Face Paradigm. Subject presses a button on a response controller during a fixation period after viewing a neural facial expression video clip.

4.2.5 Acquisition protocol

The acquisition protocol used the standard MIT Center for Brain Mind and Machines scanning parameters. The scanner used was a Siemens 3-T MAGNETOM Prisma with a 32-channel head coil. Before collecting functional data, a high-resolution (1 × 1 × 1 mm3) T1-weighted MPRAGE sequence was performed (sagittal slice orientation, field of view read = 256mm, field of view phase = 100%, 176 partitions with 1-mm thickness, GRAPPA acquisition with acceleration factor PE = 2, duration = 5.36 min, repetition time = 2500, echo time = 2.96, TI = 1070 ms, 8° flip angle). Functional data were collected using an echo-planar 2D imaging sequence with phase oversampling 0% (repetition time = 2000 ms, echo time = 30 ms, flip angle = 90°, slice thickness = 2.6 mm, with 2.6 × 2.6 mm in plane resolution).

4.2.6 fMRI data preprocessing

All data were preprocessed using fMRIPrep (`https://fmriprep.org/en/latest/index.html`). fMRIprep is a robust and easy to use pipeline built to preprocess fMRI data (Esteban et al., 2019). The anatomical images were skull-stripped with ANTs (`http://stnava.github.io/ANTs`). This was followed using FSL FAST for tissue segmentation. The functional images were corrected for head movements using FSL MCFLIRT (`https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT`), and then the images were coregistered to the corresponding anatomical scans using FSL FLIRT. The pipeline included brain extraction, tissue segmentation, spatial normalization for the anatomical scans, followed by bold estimation with head-motion and slice time correction, and resampling to native space. Denoising was done using the derived aCompCorr (anatomical Component-Based Noise Correction) confound regressors (Muschelli et al., 2014).

4.2.7 ROI localization

Functional localizers (Experiment 1) were analyzed using General Linear Models (GLM) in SPM12 (Ashburner et al., 2014), with boxcar predictors for different stimulus types, convolved with a standard hemodynamic response function (HRF). Human face-sensitive areas were defined as clusters of voxels which respond more to human faces than to objects, scenes, and body parts (uncorrected $p < 0.05$). The voxel cluster size was set to 0. For each participant and hemisphere, visually responsive voxels were identified based on the functional localizer scans in which responses to videos of faces were contrasted with videos of body parts performing actions, videos of artifacts and videos of scenes. ROIs were determined using a standard procedure for spherical ROI creation: identifying coordinates of peak voxels from the above steps, drawing a 9mm radius spherical ROI around the peak voxel as the sphere's center, and selecting the top 80 voxels within the spherical ROI based on the t-values from the relevant contrast. Voxels were not included in the spherical ROI if they had a t-value less than 0. The identified voxels were then used to create a binary mask to select the subset of voxels that lay within the rOFA, lOFA, rFFA, lFFA, rpSTS, and lpSTS for Experiment 2.

4.2.8 Analyzing BOLD responses to dynamic faces

In Experiment 2 initially, an analysis building a GLM for each trial where each trial was compared to all other trials was done (Mumford et al., 2012). However, after evaluating the test-retest reliability, the betas obtained appeared to be noisy. Instead, a GLM was implemented for each subject and run designed to characterize each face identity and facial expression as its own predictor using SPM12 (Ashburner et al., 2014). Trials of the same identity and trials of the same expression

were grouped together and included as multiple regressors. This yielded 11 predictors for each run (neutral removed due it its relevance to the task the subjects were told to perform). Thus, there was a resulting estimated beta image corresponding to each identity category and each facial expression category. The purpose of modeling each face identity and each facial expression instead of modeling each combination of the face identity category and the facial expression category was to avoid overfitting the model. This can happen when using too many parameters and not enough data points. Additionally, the top five aCompCorr components were included as regressors in the GLM for each subject.

4.2.9 Representational dissimilarity matrices: neural responses

Representational dissimilarity matrices were created using the face identity and facial expression beta values obtained from the GLM analysis of the fMRI data. For each ROI, the beta values from each of the 11 predictor beta images were extracted from the fMRI data using ROI masks created via the localizer study data. In other words, the full pattern of voxel data was extracted from each beta image that corresponded to a particular predictor, and this yielded an ROI-specific vector for each predictor. For each run, an RDM was creating by calculating the pairwise correlation (1-Pearson's *r*) between the ROI-specific beta vectors for each identity and expression predictor. This was repeated for all four subject RDM runs separately. Following this, for each subject, the RDM runs were then averaged to get a single subject RDM per ROI averaged across runs.

4.2.10 Reliability within subject runs

To evaluate reliability across the four runs for each subject, the smallest sized stimuli were averaged with the second largest stimuli to form one RDM pair. Similarly, the second smallest stimuli were averaged with the largest stimuli to form another RDM pair. These two averaged RDMs were then correlated using Kendall's τ_B. To verify that this was different than noise, the procedure was repeated with RDMs that were randomly shuffled before averaging, and the correlation between the averaged shuffled RDMs was obtained.

4.2.11 Noise ceilings

A preliminary analysis was performed to calculate the noise ceiling for the data in Experiment 2. Across runs, stimuli were presented multiple times with variations in size to accommodate a greater number of stimuli. Considering the repeated presentation of stimuli across runs with variations in size, RDMs for each subject were created that contained the pairwise dissimilarities between the responses for different predictors. To derive an unbiased noise ceiling, each subject's RDM was correlated with the average RDM computed from the data of the other subjects (using a "leave-one-out" approach), as well as a noise ceiling where each subject's RDM was correlated with the average RDM computed from all subject's data.

4.2.12 Two-stream deep convolutional neural network models

Two-stream DCNNs were implemented to model the neural data. The models were trained to perform either video recognition of face identity or video recognition of face expression using the Hidden Two-Stream Architecture method by Zhu

et al. (2019). The model architecture consisted of 3 components: an unsupervised optic flow estimation (MotionNet) that will be referred to as the optic flow model, a temporal stream convolutional neural network (CNN) which will be referred to as the temporal stream, and a spatial stream CNN which will be referred to as the spatial stream. The temporal stream was trained to label the stimulus based on motion (via optic flows estimated via the MotionNet component), while the spatial stream was trained to label the stimulus based on still video frames. The overarching goal of the complete model was to implicitly capture motion information and predict class labels successfully. More details describing each component are presented below.

4.2.12.1 MotionNet

Optic flow captures apparent motion information between the frames of a video. The MotionNet was used to predict the optic flow of consecutive video frames. This was done in a window size of 11 frames for each video. The optic flow maps were then fed into the temporal stream which is detailed below.

The MotionNet worked by taking windows of frames and inputting them into a DCNN made up of four convolutional layers and four deconvolutional layers. After each convolutional layer, the dimensions of the frames were downsampled. The dimensions of the frames were then upsampled through a series of deconvolutional layers and additional convolutional layers. For each set of deconvolutional/convolutional layers and their associated output resolutions, a flow loss was calculated. This flow loss was made up of 3 different loss functions. The first loss function was a standard pixelwise reconstruction loss that is based on the pixel-level optical flow change from reconstructing the current frame from the

next frame. The second loss function was a piecewise smoothness loss. This addressed the aperture problem when estimating motion in non-textured regions that may seem ambiguous by calculating the gradients of the estimated flow fields in both the x and y directions. The third loss function was a structural similarity loss (SSIM) that was used to learn the structure of the frames, by comparing how similar patches are between the frame and the reconstructed frame. The three losses were then each weighted by their relative importance during training, and then a weighted sum was calculated. A table from Zhu et al. (2019) of the MotionNet architecture can be found in Table 4.1.

4.2.12.2 Spatial stream

The spatial stream of the model processed individual frames from the videos, treating each frame as a separate image. Each frame was assigned the corresponding label from its video and is essentially subjected to the same procedure as an image for a standard image classification model. ResNet-18 (He et al., 2016) was used as the spatial stream architecture (Table 4.2), and the model was trained to either learn to identify the face identity depicted in the single frame or the facial expression.

4.2.12.3 Temporal stream

The temporal stream used the optic flow maps generated by MotionNet as its input. The MotionNet was given 11 frames as an input and from there outputted 11 optic flow estimations. These optic flows have a specific resolution as well as x and y channels. Thus, the optic flows had a shape of $11x224x224x2$. ResNet-18 (He et al., 2016) was used as the temporal stream architecture (Table 4.2), and the

Name	Kernel	Str	Ch I/O	In Res	Out Res	Input
conv1	3×3	1	33/64	224×224	224×224	Frames
conv1_1	3×3	1	64/64	224×224	224×224	conv1
conv2	3×3	2	64/128	224×224	112×112	conv1_1
conv2_1	3×3	1	128/128	112×112	112×112	conv2
conv3	3×3	2	128/256	112×112	56×56	conv2_1
conv3_1	3×3	1	256/256	56×56	56×56	conv3
conv4	3×3	2	256/512	56×56	28×28	conv3_1
conv4_1	3×3	1	512/512	28×28	28×28	conv4
conv5	3×3	2	512/512	28×28	14×14	conv4_1
conv5_1	3×3	1	512/512	14×14	14×14	conv5
conv6	3×3	2	512/1024	14×14	7×7	conv5_1
conv6_1	3×3	1	1024/1024	7×7	7×7	conv6
flow6 (loss6)	3×3	1	1024/20	7×7	7×7	conv6_1
deconv5	4×4	2	1024/512	7×7	14×14	conv6_1
xconv5	3×3	1	1044/512	14×14	14×14	deconv5+flow6+conv5_1
flow5 (loss5)	3×3	1	512/20	14×14	14×14	xconv5
deconv4	4×4	2	512/256	14×14	28×28	xconv5
xconv4	3×3	1	788/256	28×28	28×28	deconv4+flow5+conv4_1
flow4 (loss4)	3×3	1	256/20	28×28	28×28	xconv4
deconv3	4×4	2	256/128	28×28	56×56	xconv4
xconv3	3×3	1	404/128	56×56	56×56	deconv3+flow4+conv3_1
flow3 (loss3)	3×3	1	128/20	56×56	56×56	xconv3
deconv2	4×4	2	128/64	56×56	112×112	xconv3
xconv2	3×3	1	212/64	112×112	112×112	deconv2+flow3+conv2_1
flow2 (loss2)	3×3	1	64/20	112×112	112×112	xconv2

TABLE 4.1: Layers of the MotionNet Architecture from Zhu et al. (2019)

TABLE 4.2: Layers of the ResNet-18 Model.

Layer Name	Kernel Size	Input Channels	Output Channels
Conv1	7×7	3	64
Conv2_1	3×3	64	64
Conv2_2	3×3	64	64
Conv3_1	3×3	64	128
Conv3_2	3×3	128	128
Conv4_1	3×3	128	256
Conv4_2	3×3	256	256
Conv5_1	3×3	256	512
Conv5_2	3×3	512	512
Avg Pool	1×1	512	512
FC	–	512	328 (or 7)

model was trained to either learn to identify the face identity or the facial expression based on the optic flow maps generated from the 11 frames given as input into the MotionNet.

4.2.12.4 Training the MotionNet

The Human-Centric Atomic Action Dataset with Curated Videos (HAA500) database (Chung et al., 2021) was used to train the MotionNet. HAA500 is an action recognition database and consists of over 591,000 frames with 500 action labels. The database was split into training (0.8), validation (0.05), and testing sets (0.15). Frames were cropped and resized to 224 $\times$ 224. The batch size was 64 frames and the model was trained to minimize the weighted sum of the pixelwise reconstruction loss, the piecewise smoothness loss, and the SSIM loss at each resolution scale.

4.2.12.5 Training the spatial stream

For the identity recognition DCNN, a subset of 16,379 videos were used from VoxCeleb2 (Nagrani et al., 2020), a video dataset made up of over 150,000 videos.

Videos to include were chosen at random. However, each identity had to have at least 50 videos. The subset was broken down into a training set made up of 15,068 videos and each video was clipped to have 51 frames. A validation set was made up of 983 videos, and a testing set made up of 328 videos. Given 224 $\times$ 224 videos from a subset of the VoxCeleb2 dataset, the identity network was trained to recognize 328 face identities varying in pose and age. The batch size was 64 images. The network was trained to minimize the cross-entropy loss between the outputs and true labels using the adaptive movement estimation (Adam) optimization algorithm. Adam adapts learning rate to improve speed and model convergence (Kingma and Ba, 2014). The learning rate was specified at 0.0001 with initial decay rate betas ranging from 0.9 to 0.999. The training was set for 40 epochs, however, early stopping was implemented to prevent the models from overfitting. A validation set was used to test the accuracy of the model every 500 batches. If the validation accuracy was not greater than the previous max accuracy after 10 consecutive validations, early stopping of the model was implemented. The model converged and stopped running during the second epoch on batch 7000, with a validation accuracy of 80.3%. The model was then again evaluated on a small testing set and performed with an accuracy of 78.5%. Training and validation loss can be seen in Figure 4.3A.

For the expression recognition DCNN, videos were used from Dynamic Facial Expression in-the-Wild (DFEW, Jiang et al., 2020). DFEW contains more than 16,000 videos. However, due to the nature of the videos of the dataset, face detection was performed on each video to crop closer to the face. This reduced the usable videos, and thus, a subset of 6,076 videos were used for expression training. On average, each video of DFEW was made up of more frames (mean = 72.11

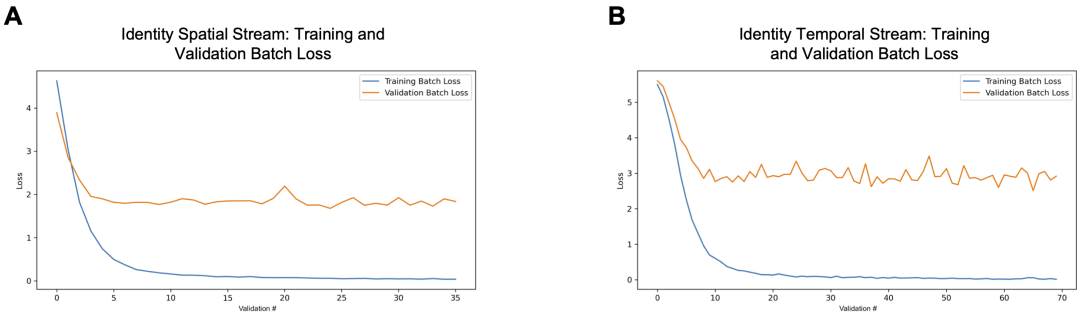

FIGURE 4.3: Training performance for identity two-stream DCNN. A) Training and validation loss of the spatial stream when training with VoxCeleb2 for identity recognition. The x-axis represents the number of times validation performance was calculated while training. Losses were plotted every 500 batches of training. B) Training and validation loss of the temporal stream when training with VoxCeleb2 for identity recognition. Losses were plotted every 500 batches of training. Losses were plotted every 500 batches of training.

frames) compared to the identity network (all videos clipped to 51 frames). The batch size was 64 images. The network was trained to minimize the cross-entropy loss between the outputs and true labels using the adaptive movement estimation (Adam) optimization algorithm (Kingma and Ba, 2014). The learning rate was specified at 0.0001 with initial decay rate betas ranging from 0.9 to 0.999. The training was set for 40 epochs, however, early stopping was implemented to prevent the models from overfitting. A validation set was used to test the accuracy of the model every 500 batches. If the validation accuracy was not greater than the previous max accuracy after 10 consecutive validations, early stopping of the model was implemented. The model ran for all 40 epochs, with a validation accuracy of 91.0%. The model was then again evaluated on a small testing set and performed with an accuracy of 87.7%. Training and validation loss can be seen in Figure 4.4.

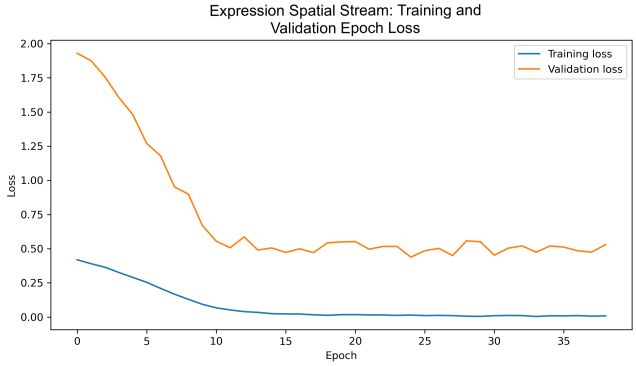

FIGURE 4.4: Training performance for the expression spatial stream DCNN when training with DFEW for expression recognition. The x-axis represents the number epochs where the training and validation losses were calculated. Losses were obtained every 500 batches of training. A plot for the expression temporal stream was not included due to incomplete training.

4.2.12.6 Training the temporal stream

Given 224 $\times$ 224 videos from a subset of the VoxCeleb2 dataset, the identity DCNN was trained to recognize 328 face identities varying in pose and age. Batch size consisted of 16 outputs from MotionNet. The network was trained to minimize the cross-entropy loss between the outputs and true labels using the adaptive movement estimation (Adam) optimization algorithm. The learning rate was specified at 0.0001 with initial decay rate betas ranging from 0.9 to 0.999. The training was set for 20 epochs, however, early stopping was implemented to prevent the models from overfitting. A validation set was used to test the accuracy of the model every 500 batches. If the validation accuracy was not greater than the previous max accuracy after 10 consecutive validations, early stopping of the model was implemented. The model converged and stopped running in its forth epoch on batch 6500, with a validation accuracy of 66.5%. The model was then again evaluated on a small testing set and performed with an accuracy of 60.2%. Training and

validation loss can be seen in Figure 4.3B.

Given 224 $\times$ 224 videos from a subset of the DFEW dataset, the expression temporal DCNN was trained to recognize 7 facial expressions (neutral, happy, sad, disgust, surprise, anger, and fear). Batch size consisted of 16 outputs from MotionNet. The network was trained to minimize the cross-entropy loss between the outputs and true labels using the adaptive movement estimation (Adam) optimization algorithm. The expression temporal stream DCNN needed a smaller learning rate than its identity counterpart. The learning rate was set to 0.00001 with initial decay rate betas ranging from 0.9 to 0.999. Due to time constraints, the training was set for 10 epochs with early stopping implemented if needed. A validation set was used to test the accuracy of the model every 500 batches. If the validation accuracy was not greater than the previous max accuracy after 10 consecutive validations, early stopping of the model was implemented. However, due to time constraints, a partially trained model is presented here. The model was evaluated on a testing set and performed with an accuracy of 46.3%.

4.2.12.7 Hidden-two stream model performance

After training each piece of the hidden-two stream model for identity recognition, the outputs of the identity spatial stream DCNN and identity temporal stream DCNN were fused to get the final model accuracy. This was done by summing the output vectors from the two streams and then running an argmax function. Argmax finds the position in a vector containing the element with the largest value. The final accuracy of the complete fused model for identity recognition was 73.1%.

After training each piece of the hidden-two stream model for expression recognition, the outputs of the expression spatial stream DCNN and the partially-trained

expression temporal stream DCNN were fused to get the final model accuracy. The final accuracy of the fused model for expression recognition was 92.9%.

4.2.13 Representational dissimilarity matrices: two-stream DCNNs

Feature representations for each stimulus used in the fMRI study were extracted from the MotionNet, the spatial stream DCNN, and the temporal stream DCNN. For the MotionNet, this corresponded to the last flow block of the model, yielding outputs of 20 channels (10 for the x gradient and 10 for the y gradient) with 224x224 resolution. For both the spatial stream and temporal stream DCNNs, outputs were extracted after the last convolutional layer of the last residual block before the fully connected layer of the ResNet-18 architecture.

Since there were multiple sizes of the same fMRI stimuli, feature representations were extracted from the DCNNs for each size of the stimulus, and then averaged to obtain a single representation for each unique stimulus. To get stimulus representations for each identity and for each expression, the stimulus feature representations were averaged over the facial expressions to get an identity representation, and averaged over the face identities to get a facial expression representation. Prior to doing this, all neutral facial expression stimuli were removed. For all pairs of expressions and identities, correlation distance was calculated between the feature vectors (correlation distance is $1 - r$ where r is Pearson's correlation) to create a representational dissimilarity matrix (RDM). The pairwise correlations were implemented in the same order for all of the following RDMs.

TABLE 4.3: RDM comparisons for evaluation.

Model	**ROI**		
Spatial stream	rOFA	rFFA	rpSTS
	lOFA	lFFA	lpSTS
Temporal stream	rOFA	rFFA	rpSTS
	lOFA	lFFA	lpSTS
Spatial + temporal stream	rOFA	rFFA	rpSTS
	lOFA	lFFA	lpSTS
MotionNet	rOFA	rFFA	rpSTS
	lOFA	lFFA	lpSTS

4.2.13.1 Representational similarity analysis: comparison between two-stream DCNNs and neural activity

To calculate the concordance between the two-stream DCNN RDMs and the neural RDMs (detailed in section 4.2.9), correlations were calculated using Kendall Tau's rank correlation coefficient, τ_B. Table 4.3 shows the different comparisons made. Semi-partial τ_B correlations were also calculated to get a combined spatial and temporal stream model τ_B value. As mentioned in the Methods section of Chapter 3, semi-partial correlations assess the strength of the relationship between two variables (e.g., between the neural RDM and in this case one of the DCNN RDMs) while controlling for the influence of other variables (e.g., the other DCNN RDM). This was done for both the identity temporal stream DCNN model and the expression temporal stream DCNN model.

As mentioned in the Methods section of Chapter 3, Frequentist tests are not designed to test for the absence of an interaction. However, Bayesian tests can be used to evaluate the strength of evidence for the absence of an effect. Thus, a Bayesian approach was once again implemented. Here, it was used for multiple

analyses. First, to examine the relative similarity of ventral and lateral ROI RDMs to the identity and expression RDMs for both spatial stream and temporal stream DCNNs. Second, to evaluate the relative support for a model in which the lateral ROIs may have a slope aligning closer to the axis representing a combination of the spatial and temporal streams compared to the ventral ROIs.

First, we aimed to test the hypothesis that ventral responses are predominantly characterized by identity information, and lateral responses by expression information. In order to evaluate this, we sought to statistically determine whether ventral face-selective ROIs (rOFA, rFFA, lOFA, lFFA) and lateral face-selective ROIs (rpSTS, lpSTS) fall on lines with different slopes in terms of their correlations with the identity DCNN and the expression DCNN. To this end, the data was fit with two competing linear regression models: one model with two separate slopes for the ventral and lateral ROIs, and one model with a single slope. This was done using the spatial stream DCNNs, and the temporal stream DCNNs separately.

Second, we aimed to test the hypothesis that ventral regions encode predominantly static information, while lateral regions encode a combination of static and dynamic information. To statistically test if ventral face-selective ROIs (rOFA, rFFA, lOFA, lFFA) and lateral face-selective ROIs (rpSTS, lpSTS) fall on lines with different slopes in terms of their correlations with the static stream model and the static and temporal model combined, the data was fit with two competing linear regression models: one model with two separate slopes for the ventral and lateral ROIs, and one model with a single slope. This was done separately for the identity DCNN model and for the expression DCNN model.

For all of these comparisons, model selection was then performed using Bayesian Information Criterion (BIC) to determine which linear regression model provided

a better account for the data. A lower BIC score signified the better model. The difference between BIC scores, $\delta = BIC_{separate} - BIC_{combined}$, determined the size of the effect, with a difference greater than 10 denoting strong evidence for the better model (Raftery, 1995). Additionally, to evaluate if combined static and dynamic feature RDMs showed significantly greater correlations with the lateral pathway compared to RDMs using only static features from the spatial stream DCNN model, Wilcoxon signed-rank test was implemented. This will be done by taking the difference of the τ_B values (combined-single stream) between subjects. The Wilcoxon signed-rank test is a non-parametric version of the paired t-test. A non-parametric test was chosen due to using correlation values which are constrained between -1 and 1, and therefore, would not be normally distributed.

4.3 Results

4.3.1 Behavioral performance on neutral face task

Subjects completed a dynamic face viewing task where they were instructed to press a button after every video of a neutral facial expression. Subjects, on average, were able to detect the neutral facial expression with an accuracy of 84% (SD: 17.7).

4.3.2 Comparison of neural RDM runs

To assess reliability within the four runs, the runs with the smallest size stimuli were averaged with those of the second largest size stimuli to create one RDM pair. Similarly, the runs with the second smallest stimuli were averaged with those of the largest stimuli to create another RDM pair. These two averaged RDMs were then correlated using Kendall's τ_B. To verify the robustness of this method, the

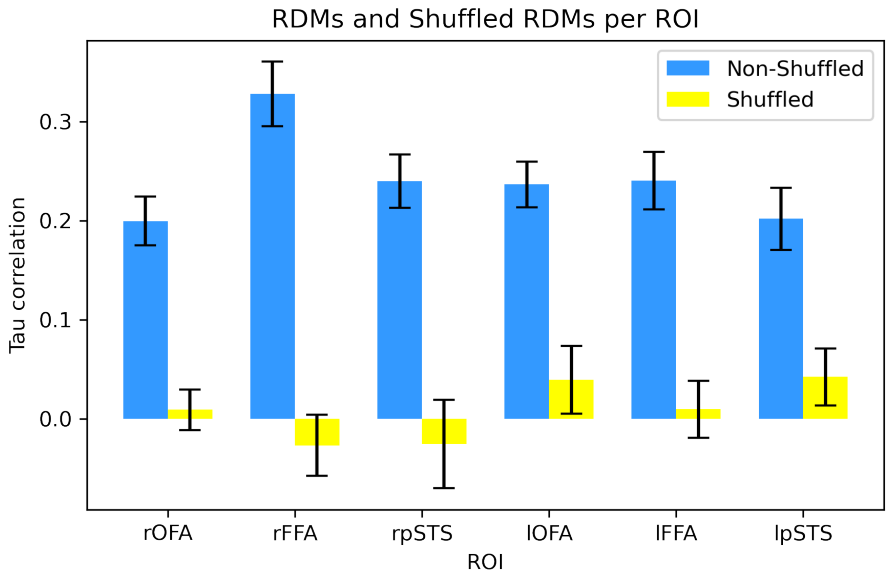

FIGURE 4.5: Within-subject correlations to evaluate reliability between pair runs of neural RDMs compared to randomly shuffled pair run RDMs.

procedure was repeated with RDMs that were randomly shuffled prior to averaging. The correlation between the averaged shuffled RDMs was obtained. Figure 4.5 compares the RDMs to their randomly shuffled counterparts for each ROI. Figure 4.6 shows the ROI RDMs averaged across subjects ($n = 19$).

4.3.3 Comparison between face-selective ROIs and hidden two-stream neural networks

A hidden two-stream neural network was trained to perform identity recognition on a video databases of faces. The hidden two-stream neural network can be broken down into three components: the spatial stream DCNN model, the optic flow DCNN model (MotionNet) and the temporal stream DCNN model. Figure 4.7 A and C show the DCNN model RDMs. RDMs from each model were correlated

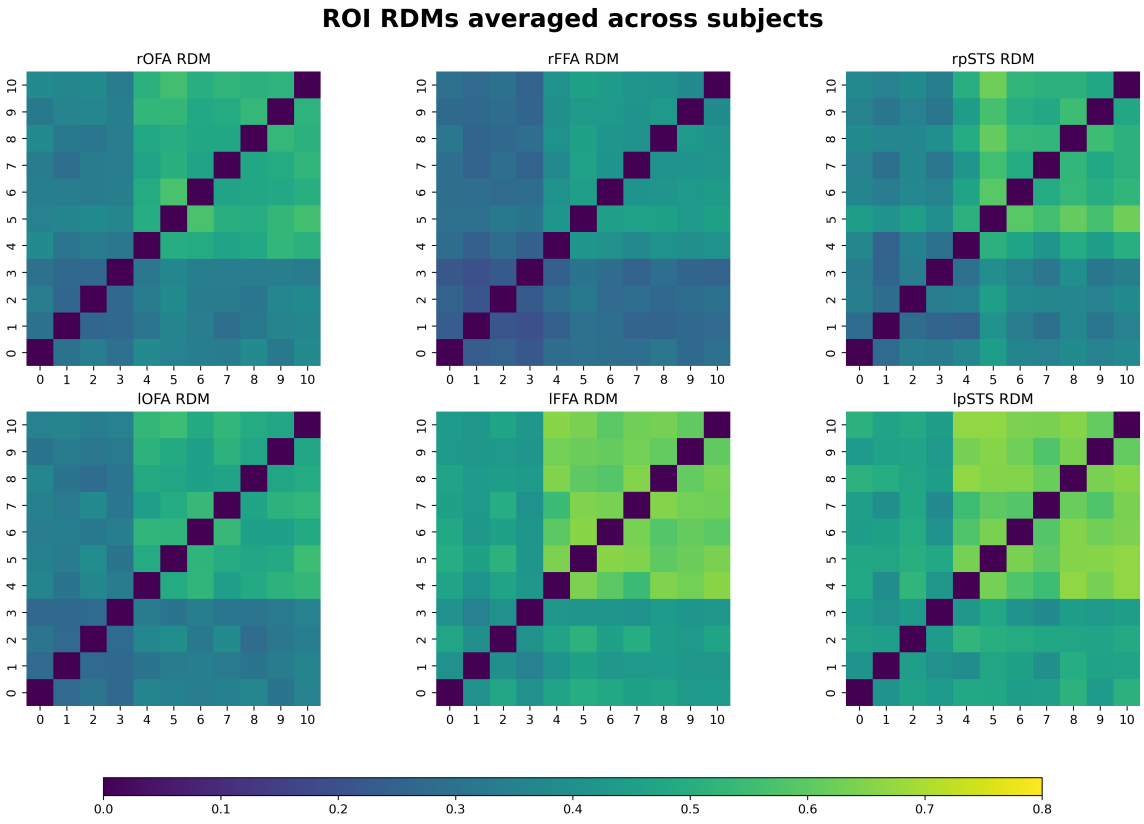

FIGURE 4.6: RDMs averaged across subjects for each face-selective ROI.

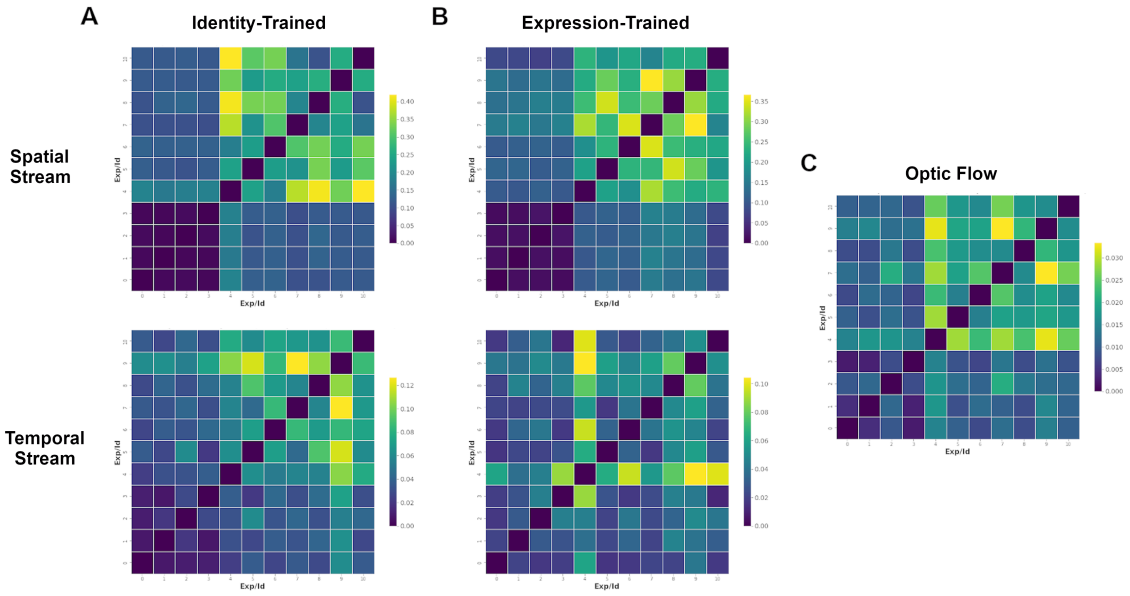

FIGURE 4.7: RDMs from identity and expression two-stream models and the optic flow model. A) RDMs of the fMRI stimuli representations extracted from the identity spatial stream DCNN and the identity temporal stream DCNN models. B) RDMs of the fMRI stimuli representations extracted from the expression spatial stream DCNN and the expression temporal stream DCNN models. C) RDMs of the fMRI stimuli optic flows extracted from the MotionNet model. For all RDMs, predictors 0-3 are facial expressions (disgust, fear/surprised, happy, sad) and 4-10 are the seven different face identities.

with subject RDMs for each ROI. Figure 4.8 shows the Kendall τ_B values averaged over subject with SEM bars and noise ceilings. All 3 identity DCNN models performed similarly for each ROI with the exception of the rpSTS that showed a smaller correlation with the optic flow model compared to the spatial and temporal stream models. The lpSTS showed the same pattern as rpSTS but to a lesser extent. Additionally, the rOFA and rFFA showed slightly greater correlations with the spatial and temporal DCNN models. These correlations were greater for the spatial and temporal stream DCNN models compared to the optic flow DCNN model within theses regions, whilst the lOFA and lFFA had greater correlations with the optic flow DCNN model compared to the spatial and temporal stream DCNN models. This could make sense due to the right hemisphere being more specialized for faces compared to the left, but the differences were small. Additionally, the rFFA had the highest correlations overall for the spatial and temporal stream DCNN model comparisons, with the spatial stream DCNN model τ_B as the largest.

A hidden two-stream neural network was trained to perform expression recognition on a video databases of faces. Figure 4.7 B and C shows the DCNN model RDMs. RDMs from each model were correlated with subject RDMs for each ROI. Figure 4.9 shows the Kendall τ_B values averaged over subject with SEM bars and noise ceilings. The expression spatial DCNN model had the highest correlations for each ROI. The rFFA showed slightly greater correlations with the spatial and optic flow DCNN models compared to the lFFA. This again makes sense due to the right hemisphere being more specialized for faces compared to the left, but the differences are small. Additionally, the rFFA had the highest correlations overall for the spatial and temporal stream DCNN model comparisons, with the spatial

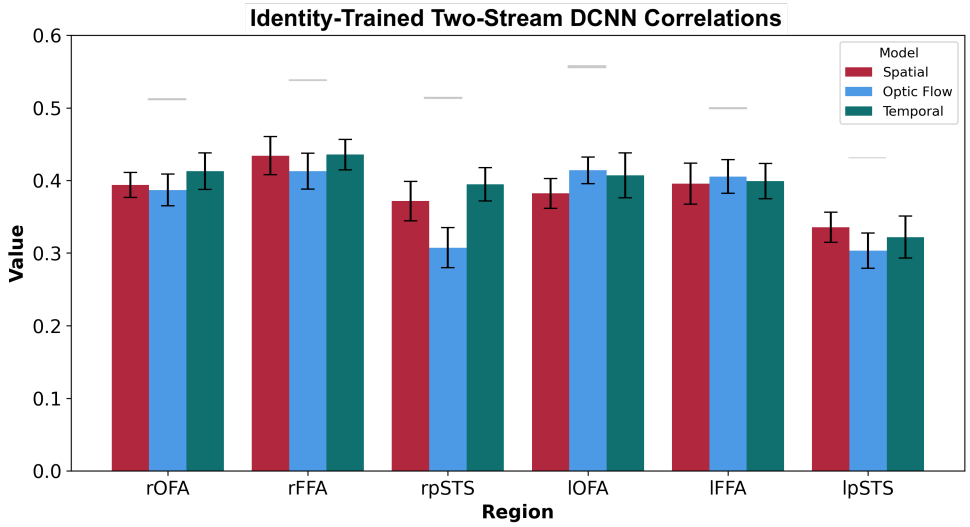

FIGURE 4.8: Correlations between neural RDMs and identity hidden two-stream model RDMs. Per each face-selective ROI, Kendall's τ_B was calculated for the spatial stream, the optic flow MotionNet, and the temporal stream, averaging over subjects ($n = 19$). SEM bars are depicted in black. The shaded grey regions represent the lower and upper bound of the noise ceiling for each ROI.

stream DCNN model as the largest. The expression temporal stream DCNN model had much lower correlations compared to the other two expression models as well as the identity stream models. However, the expression temporal stream was only partially trained, so it could be expected that these values would increase with a fully trained version of the expression temporal DCNN.

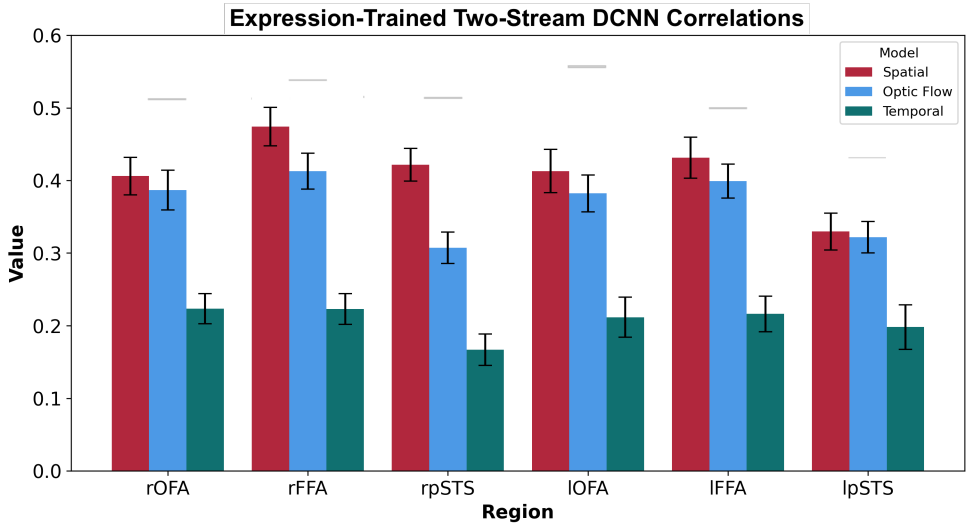

FIGURE 4.9: Correlations between neural RDMs and expression hidden two-stream model RDMs. Per each face-selective ROI, Kendall's τ_B was calculated for the spatial stream, the optic flow MotionNet, and the temporal stream, averaging over subjects ($n = 19$). SEM bars are depicted in black. The shaded grey regions represent the lower and upper bound of the noise ceiling for each ROI.

To test if the spatial streams of both identity and expression DCNNs similarly correlated with the ventral (rOFA, rFFA, lOFA, lFFA) and lateral regions (rpSTS, lpSTS), the same analysis implemented in Chapter 3 was done. If ventral ROIs have comparatively higher Kendall τ_B with the identity spatial stream DCNN, and lateral ROIs have comparatively higher Kendall τ_B with the expression spatial stream DCNN, the ventral and lateral ROIs should fall on two different lines with different slopes. However, if the findings from Chapter 3 were to be replicated using

these models and the fMRI data, then the ROIs should fall along one slope. Like before, the data points from the ventral ROI and the lateral ROI were located along one line (Figure 4.10 A). The BIC analysis confirmed this when modeling the ventral and lateral ROIs separately ($BIC_{separate} = -162.927$), and when combining the ventral and lateral regions($BIC_{combined} = -520.140$, $BIC_{separate} - BIC_{combined} = 357.213$). Differences greater than 10 in BIC values were interpreted as providing strong evidence in favor of the model with lower BIC (i.e. the combined model; Raftery 1995). This corresponded to the ventral and lateral ROI having a similar ratio of expression τ_B to identity τ_B (Figure 4.10 A).

This was then repeated using the temporal streams of both identity and expression DCNNs. Again, the data was better fitted when using a single slope for all ROIs. This is seen in the BIC analysis ($BIC_{separate} = -168.616$, $BIC_{combined} = -549.062$, $BIC_{separate} - BIC_{combined} = 380.446$). The scatter plot is shown in Figure 4.10 B.

For each subject, each ROI was evaluated to determine how similar it was to the identity spatial stream DCNN model RDM, and then how similar it was to the combined identity spatial and temporal stream DCNN model RDMs. The two values were used as coordinates for a scatter plot (Figure 4.11). If ventral ROIs have a comparatively higher Kendall's τ_B with the spatial stream DCNN model, and lateral ROIs have a comparatively higher Kendall τ_B with the combined spatial and temporal stream DCNN model, the two sets of ROIs should fall on two different lines with different slopes. Instead, for the identity DCNN model, all observed ROIs were located along one line - showing a similar ratio of the spatial model τ_B to the combined spatial and temporal DCNN model τ_B (Figure 4.11). To quantify this, BIC was used to compare a model with separate slopes for the ventral and lateral

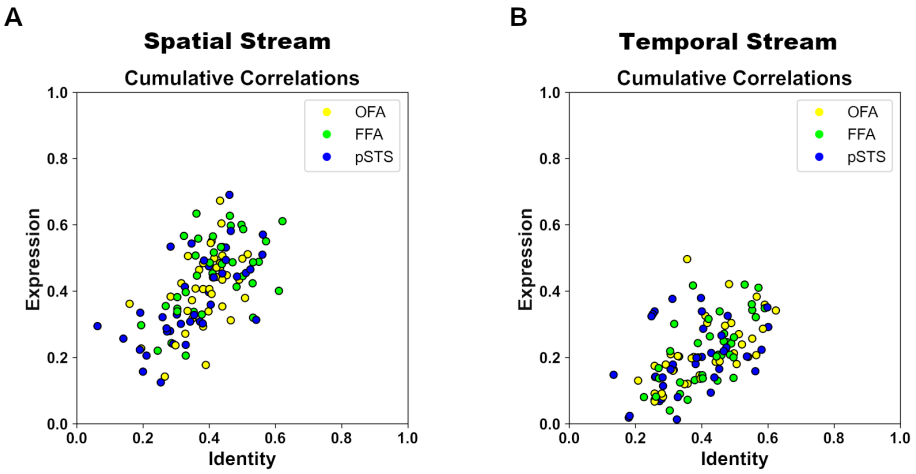

FIGURE 4.10: Relative contributions of identity and expression models within streams. A) Scatter plot comparing correlation values from the identity spatial stream and expression spatial stream using Kendall τ_B matched on ROI per subject (n=19 subjects). B) Scatter plot comparing correlation values from the identity temporal stream and expression temporal stream using Kendall τ_B matched on ROI per subject (n=19 subjects). For each ROI, the right and left hemispheres were included as individual data points in both plots.

ROIs separately ($BIC_{separate} = -166.846$) to a model with a single slope for both the ventral and lateral ROIs ($BIC_{combined} = -523.180$, $BIC_{separate} - BIC_{combined} = 356.334$). All of the ROIs fall along the same slope (Figure 4.11), suggesting that they have a similar ratio of match to the identity spatial stream and the combined identity spatial and temporal streams. To statistically test if the combined static and temporal τ_B correlations were greater than the static stream τ_B correlations for the lateral regions, a Wilcoxon signed-rank test was run. The combined static and temporal stream DCNN model showed significantly greater correlations with the lateral regions than the spatial stream DCNN model alone ($p < 0.001$).

This was then repeated using the expression DCNN model by comparing a model with separate slopes for the ventral and lateral ROIs separately ($BIC_{separate} = -161.087$) to a model with a single slope for both the ventral and lateral ROIs ($BIC_{combined} = -527.212$, $BIC_{separate} - BIC_{combined} = 366.125$). All of the ROIs fall along the same slope (Figure 4.12), suggesting that they have a similar ratio of match to the spatial stream and the combined spatial and temporal streams. Similar to the identity model analysis above, all of the ROIs fall along the same slope (Figure 4.12), suggesting that they have a similar ratio of match to the expression spatial stream and the combined expression spatial and temporal streams as well. To statistically test if the combined static and temporal τ_B correlations were greater than the static stream τ_B correlations for the lateral regions, a Wilcoxon signed-rank test was run. The combined static and temporal stream DCNN model showed significantly greater correlations with the lateral regions than the spatial stream DCNN model alone ($p < 0.01$).

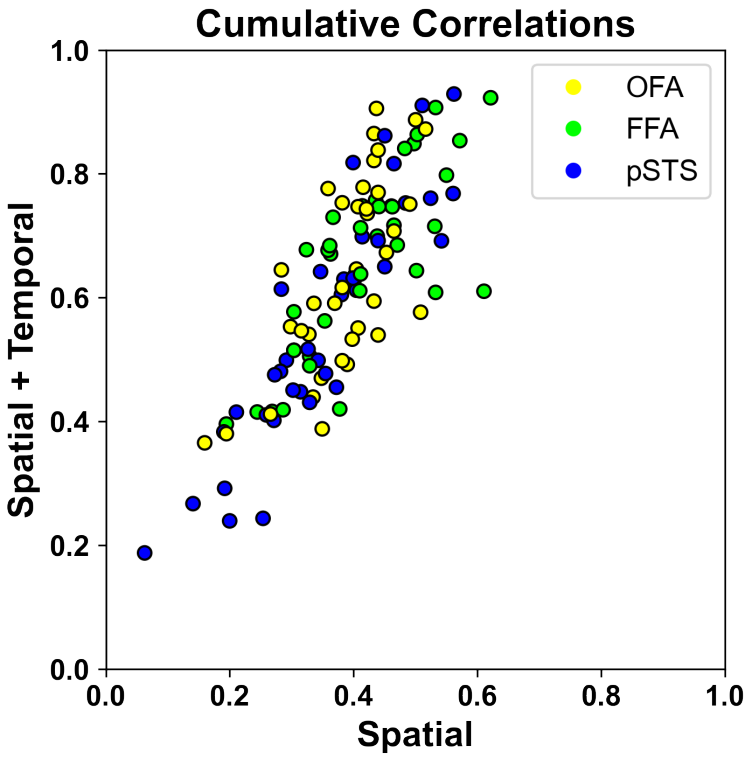

FIGURE 4.11: Variation across ROIs for identity model streams. A: Scatter plot comparing correlation values from the identity spatial stream and combined spatial and temporal streams using semi-partial Kendall τ_B matched on ROI per subject (n=19 subjects). For each ROI, the right and left hemispheres were included as individual data points.

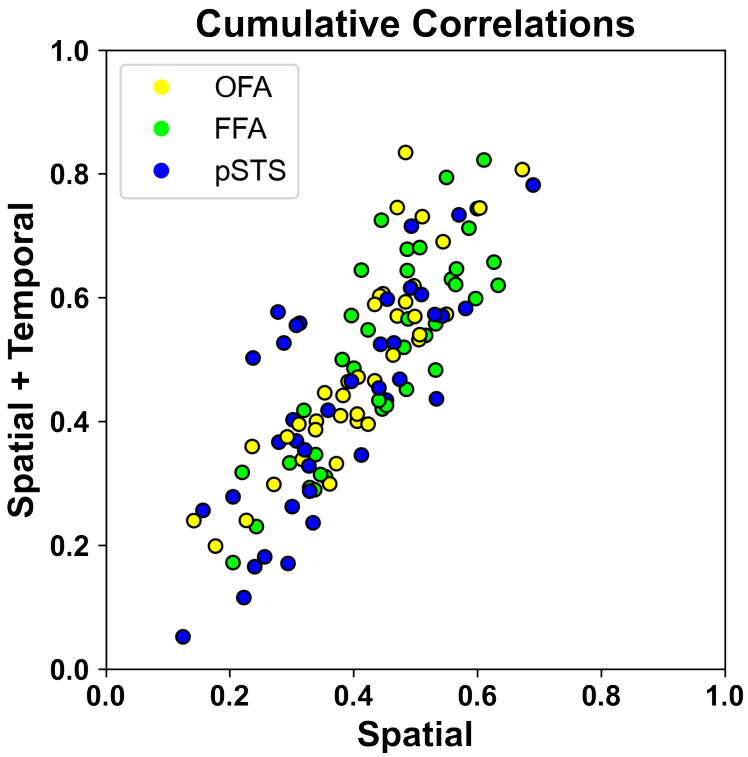

FIGURE 4.12: Variation across ROIs for expression model streams. A: Scatter plot comparing correlation values from the expression spatial stream and combined spatial and temporal streams using semi-partial Kendall τ_B matched on ROI per subject (n=19 subjects). For each ROI, the right and left hemispheres were included as individual data points.

4.4 Discussion

Originally, the distinction between the ventral and lateral temporal pathways in face perception was thought to be due to discrete neural mechanisms for processing invariant aspects of the face like face identity recognition, and for processing changeable aspects of the face like facial expression recognition. However, recent evidence had suggested something else: face identity recognition and facial expression recognition rely on shared neural mechanisms (Duchaine and Yovel, 2015; Bernstein and Yovel, 2015; Li, Richardson, and Ghuman, 2019; Schwartz et al., 2023b). In previous work, this finding was mostly restricted to static features due to evaluating static images of faces rather than dynamic faces. Thus, the first goal of this study was to determine if integrated representations of face identity and facial expression occur not only when processing static elements of face perception, but dynamic elements as well.

Since we used both a two-stream DCNN model trained for identity recognition and a two-stream DCNN model trained for expression recognition, the contribution of the static streams in isolation was evaluated first to see if the findings in Chapter 3 replicate when using a different neuroimaging modality. This was found to be the case: both identity and expression static stream DCNN models were able to explain neural responses in ventral and lateral regions. There was no interaction between the DCNN model type and the neural pathway, indicating that each model did not specifically correlate more strongly with a different pathway. Differing from what was found in Chapter 3, the identity spatial DCNN model did not outperform the expression stream DCNN model overall. This could be due to a variety of reasons. For instance, it may be due to a feature of the databases

or the change in neuroimaging modality from iEEG to fMRI. Importantly, this did not alter the interpretation of the results in terms of the relative contribution for the models for both sets of regions.

Next, to test if this phenomenon extended to dynamic attributes of face processing, the analysis was repeated using the temporal stream of both identity and expression models. We found that representations of identity and expression coexisting within the same regions occurred when processing face dynamics with both identity and expression temporal stream DCNN models being able to explain neural responses in ventral and lateral regions. Numerically, the identity model outperformed the expression model. However, further tests are needed to determine whether the difference is significant. Given the fact that the temporal stream for the expression model was only partially trained, this may be why.

Since we found that shared mechanisms for face identity and facial expression held true for dynamic faces, what could be an alternative dimension that may be able to explain the ventral and lateral functional roles? Based on a current theory of social perception, there is one pathway specialized for processing static (e.g., form) information, and another pathway specialized for processing both static and dynamic (e.g., velocity) information (Pitcher and Ungerleider, 2021). Static information is represented in regions in the ventral temporal cortex, while both static and dynamic information are represented in the lateral temporal cortex where both types of information can be integrated for social understanding. If this is the case, features optimized to recognize static information should better capture the similarity between neural responses in ventral regions, and features optimized to recognize static and dynamic information should better capture responses in lateral regions. Thus, the current theory would predict that RDMs from a model trained

to recognize identities from spatial inputs should correlate with RDMs from the ventral regions, while RDMs from both a model trained to recognize identities from spatial inputs as well as a model trained to recognize identity from optic flow estimation or motion inputs should correlate with RDMs from the lateral brain regions. However, this was not what was found: both ventral and lateral regions had comparable Kendall's τ_B values with each type of model, and were similarly explained by a combination of static and dynamic information.

The results presented here cannot be attributed to noise that could hide potential differences between ventral and lateral temporal face-selective ROIs. In fact, the correlation values obtained between the neural regions and the models would not be considered low, are on par with previous studies (Yamins et al., 2013). As an additional check for noise and consistency within a subject's neural data, Kendall's τ_B was calculated comparing subject run pairs with randomly shuffled RDMs. Figure 4.5 demonstrated that for each ROI the mean correlation for real data RDMs was higher than the shuffled data RDMs. The BIC analysis also showed strong support for the ventral and lateral ROIs to be modeled together with a single slope to account for the relative contribution of either static or static and dynamic information combined. It was also important to control for shared variance between the RDMs obtained from the spatial and temporal models when combining the two to get a cumulative value for combined static and dynamic information. The analysis used semi-partial correlations by computing the correlation of the residuals for the temporal stream model while controlling for the spatial stream model to account for redundant information that may be present in both spatial and temporal stream representations.

Face perception has typically been regarded as a specialized process that differs from all other object recognition, but it might have more in common with the mechanisms involved in perceiving bodies and actions than has been previously considered. Both macaques and humans have adjacent face-selective and body-selective brain areas (Pinsk et al., 2009; Arcaro et al., 2020), suggesting commonalities in the organization. Furthermore, a similar study done by the lab (Karimi et al., 2023) has been using fMRI data and two-stream video models to evaluate static and dynamic information in ventral and lateral regions for bodies and actions. Interestingly, the findings fit with the results reported here, also suggesting similar amounts of spatial and dynamic information for ventral and lateral regions. It is intriguing to see that this phenomenon seems to generalize for both action and face perception. Thus, face and body perception may not only be similar in terms of their large-scale architecture, but also in terms of how the static and dynamic information they both encode are organized.

There are a few documented cases in patient studies that have indicated motion-related deficits from ventral lesions (Gilaie-Dotan et al., 2013). These deficits, however, appear not to be crucial for processing biological motion, but rather pertain to the structure of the moving stimuli (Gilaie-Dotan et al., 2015). Conversely, there are also many cases where patients have ventral lesions and face recognition deficits, but do not have impairments for motion. Although, motion perception seemed to be spared in many of these patient cases, it would be unlikely that motion information is in the ventral regions accidentally. If not an accidental byproduct, what is the functional role of motion information in the ventral pathway? Work in

object segmentation, and understanding two dimensional (2D) and three dimensional (3D) layouts may be able to provide useful clues. Shape is typically considered an important static feature. However, it it difficult to perceive an object's 3D shape from a static image. Behavioral work has shown that people need to have seen a motion sequence in order to perceive the 3D shape of an object from its 2D image (Sinha and Poggio, 1996; Caudek and Rubin, 2001). Humans perceive structure from motion, a phenomena that has been well documented in work related to face and body perception as well (O'Toole, Roark, and Abdi, 2002; O'Toole et al., 2011). It may be possible that motion information is used in ventral temporal regions in order to extract necessary shape attributes, and contributes to learning the structure of visual stimuli. Predictive coding strategies related to learning spatiotemporal relationships may also play a role for learning invariance in recognition. Exploring the role of dynamic features in the ventral pathway will be an important direction to evaluate in future research.

Chapter 5

Discussion

5.1 Summary of findings

The aims of this dissertation were to 1) determine if it is possible to acquire face identity information without discarding facial expression information and vice versa, 2) evaluate the relative contributions of both identity and expression information in face-selective brain regions for static face stimuli and 3) extend this to dynamic faces, investigating whether and where both static and dynamic features are encoded in the ventral and lateral temporal neural pathways. A combination of neuroimaging techniques and deep learning methods allowed a nuanced assessment of the functional roles of the ventral and lateral temporal pathways in the brain, specifically within the realm of face perception. In Chapter 2, a proof-of-concept experiment was described to demonstrate that it is possible to learn identity and expression information together, without needing to discard information for one. Computational models were developed that can undermine the traditional view of abstraction in psychology (Posner, 1970), and within the scope of neuroscience research, implemented a novel techniques to further analyze RDMs using semi-partial correlations. In Chapter 3, the computational models were applied to neural data, and demonstrated that within face-selective regions of both the ventral and lateral temporal pathways, there were both face identity and facial expression information. The iEEG results, thus, argue against the weak explanation of the classical theory of face perception (where only a small amount of information for either identity or expression is leftover as a by-product). Chapter 4 further showed that integrated representations of identity and expression occur in dynamic face stimuli as well, and examined the functional roles that may differentiate between the ventral and lateral temporal pathways. Using fMRI, neural

responses were collected while subjects viewed videos of faces constructed from different identities and expressions. To investigate the representational content of the responses in the brain, computational models specialized for video tasks were trained to recognize face identity and compared to the neural data. These models had the unique property of implementing two streams for visual processing: a spatial stream and a temporal stream. The optimized streams were then used to investigate questions related to static and dynamic face processing in the brain. Unlike what was hypothesized, the results demonstrated that there was a similar amount of dynamic information represented in both ventral and lateral temporal face-selective regions.

In the following sections, the implications and limitations of the three studies are discussed in a broader context. More specifically, the discussion expands on how they contribute to our understanding of the functional organization of the brain in relation to vision and social perception. The brain functions as an optimized computational system, developing efficient mechanisms that enable our day-to-day functions. Given this, it is important to understand how the visual system develops to function as it does. It will be briefly discussed why we may have integrated representations in the brain to enable more efficient and accurate processing of complex information, and why this may hold true for both spatial and temporal modalities. The last section discusses how these findings may affect the way we think of the distinction between the ventral and lateral temporal neural pathways.

5.2 For what reason might we have representations of identity and expression in common brain regions?

Chapter 2 suggests that information about face identity and facial expression is progressively disentangled from layer to layer within DCNN models. A similar idea of untangling has been proposed before in object recognition which suggests that objects are entangled manifolds within a space, and that variations of the object are a point on a manifold (DiCarlo and Cox, 2007). This type of perspective is part of broader research efforts suggesting that information about object category, along with other properties such as position and aspect ratio, are disentangled while coexisting in IT representations (Hong et al., 2016). Audition studies have demonstrated that both speaker identity and speech content can be decoded in the superior temporal cortex (Formisano et al., 2008; Bonte et al., 2014), indicating similar phenomena are observed in other cognitive domains. Taken together, these studies reveal that some sets of complementary tasks — such as recognition of face identity and facial expression or speaker identity and speech content — rely on common brain regions, while others — like the recognition of faces and places — are implemented by distinct neural substrates. It can be speculated that this pattern of integrated representations might be driven by constraints derived from computational efficiency (e.g., the number of neurons allocated, the amount of training input needed). Recent work has begun to investigate these computational constraints in the field of deep learning and computer science, testing what are the optimal ways of structuring and sharing representations across multiple tasks (Zamir et al., 2018). This proposal is broadly related to the concept of a taxonomy of tasks ('Taskonomy', Zamir et al., 2018, Wang, Tarr, and Wehbe, 2019)

which describes a structural space to relate tasks to one another. Importantly, this framework is not limited to the domain of vision but can be applied across various fields. An interesting future direction would include modeling neural responses using multi-task networks that learn to recognize both face identity and facial expression as mentioned in the Discussion section of Chapter 2.

In the context of using machine learning as a tool to model the brain, it has been shown that object-trained DCNNs can correlate with face regions to a larger extent than DCNNs trained for identity recognition (Grossman et al., 2019; Chang et al., 2021). These findings might lead to the explanation that the neural regions are not partaking in domain-specific processing of categories, but rather, are doing domain-general processing (Vinken, Konkle, and Livingstone, 2022). However, as mentioned in the Discussion of Chapter 3, this is not a necessary conclusion given the results. There is an alternative interpretation according to which these regions would be expected to be better predicted by the object-trained DCNNs even if neural processing is domain-specific. In particular, brain regions might be involved in supporting multiple different face perception tasks (e.g., recognition of face identity, face viewpoint, facial expressions, age, eye gaze). A model that is trained to do only one of these tasks would only capture part of the information encoded by these regions, and therefore, its performance at predicting neural responses may be lacking. By contrast, because non-face objects have a wide variety of features and shapes, models trained to recognize non-face objects might learn representations that are sufficiently varied that could be used to perform quite accurately at multiple different kinds of face tasks. In line with this, Chang et al. (2020) studying macaque monkeys found that neural networks trained to perform face recognition did not predict neural responses to faces well. Instead, their work suggested that

generative models of faces best explain neural responses. They proposed that face-selective cells in macaque monkeys have high-level information associated with different face features and that these features are filtered out in DCNNs trained to specifically perform face identification. However, generative face models learn latent representations that do not necessarily relate distinctly to identity. Thus, the generative models would better incorporate explained variance, and thereby predict with higher accuracy due to the models having learned additional information that is not directly related to face identity. This finding is in concordance with the hypothesis that identity and expression are processed by shared mechanisms, in line with the computational results observed in Chapter 2. Importantly, face processing in the ventral stream may not follow the traditional model of abstraction that sheds non-target information (Posner, 1970), but rather it may build representations of multiple different aspects of the face.

This research demonstrated that spontaneous learning of expression representations occur when DCNN models are trained to label identity (and vice versa). Importantly, the integrated representations found in both DCNN models were able to correlate with sets of regions previously thought to encode separate information related to faces. As mentioned earlier, similar phenomena related to the Integrated Representation of Identity and Expression Hypothesis (IRIEH, described in Chapter 2) might be in play in other cognitive areas. Moving forward, there are various ways this can be extended to other modalities and social processes. More broadly, one could speculate that implementation of more integrated computations might be a large-scale principle of organization of human cortex, determining which sets of cognitive processes are represented within the same neural systems. As such, disentanglement of shared mechanisms could apply to cases as diverse as word

recognition and speaker recognition in speech processing, syntax and semantics in language, and the inference of mental states and traits in social cognition.

5.3 Which dimensions may serve to define the functional roles of the ventral temporal and lateral temporal pathways?

If ventral and lateral temporal regions are not specifically specialized for the recognition of identity and expression respectively, but rather consist of integrated representations for related tasks, do these two pathways serve the same functional role? Otherwise, what are the functional differences between them? Studies combining TMS and fMRI have suggested that the pSTS receives inputs from both motion-responsive regions and form-encoding regions (Pitcher, Duchaine, and Walsh, 2014). Furthermore, the pSTS is also involved in audiovisual integration (Nath and Beauchamp, 2012; Anzellotti and Caramazza, 2017; Rennig and Beauchamp, 2021), which is reasonable given that audition information also has a temporal component. Based on the evidence, lateral regions along the STS might integrate static visual information, dynamic visual information, and auditory information, whereas regions in the ventral pathway are specialized for shape information (Duchaine and Yovel, 2015). Chapter 4 aimed to elucidate the functional role of these two pathways.

Curiously, the findings in Chapter 4 did not support a differentiation of static and dynamic information in the ventral temporal and lateral temporal pathways when processing the neural representations of dynamic face stimuli. In concordance with this, recent work from Karimi et al. (2023) has also shown this to be

the case when processing videos of bodies performing actions. Therefore, it is important to investigate why there could be motion information in ventral temporal regions, and what may be the differentiating feature between ventral and lateral temporal pathways if not dynamic processing.

As mentioned in the Discussion of Chapter 4, research in object segmentation and the way humans comprehend 2D and 3D layouts might offer valuable insights. Shape is generally thought of as a static feature, yet it is challenging to perceive an object's 3D shape from a static image alone. Structure from motion is the process of estimating 3D structure from 2D images. In fact, people typically need to observe a motion sequence to discern the 3D shape of an object from its 2D representation (Sinha and Poggio, 1996). This fits with studies showing that effects of face familiarity increase with motion (O'Toole, Roark, and Abdi, 2002). Thus, motion information might be represented in the ventral temporal regions to extract crucial shape features and aid in learning the structure of visual stimuli.

5.4 Relevance to other research areas

The findings here may be able to be applied to clinical areas as well, particularly for understanding impairments in prosopagnosia and potentially in addressing deficits in face and biological motion perception found in some individuals with autism spectrum disorder. In a different direction, this work is valuable for exploring how artificial intelligence (AI) can enhance our understanding of the human brain. Similarly, insights into vision and the brain can also help to inform advancements in AI, allowing AI to potentially better emulate human cognitive functioning.

5.5 Conclusion

These approaches enabled a more detailed investigation of the functional roles of the ventral and lateral temporal pathways in the brain, particularly in the context of face-specificity by evaluating the representational content in face-selective brain regions. The research here provided essential groundwork for understanding perceptual representations and their organization in the brain, potentially elucidating how the brain effectively perceives faces and other aspects of the social world.

Appendix A

Chapter 2 Supplementary Materials

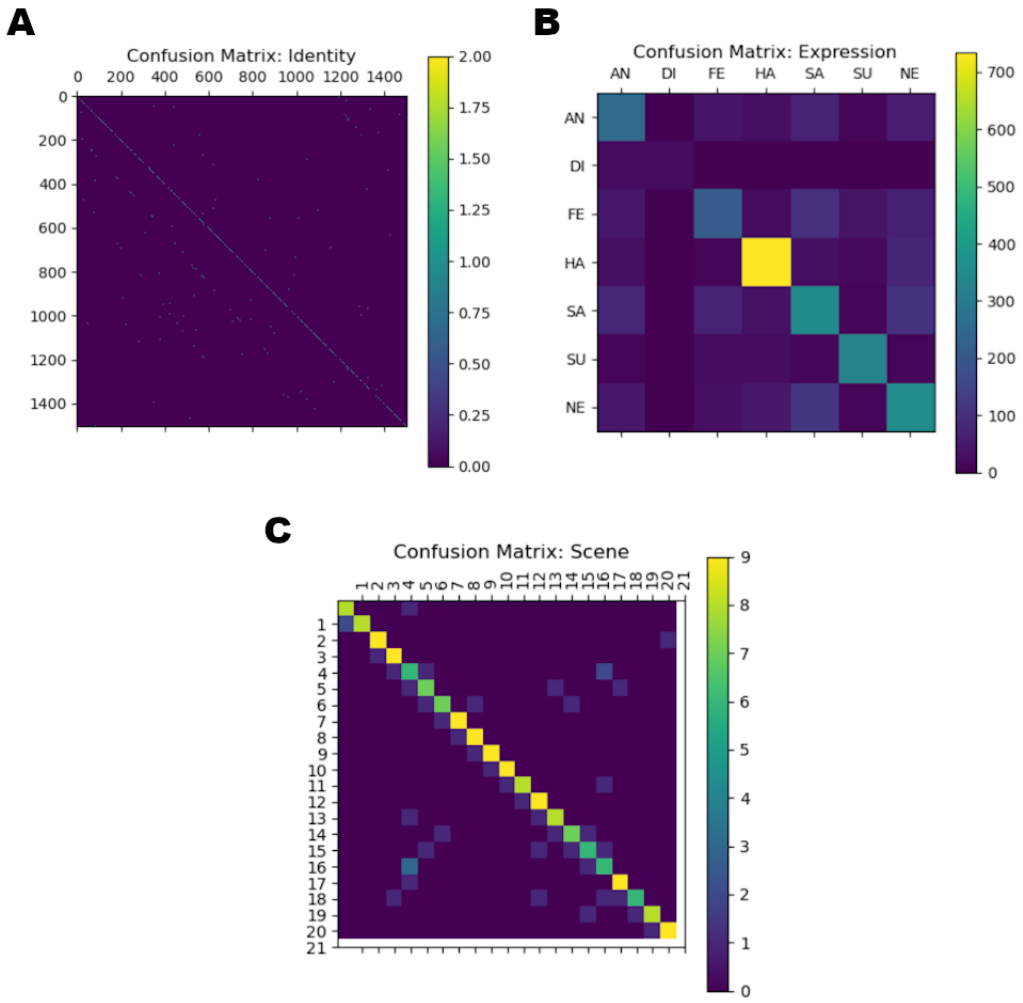

FIGURE A.1: Identity, expression, and scene DCNN confusion matrices. A) The confusion matrix of the identity-trained DCNN based on identity classification performance on the identity validation test set. B) The confusion matrix of the expression-trained DCNN based on expression classification performance on the expression validation test set. C) The confusion matrix of the scene-trained DCNN based on scene classification performance on the scene validation test set.

Bibliography

Allison, Truett et al. (1994). "Human extrastriate visual cortex and the perception of faces, words, numbers, and colors". In: *Cerebral cortex* 4.5, pp. 544–554.

Andrews, Timothy J and Michael P Ewbank (2004). "Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe". In: *Neuroimage* 23.3, pp. 905–913.

Anzellotti, Stefano and Alfonso Caramazza (2016). "From parts to identity: invariance and sensitivity of face representations to different face halves". In: *Cerebral Cortex* 26.5, pp. 1900–1909.

— (2017). "Multimodal representations of person identity individuated with fMRI". In: *Cortex* 89, pp. 85–97.

Anzellotti, Stefano, Scott L Fairhall, and Alfonso Caramazza (2013). "Decoding representations of face identity that are tolerant to rotation". In: *Cerebral Cortex* 24.8, pp. 1988–1995.

— (2014). "Decoding representations of face identity that are tolerant to rotation". In: *Cerebral Cortex* 24.8, pp. 1988–1995.

Anzellotti, Stefano and Liane L Young (2020). "The Acquisition of Person Knowledge". In: *Annual Review of Psychology* 71, pp. 613–634.

Arcaro, Michael J and Margaret S Livingstone (2017). "A hierarchical, retinotopic proto-organization of the primate visual system at birth". In: *Elife* 6, e26196.

Arcaro, Michael J et al. (2017). "Seeing faces is necessary for face-domain formation". In: *Nature neuroscience* 20.10, pp. 1404–1412.

Arcaro, Michael J et al. (2020). "Anatomical correlates of face patches in macaque inferotemporal cortex". In: *Proceedings of the National Academy of Sciences* 117.51, pp. 32667–32678.

Ashburner, John et al. (2014). "SPM12 manual". In: *Wellcome Trust Centre for Neuroimaging, London, UK* 2464.4.

Atkinson, Anthony P, Quoc C Vuong, and Hannah E Smithson (2012). "Modulation of the face-and body-selective visual regions by the motion and emotion of point-light face and body stimuli". In: *Neuroimage* 59.2, pp. 1700–1712.

Aviezer, Hillel, Yaacov Trope, and Alexander Todorov (2012). "Body cues, not facial expressions, discriminate between intense positive and negative emotions". In: *Science* 338.6111, pp. 1225–1229.

Axelrod, Vadim and Galit Yovel (2015). "Successful decoding of famous faces in the fusiform face area". In: *PloS one* 10.2, e0117126.

Barbeau, Emmanuel J et al. (2008). "Spatio temporal dynamics of face recognition". In: *Cerebral Cortex* 18.5, pp. 997–1009.

Beauchamp, Michael S, Audrey R Nath, and Siavash Pasalar (2010). "fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect". In: *Journal of Neuroscience* 30.7, pp. 2414–2417.

Bentin, Shlomo et al. (1996). "Electrophysiological studies of face perception in humans". In: *Journal of cognitive neuroscience* 8.6, pp. 551–565.

Bernstein, Michal and Galit Yovel (2015). "Two neural pathways of face processing: a critical evaluation of current models". In: *Neuroscience & Biobehavioral Reviews* 55, pp. 536–546.

Bombari, Dario et al. (2013). "Emotion recognition: The role of featural and configural face information". In: *Quarterly Journal of Experimental Psychology* 66.12, pp. 2426–2442.

Bonte, Milene et al. (2014). "Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns". In: *Journal of Neuroscience* 34.13, pp. 4548–4557.

Boring, Matthew J et al. (2021). "Multiple adjoining word-and face-selective regions in ventral temporal cortex exhibit distinct dynamics". In: *Journal of Neuroscience* 41.29, pp. 6314–6327.

Bracci, Stefania and Hans Op de Beeck (2016). "Dissociations and associations between shape and category representations in the two visual pathways". In: *Journal of Neuroscience* 36.2, pp. 432–444.

Bracci, Stefania et al. (2019). "The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks". In: *Journal of Neuroscience* 39.33, pp. 6513–6525.

Brett, Matthew et al. (2002). "Region of interest analysis using an SPM toolbox". In: *8th international conference on functional mapping of the human brain*. Vol. 16. 2. Sendai, p. 497.

Bruce, Vicki (1982). "Changing faces: Visual and non-visual coding processes in face recognition". In: *British journal of psychology* 73.1, pp. 105–116.

Bruce, Vicki and Andy Young (1986). "Understanding face recognition". In: *British journal of psychology* 77.3, pp. 305–327.

Burton, A Mike et al. (1999). "Face recognition in poor-quality video: Evidence from security surveillance". In: *Psychological Science* 10.3, pp. 243–248.

Butcher, Natalie and Karen Lander (2017). "Exploring the motion advantage: Evaluating the contribution of familiarity and differences in facial motion". In: *The Quarterly Journal of Experimental Psychology* 70.5, pp. 919–929.

Calder, Andrew J (2011). "Does facial identity and facial expression recognition involve separate visual routes". In: *The Oxford handbook of face perception*, pp. 427–448.

Calder, Andrew J and Andrew W Young (2005). "Understanding the recognition of facial identity and facial expression". In: *Nature Reviews Neuroscience* 6.8, p. 641.

Calvo, Manuel G and Daniel Lundqvist (2008). "Facial expressions of emotion (KDEF): Identification under different display-duration conditions". In: *Behavior research methods* 40.1, pp. 109–115.

Caramazza, Alfonso and Jennifer R Shelton (1998). "Domain-specific knowledge systems in the brain: The animate-inanimate distinction". In: *Journal of cognitive neuroscience* 10.1, pp. 1–34.

Caramazza, Alfonso et al. (1990). "The multiple semantics hypothesis: Multiple confusions?" In: *Cognitive neuropsychology* 7.3, pp. 161–189.

Caudek, Corrado and Nava Rubin (2001). "Segmentation in structure from motion: modeling and psychophysics". In: *Vision Research* 41.21, pp. 2715–2732.

Chang, Le et al. (2020). "What computational model provides the best explanation of face representations in the primate brain?" In: *bioRxiv*.

Chang, Le et al. (2021). "Explaining face representation in the primate brain using different computational models". In: *Current Biology* 31.13, pp. 2785–2795.

Chao, Linda L, James V Haxby, and Alex Martin (1999). "Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects". In: *Nature neuroscience* 2.10, pp. 913–919.

Chung, Jihoon et al. (2021). "Haa500: Human-centric atomic action dataset with curated videos". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13465–13474.

Colón, Y Ivette, Carlos D Castillo, and Alice J O'Toole (2021). "Facial expression is retained in deep networks trained for face identification". In: *Journal of Vision* 21.4, pp. 4–4.

Connolly, Hannah L, Andrew W Young, and Gary J Lewis (2019). "Recognition of facial expression and identity in part reflects a common ability, independent of general intelligence and visual short-term memory". In: *Cognition and Emotion* 33.6, pp. 1119–1128.

Conway, Bevil R (2018). "The organization and operation of inferior temporal cortex". In: *Annual review of vision science* 4, pp. 381–402.

Dahl, George E, Tara N Sainath, and Geoffrey E Hinton (2013). "Improving deep neural networks for LVCSR using rectified linear units and dropout". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 8609–8613.

Deen, Ben et al. (2015). "Functional organization of social perception and cognition in the superior temporal sulcus". In: *Cerebral cortex* 25.11, pp. 4596–4609.

DiCarlo, James J and David D Cox (2007). "Untangling invariant object recognition". In: *Trends in cognitive sciences* 11.8, pp. 333–341.

DiCarlo, James J, Davide Zoccolan, and Nicole C Rust (2012). "How does the brain solve visual object recognition?" In: *Neuron* 73.3, pp. 415–434.

Dobs, Katharina, Isabelle Bülthoff, and Johannes Schultz (2016). "Identity information content depends on the type of facial movement". In: *Scientific reports* 6, p. 34301.

— (2018). "Use and usefulness of dynamic face stimuli for face perception studies—A review of behavioral findings and methodology". In: *Frontiers in psychology* 9, p. 1355.

Dobs, Katharina, Wei Ji Ma, and Leila Reddy (2017). "Near-optimal integration of facial form and motion". In: *Scientific reports* 7.1, p. 11002.

Dobs, Katharina et al. (2018). "Task-dependent enhancement of facial expression and identity representations in human cortex". In: *NeuroImage* 172, pp. 689–702.

Dobs, Katharina et al. (2019). "Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational Hypotheses with Deep Convolutional Neural Networks". In: Oral presentation at Cognitive Computational Neuroscience Conference, Berlin ...

Dobs, Katharina et al. (2022). "Brain-like functional specialization emerges spontaneously in deep neural networks". In: *Science advances* 8.11, eabl8913.

Dobs, Katharina et al. (2023). "Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition". In: *Proceedings of the National Academy of Sciences* 120.32, e2220642120.

Doshi, Fenil R and Talia Konkle (2023). "Cortical topographic motifs emerge in a self-organized map of object space". In: *Science Advances* 9.25, eade8187.

Downing, Paul E et al. (2001). "A cortical area selective for visual processing of the human body". In: *Science* 293.5539, pp. 2470–2473.

Duchaine, Brad and Galit Yovel (2015). "A revised neural framework for face processing". In: *Annual review of vision science* 1, pp. 393–416.

Edelman, Shimon et al. (1998). "Toward direct visualization of the internal shape representation space by fMRI". In: *Psychobiology* 26, pp. 309–321.

Eimer, Martin (2000). "Event-related brain potentials distinguish processing stages involved in face perception and recognition". In: *Clinical neurophysiology* 111.4, pp. 694–705.

Epstein, Russell and Nancy Kanwisher (1998). "A cortical representation of the local visual environment". In: *Nature* 392.6676, pp. 598–601.

Epstein, Russell A (2008). "Parahippocampal and retrosplenial contributions to human spatial navigation". In: *Trends in cognitive sciences* 12.10, pp. 388–396.

Esteban, Oscar et al. (2019). "fMRIPrep: a robust preprocessing pipeline for functional MRI". In: *Nature methods* 16.1, pp. 111–116.

Etcoff, Nancy L (1984). "Selective attention to facial identity and facial emotion". In: *Neuropsychologia* 22.3, pp. 281–295.

Fang, Mengting, Craig Poskanzer, and Stefano Anzellotti (2022). "Pymvpd: A toolbox for multivariate pattern dependence". In: *Frontiers in Neuroinformatics* 16, p. 835772.

Feather, Jenelle et al. (2019). "Metamers of neural networks reveal divergence from human perceptual systems". In: *Advances in Neural Information Processing Systems*, pp. 10078–10089.

Felleman, Daniel J and David C Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex." In: *Cerebral cortex (New York, NY: 1991)* 1.1, pp. 1–47.

Finzi, Dawn et al. (2023). “A single computational objective drives specialization of streams in visual cortex”. In: *bioRxiv*, pp. 2023–12.

Formisano, Elia et al. (2008). “" Who" is saying" what"? Brain-based decoding of human voice and speech”. In: *Science* 322.5903, pp. 970–973.

Fox, Christopher J et al. (2011). “Perceptual and anatomic patterns of selective deficits in facial identity and expression processing”. In: *Neuropsychologia* 49.12, pp. 3188–3200.

Furl, Nicholas et al. (2013). “Top-down control of visual responses to fear by the amygdala”. In: *Journal of Neuroscience* 33.44, pp. 17435–17443.

Gauthier, Isabel and Michael J Tarr (1997). “Becoming a “Greeble” expert: Exploring mechanisms for face recognition”. In: *Vision research* 37.12, pp. 1673–1682.

Gauthier, Isabel et al. (1999). “Activation of the middle fusiform’face area’increases with expertise in recognizing novel objects”. In: *Nature neuroscience* 2.6, pp. 568–573.

Gauthier, Isabel et al. (2000). “The fusiform “face area” is part of a network that processes faces at the individual level”. In: *Journal of cognitive neuroscience* 12.3, pp. 495–504.

Ghuman, Avniel Singh et al. (2014). “Dynamic encoding of face information in the human fusiform gyrus”. In: *Nature communications* 5.1, pp. 1–10.

Gilaie-Dotan, Sharon et al. (2013). “The role of human ventral visual cortex in motion perception”. In: *Brain* 136.9, pp. 2784–2798.

Gilaie-Dotan, Sharon et al. (2015). “Ventral aspect of the visual form pathway is not critical for the perception of biological motion”. In: *Proceedings of the National Academy of Sciences* 112.4, E361–E370.

Goeleven, Ellen et al. (2008). “The Karolinska directed emotional faces: a validation study”. In: *Cognition and emotion* 22.6, pp. 1094–1118.

Goodale, Melvyn A and A David Milner (1992). “Separate visual pathways for perception and action”. In: *Trends in neurosciences* 15.1, pp. 20–25.

Goodale, Melvyn A et al. (1991). “A neurological dissociation between perceiving objects and grasping them”. In: *Nature* 349.6305, pp. 154–156.

Goodale, Melvyn A et al. (1994). “Separate neural pathways for the visual analysis of object shape in perception and prehension”. In: *Current Biology* 4.7, pp. 604–610.

Goodfellow, Ian J et al. (2013). “Challenges in representation learning: A report on three machine learning contests”. In: *International Conference on Neural Information Processing*. Springer, pp. 117–124.

Grosbras, Marie-Hélène, Susan Beaton, and Simon B Eickhoff (2012). “Brain regions involved in human movement perception: A quantitative voxel-based meta-analysis”. In: *Human brain mapping* 33.2, pp. 431–454.

Grossman, Emily et al. (2000). “Brain areas involved in perception of biological motion”. In: *Journal of cognitive neuroscience* 12.5, pp. 711–720.

Grossman, Emily D, Lorella Battelli, and Alvaro Pascual-Leone (2005). “Repetitive TMS over posterior STS disrupts perception of biological motion”. In: *Vision research* 45.22, pp. 2847–2853.

Grossman, Shany et al. (2019). “Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks”. In: *Nature communications* 10.1, pp. 1–13.

Gschwind, Markus et al. (2012). “White-matter connectivity between face-responsive regions in the human brain”. In: *Cerebral cortex* 22.7, pp. 1564–1576.

Guo, Kun (2012). "Holistic gaze strategy to categorize facial expression of varying intensities". In.

Harel, Assaf et al. (2016). "The temporal dynamics of scene processing: A multi-faceted EEG investigation". In: *Eneuro* 3.5.

Hasan, Bashar Awwad Shiekh et al. (2016). ""Hearing faces and seeing voices": Amodal coding of person identity in the human brain". In: *Scientific reports* 6, p. 37494.

Haxby, James V, Elizabeth A Hoffman, and M Ida Gobbini (2000). "The distributed human neural system for face perception". In: *Trends in cognitive sciences* 4.6, pp. 223–233.

Haxby, James V et al. (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex". In: *Science* 293.5539, pp. 2425–2430.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hemond, Christopher C, Nancy G Kanwisher, and Hans P Op de Beeck (2007). "A preference for contralateral stimuli in human object-and face-selective cortex". In: *PLoS one* 2.6, e574.

Hendry, Stewart HC and R Clay Reid (2000). "The koniocellular pathway in primate vision". In: *Annual review of neuroscience* 23.1, pp. 127–153.

Hermes, Dora et al. (2010). "Automated electrocorticographic electrode localization on individually rendered brain surfaces". In: *Journal of neuroscience methods* 185.2, pp. 293–298.

Higgins, Irina et al. (2021). "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons". In: *Nature communications* 12.1, pp. 1–14.

Hoffman, Elizabeth A and James V Haxby (2000). "Distinct representations of eye gaze and identity in the distributed human neural system for face perception". In: *Nature neuroscience* 3.1, p. 80.

Hong, Ha et al. (2016). "Explicit information for category-orthogonal object properties increases along the ventral stream". In: *Nature neuroscience* 19.4, pp. 613–622.

Hornak, J, ET Rolls, and D Wade (1996). "Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage". In: *Neuropsychologia* 34.4, pp. 247–261.

Huang, Gao et al. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167*.

Isik, Leyla et al. (2017). "Perceiving social interactions in the posterior superior temporal sulcus". In: *Proceedings of the National Academy of Sciences* 114.43, E9145–E9152.

Jansari, Ashok et al. (2015). "The man who mistook his neuropsychologist for a popstar: when configural processing fails in acquired prosopagnosia". In: *Frontiers in Human Neuroscience* 9, p. 390.

Jiang, Xingxun et al. (2020). "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild". In: *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889.

Julian, Joshua B, Jack Ryan, and Russell A Epstein (2017). "Coding of object size and object category in human visual cortex". In: *Cerebral cortex* 27.6, pp. 3095–3109.

Julian, Joshua B et al. (2012). "An algorithmic method for functionally defining regions of interest in the ventral visual pathway". In: *Neuroimage* 60.4, pp. 2357–2364.

Kanwisher, Nancy, Josh McDermott, and Marvin M Chun (1997). "The fusiform face area: a module in human extrastriate cortex specialized for face perception". In: *Journal of neuroscience* 17.11, pp. 4302–4311.

Kanwisher, Nancy, Carol Yin, and Ewa Wojciulik (1999). "Repetition Blindness for Pictures: Evidence for the Rapid Computation of Abstract Visual". In: *Fleeting memories: Cognition of brief visual stimuli*, p. 119.

Kanwisher, Nancy et al. (1997). "A locus in human extrastriate cortex for visual shape analysis". In: *Journal of Cognitive Neuroscience* 9.1, pp. 133–142.

Kar, Kohitij et al. (2019). "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior". In: *Nature neuroscience* 22.6, pp. 974–983.

Karimi, Hamed et al. (2023). "Modeling fMRI responses to complex dynamic stimuli with two-stream convolutional networks". In: *Journal of Vision* 23.9. DOI: 10.1167/jov.23.9.5348. URL: https://jov.arvojournals.org/article.aspx?articleid=2792047.

Keysers, Christian, Valeria Gazzola, and Eric-Jan Wagenmakers (2020). "Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence". In: *Nature neuroscience* 23.7, pp. 788–799.

Khaligh-Razavi, Seyed-Mahdi and Nikolaus Kriegeskorte (2014). "Deep supervised, but not unsupervised, models may explain IT cortical representation". In: *PLoS computational biology* 10.11, e1003915.

Kheradpisheh, Saeed Reza et al. (2016). "Deep networks can resemble human feed-forward vision in invariant object recognition". In: *Scientific reports* 6, p. 32672.

Kietzmann, Tim C et al. (2019). "Recurrence is required to capture the representational dynamics of the human visual system". In: *Proceedings of the National Academy of Sciences* 116.43, pp. 21854–21863.

Kietzmann, Tim Christian, Patrick McClure, and Nikolaus Kriegeskorte (2018). "Deep neural networks in computational neuroscience". In: *bioRxiv*, p. 133504.

Kim, Seongho (2015). "ppcor: an R package for a fast calculation to semi-partial correlation coefficients". In: *Communications for statistical applications and methods* 22.6, p. 665.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kliemann, Dorit et al. (2018). "Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without Autism". In: *Cortex* 103, pp. 24–43.

Komatsu, Hidehiko and ROBERT H Wurtz (1989). "Modulation of pursuit eye movements by stimulation of cortical areas MT and MST". In: *Journal of Neurophysiology* 62.1, pp. 31–47.

Konkle, Talia and Alfonso Caramazza (2013). "Tripartite organization of the ventral stream by animacy and object size". In: *Journal of Neuroscience* 33.25, pp. 10235–10242.

Konkle, Talia and Aude Oliva (2012). "A real-world size organization of object responses in occipitotemporal cortex". In: *Neuron* 74.6, pp. 1114–1124.

Kornblith, Simon et al. (2013). "A network for scene processing in the macaque temporal lobe". In: *Neuron* 79.4, pp. 766–781.

Kosakowski, Heather L et al. (2022). "Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants". In: *Current Biology* 32.2, pp. 265–274.

Kriegeskorte, Nikolaus and Rogier A Kievit (2013). "Representational geometry: integrating cognition, computation, and the brain". In: *Trends in cognitive sciences* 17.8, pp. 401–412.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25.

Krzanowski, WJ (1979). "Between-groups comparison of principal components". In: *Journal of the American Statistical Association* 74.367, pp. 703–707.

Leventhal, Audie G, RW Rodieck, and B Dreher (1981). "Retinal ganglion cell classes in the Old World monkey: morphology and central projections". In: *Science* 213.4512, pp. 1139–1142.

Li, Yuanning, R Mark Richardson, and Avniel Singh Ghuman (2019). "Posterior fusiform and midfusiform contribute to distinct stages of facial expression processing". In: *Cerebral Cortex* 29.7, pp. 3209–3219.

Liu, Ziwei et al. (2015). “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*.

— (2018). “Large-scale celebfaces attributes (celeba) dataset”. In: *Retrieved August* 15.2018, p. 11.

Livingstone, Margaret and David Hubel (1988). “Segregation of form, color, movement, and depth: anatomy, physiology, and perception”. In: *Science* 240.4853, pp. 740–749.

Livingstone, Margaret S et al. (2017). “Development of the macaque face-patch system”. In: *Nature communications* 8.1, p. 14897.

Long, Bria, Chen-Ping Yu, and Talia Konkle (2018). “Mid-level visual features underlie the high-level categorical organization of the ventral stream”. In: *Proceedings of the National Academy of Sciences* 115.38, E9015–E9024.

Lundqvist, Daniel, Anders Flykt, and Arne Öhman (1998). “The Karolinska directed emotional faces (KDEF)”. In: *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet* 91, p. 630.

Macko, Kathleen A et al. (1982). “Mapping the primate visual system with [2-14C] deoxyglucose”. In: *Science* 218.4570, pp. 394–397.

Madan, Spandan et al. (2022). “What makes domain generalization hard?” In: *arXiv preprint arXiv:2206.07802*.

Mahon, Bradford Z and Alfonso Caramazza (2009). “Concepts and categories: A cognitive neuropsychological perspective”. In: *Annual review of psychology* 60, pp. 27–51.

Margalit, Eshed et al. (2024). “A unifying framework for functional organization in early and higher ventral visual cortex”. In: *Neuron*.

Mavadati, S Mohammad et al. (2013). "Disfa: A spontaneous facial action intensity database". In: *IEEE Transactions on Affective Computing* 4.2, pp. 151–160.

McCarthy, Gregory et al. (1997). "Face-specific processing in the human fusiform gyrus". In: *Journal of cognitive neuroscience* 9.5, pp. 605–610.

Mende-Siedlecki, Peter, Yang Cai, and Alexander Todorov (2012). "The neural dynamics of updating person impressions". In: *Social cognitive and affective neuroscience* 8.6, pp. 623–631.

Minkowski, Mieczyslaw (1920). *Über den Verlauf, die Endigung und die zentrale Repräsentation von gekreuzten und ungekreuzten Sehnervenfasern bei einigen Säugetieren und beim Menschen*. O. Füssli.

Mishkin, Mortimer, Leslie G Ungerleider, and Kathleen A Macko (1983). "Object vision and spatial vision: two cortical pathways". In: *Trends in neurosciences* 6, pp. 414–417.

Mountcastle, VB et al. (1987). "Common and differential effects of attentive fixation on the excitability of parietal and prestriate (V4) cortical visual neurons in the macaque monkey". In: *Journal of Neuroscience* 7.7, pp. 2239–2255.

Mumford, Jeanette A et al. (2012). "Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses". In: *Neuroimage* 59.3, pp. 2636–2643.

Muschelli, John et al. (2014). "Reduction of motion-related artifacts in resting state fMRI using aCompCor". In: *Neuroimage* 96, pp. 22–35.

Nagrani, Arsha et al. (2020). "Voxceleb: Large-scale speaker verification in the wild". In: *Computer Speech & Language* 60, p. 101027.

Nath, Audrey R and Michael S Beauchamp (2012). "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion". In: *Neuroimage* 59.1, pp. 781–787.

Natu, Vaidehi S et al. (2010). "Dissociable neural patterns of facial identity across changes in viewpoint". In: *Journal of Cognitive Neuroscience* 22.7, pp. 1570–1582.

Nestor, Adrian, David C Plaut, and Marlene Behrmann (2011). "Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis". In: *Proceedings of the National Academy of Sciences* 108.24, pp. 9998–10003.

Nummenmaa, Lauri and Andrew J Calder (2009). "Neural mechanisms of social attention". In: *Trends in cognitive sciences* 13.3, pp. 135–143.

O'Toole, Alice J, Dana A Roark, and Hervé Abdi (2002). "Recognizing moving faces: A psychological and neural synthesis". In: *Trends in cognitive sciences* 6.6, pp. 261–266.

O'Toole, Alice J et al. (2011). "Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach". In: *Vision research* 51.1, pp. 74–83.

Parkhi, Omkar M, Andrea Vedaldi, Andrew Zisserman, et al. (2015). "Deep face recognition." In: *bmvc*. Vol. 1. 3, p. 6.

Paszke, Adam et al. (2017). "Automatic differentiation in PyTorch". In.

Peelen, Marius V, Anthony P Atkinson, and Patrik Vuilleumier (2010). "Supramodal representations of perceived emotions in the human brain". In: *Journal of Neuroscience* 30.30, pp. 10127–10134.

Peelen, Marius V, Alison J Wiggett, and Paul E Downing (2006). "Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion". In: *Neuron* 49.6, pp. 815–822.

Perenin, M-T and A Vighetto (1988). "Optic ataxia: A specific disruption in visuomotor mechanisms: I. Different aspects of the deficit in reaching for objects". In: *Brain* 111.3, pp. 643–674.

Pinsk, Mark A et al. (2009). "Neural representations of faces and body parts in macaque and human cortex: a comparative FMRI study". In: *Journal of neurophysiology* 101.5, pp. 2581–2600.

Pitcher, David, Bradley Duchaine, and Vincent Walsh (2014). "Combined TMS and fMRI reveal dissociable cortical pathways for dynamic and static face perception". In: *Current Biology* 24.17, pp. 2066–2070.

Pitcher, David and Leslie G Ungerleider (2020). "Evidence for a Third Visual Pathway Specialized for Social Perception". In: *Trends in Cognitive Sciences*.

— (2021). "Evidence for a third visual pathway specialized for social perception". In: *Trends in Cognitive Sciences* 25.2, pp. 100–110.

Pitcher, David et al. (2007). "TMS evidence for the involvement of the right occipital face area in early face processing". In: *Current Biology* 17.18, pp. 1568–1573.

Pitcher, David et al. (2011). "Differential selectivity for dynamic versus static information in face-selective cortical regions". In: *Neuroimage* 56.4, pp. 2356–2363.

Poggio, Tomaso and Shimon Edelman (1990). "A network that learns to recognize three-dimensional objects". In: *Nature* 343.6255, p. 263.

Popivanov, Ivo D et al. (2014). "Heterogeneous single-unit selectivity in an fMRI-defined body-selective patch". In: *Journal of Neuroscience* 34.1, pp. 95–111.

Posner, Michael I (1970). "Abstraction and the process of recognition". In: *Psychology of learning and motivation*. Vol. 3. Elsevier, pp. 43–100.

Raftery, Adrian E. (1995). "Bayesian Model Selection in Social Research". In: *Sociological Methodology* 25, pp. 111–163. ISSN: 00811750, 14679531. URL: http://www.jstor.org/stable/271063.

Ratan Murty, N Apurva et al. (2021). "Computational models of category-selective brain regions enable high-throughput tests of selectivity". In: *Nature communications* 12.1, pp. 1–14.

Rennig, Johannes and Michael S Beauchamp (2021). "Intelligibility of Audiovisual Sentences Drives Multivoxel Response Patterns in Human Superior Temporal Cortex". In: *NeuroImage*, p. 118796.

Rezlescu, Constantin et al. (2014). "Normal acquisition of expertise with greebles in two cases of acquired prosopagnosia". In: *Proceedings of the National Academy of Sciences* 111.14, pp. 5123–5128.

Rumiati, Raffaella I et al. (1994). "Visual object agnosia without prosopagnosia or alexia: Evidence for hierarchical theories of visual recognition". In: *Visual Cognition* 1.2-3, pp. 181–225.

Saxe, Andrew M, James L McClelland, and Surya Ganguli (2019). "A mathematical theory of semantic development in deep neural networks". In: *Proceedings of the National Academy of Sciences* 116.23, pp. 11537–11546.

Saxe, Rebecca and Sean Dae Houlihan (2017). "Formalizing emotion concepts within a Bayesian model of theory of mind". In: *Current opinion in Psychology* 17, pp. 15–21.

Schwartz, Emily et al. (2022). "Spontaneous Learning of Face Identity in Expression-Trained Deep Nets". In: Accepted to Cognitive Computational Neuroscience.

Schwartz, Emily et al. (2023a). "Challenging the Classical View: Recognition of Identity and Expression as Integrated Processes". In: *Brain Sciences* 13.2, p. 296.

Schwartz, Emily et al. (2023b). “Intracranial electroencephalography and deep neural networks reveal shared substrates for representations of face identity and expressions”. In: *Journal of Neuroscience* 43.23, pp. 4291–4303.

Selvaraju, Ramprasaath R et al. (2017). “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Sergent, Justine and Jean-Louis Signoret (1992). “Functional and anatomical decomposition of face processing: evidence from prosopagnosia and PET study of normal subjects”. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335.1273, pp. 55–62.

Sheinberg, David L and Nikos K Logothetis (1997). “The role of temporal cortical areas in perceptual organization”. In: *Proceedings of the National Academy of Sciences* 94.7, pp. 3408–3413.

Shipp, Stewart and Semir Zeki (1985). “Segregation of pathways leading from area V2 to areas V4 and V5 of macaque monkey visual cortex”. In: *Nature* 315.6017, pp. 322–324.

Shultz, Sarah et al. (2011). “The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions”. In: *Social cognitive and affective neuroscience* 6.5, pp. 602–611.

Sigala, Natasha and Nikos K Logothetis (2002). “Visual categorization shapes feature selectivity in the primate temporal cortex”. In: *Nature* 415.6869, pp. 318–320.

Silver, Michael A and Sabine Kastner (2009). “Topographic maps in human frontal and parietal cortex”. In: *Trends in cognitive sciences* 13.11, pp. 488–495.

Sinha, Pawan and Tomaso Poggio (1996). "Role of learning in three-dimensional form perception". In: *Nature* 384.6608, pp. 460–463.

Skerry, Amy E and Rebecca Saxe (2014). "A common neural code for perceived and inferred emotion". In: *Journal of Neuroscience* 34.48, pp. 15997–16008.

Steeves, Jennifer KE et al. (2006). "The fusiform face area is not sufficient for face recognition: evidence from a patient with dense prosopagnosia and no occipital face area". In: *Neuropsychologia* 44.4, pp. 594–609.

Tarnowski, Paweł et al. (2017). "Emotion recognition using facial expressions". In: *Procedia Computer Science* 108, pp. 1175–1184.

Thomas, Cibu et al. (2009). "Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia". In: *Nature neuroscience* 12.1, pp. 29–31.

Thornton, Chris (1996). "Re-presenting representation". In: *Forms of representation: An interdisciplinary theme for cognitive science*, pp. 152–62.

Tsao, Doris Y et al. (2006). "A cortical region consisting entirely of face-selective cells". In: *Science* 311.5761, pp. 670–674.

Ungerleider, Leslie G and Robert Desimone (1986). "Cortical connections of visual area MT in the macaque". In: *Journal of Comparative Neurology* 248.2, pp. 190–222.

Van Essen, David C, Charles H Anderson, and Daniel J Felleman (1992). "Information processing in the primate visual system: an integrated systems perspective". In: *Science* 255.5043, pp. 419–423.

Van Essen, David C et al. (2001). "Mapping visual cortex in monkeys and humans using surface-based atlases". In: *Vision research* 41.10-11, pp. 1359–1378.

Van Grootel, Tom J et al. (2017). "Development of visual cortical function in infant macaques: A BOLD fMRI study". In: *PloS one* 12.11, e0187942.

Vinken, Kasper, Talia Konkle, and Margaret Livingstone (2022). "The neural code for'face cells' is not face specific". In: *bioRxiv*.

Wagner, HL, CJ MacDonald, and AS Manstead (1986). "Communication of individual emotions by spontaneous facial expressions." In: *Journal of Personality and Social Psychology* 50.4, p. 737.

Wang, Aria, Michael Tarr, and Leila Wehbe (2019). "Neural taskonomy: Inferring the similarity of task-derived representations from brain activity". In: *Advances in Neural Information Processing Systems*, pp. 15501–15511.

Wang, Yin et al. (2017). "Dynamic neural architecture for social knowledge retrieval". In: *Proceedings of the National Academy of Sciences* 114.16, E3305–E3314.

Warrington, Elizabeth K and Rosaleen McCarthy (1983). "Category specific access dysphasia". In: *Brain* 106.4, pp. 859–878.

Warrington, Elizabeth K and Rosaleen A McCarthy (1987). "Categories of knowledge: Further fractionations and an attempted integration". In: *Brain* 110.5, pp. 1273–1296.

Warrington, Elizabeth K and Tim Shallice (1984). "Category specific semantic impairments". In: *Brain* 107.3, pp. 829–853.

Winston, Joel S et al. (2004). "fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception". In: *Journal of neurophysiology* 92.3, pp. 1830–1839.

Wu, Yang and Laura E Schulz (2018). "Inferring beliefs and desires from emotional reactions to anticipated and observed events". In: *Child development* 89.2, pp. 649–662.

Wurm, Moritz F and Alfonso Caramazza (2022). "Two 'what'pathways for action and object recognition". In: *Trends in cognitive sciences* 26.2, pp. 103–116.

Xu, Xiaokun and Irving Biederman (2010). "Loci of the release from fMRI adaptation for changes in facial expression, identity, and viewpoint". In: *Journal of Vision* 10.14, pp. 36–36.

Xu, Yaoda and Maryam Vaziri-Pashkam (2021). "Limits to visual representational correspondence between convolutional neural networks and the human brain". In: *Nature communications* 12.1, pp. 1–16.

Yamins, Daniel L et al. (2013). "Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream". In: *Advances in neural information processing systems*, pp. 3093–3101.

Yamins, Daniel LK and James J DiCarlo (2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3, p. 356.

Yang, Tianming and John HR Maunsell (2004). "The effect of perceptual learning on neuronal responses in monkey visual area V4". In: *Journal of Neuroscience* 24.7, pp. 1617–1626.

Yang, Yi and Shawn Newsam (2010). "Bag-of-visual-words and spatial extensions for land-use classification". In: *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279.

Yang, Zetian and Winrich A Freiwald (2021). "Joint encoding of facial identity, orientation, gaze, and expression in the middle dorsal face area". In: *Proceedings of the National Academy of Sciences* 118.33, e2108283118.

Yarkoni, Tal et al. (2011). "Large-scale automated synthesis of human functional neuroimaging data". In: *Nature methods* 8.8, pp. 665–670.

Yosinski, Jason et al. (2014). "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*, pp. 3320–3328.

Young, Andrew W et al. (1993). "Face perception after brain injury: Selective impairments affecting identity and expression". In: *Brain* 116.4, pp. 941–959.

Young, Andrew W et al. (2002). "Facial expressions of emotion: Stimuli and tests (FEEST)". In: *Bury St. Edmunds: Thames Valley Test Company*.

Yovel, Galit and Alice J O'Toole (2016). "Recognizing people in motion". In: *Trends in Cognitive Sciences* 20.5, pp. 383–395.

Zamir, Amir R et al. (2018). "Taskonomy: Disentangling task transfer learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722.

Zhou, Liqin, Ming Meng, and Ke Zhou (2021). "Emerged human-like facial expression representation in a deep convolutional neural network". In: *bioRxiv*.

Zhou, Liqin et al. (2022). "Emerged human-like facial expression representation in a deep convolutional neural network". In: *Science advances* 8.12, eabj4383.

Zhu, Yi et al. (2019). "Hidden two-stream convolutional networks for action recognition". In: *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, pp. 363–378.

Zhuang, Chengxu et al. (2021). "Unsupervised neural network models of the ventral visual stream". In: *Proceedings of the National Academy of Sciences* 118.3, e2014196118.