Investigation and Modelling of Dynamical Facial Expression Perception

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. Michael Stettler

aus Aarau/Schweiz

Tübingen

2025

Investigation and Modelling of Dynamical Facial Expression Perception

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. Michael Stettler

aus Aarau/Schweiz

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	14.03.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Martin A. Giese
2. Berichterstatter:	Prof. Dr. Martin Butz

For
Marie, Isaac & Jeremy

Abstract

Facial expressions play a crucial role in daily human social interactions. Humans exhibit a remarkably strong and innate ability to perceive both novel faces and non-human head shapes, such as emojis and cartoons. However, the visual mechanisms involved in perceiving dynamic facial expressions remain largely unclear. This thesis aims to unravel the processes by which our visual system encodes dynamic facial expressions.

Employing computer graphics, we have crafted stimuli to delve into human behavioral perception through psychophysics experiments. Concurrently, we collect electrophysiological recordings from rhesus macaques. Leveraging these data sets, we design, test, and validate our models. Our innovative framework, based on norm-referenced encoding—a mechanism initially proposed for encoding facial identity—proves to be a valid mechanism for encoding facial expressions.

Expanding norm-referenced encoding into multi-domain applications, we discover that utilizing an updatable reference frame allows us to split the learning procedure, making it a data-efficient mechanism. This enables the transfer of facial expressions across basic face shapes with only a single training data point. Extending our model to larger datasets, we demonstrate its robustness and generalization capabilities. Notably, norm-referenced encoding yields perceptual results closer to human perception compared to other computer vision models.

In summary, our research sheds light on a mechanism well-known in neuroscience but relatively unexplored in computer vision. We demonstrate its potential application in developing computer systems that interact and understand human facial expressions more similarly to current techniques. Norm-referenced encoding holds promise for the advancement of computer systems in this domain.

Kurzfassung

Gesichtsausdrücke spielen eine entscheidende Rolle in der sozialen Interaktion zwischen Menschen im täglichen Leben. Die menschliche Wahrnehmung ist dabei außergewöhnlich robust bei der Vearbeitung unbekannter Gesichter, und sogar gegenüber nicht-menschlichen Kopfformen (z. B. Emojis, Cartoons). Der visuelle Mechanismus, der zur Wahrnehmung dynamischer Gesichtsausdrücke führt, ist weitgehend unklar. In dieser Arbeit zielen wir darauf ab, zu verstehen, wie unser visuelles System dynamische Gesichtsausdrücke codiert. Unter Verwendung von Methoden aus der Computergrafik haben wir Stimuli entwickelt, um die menschliche Verhaltenswahrnehmung durch psychophysische Experimente zu untersuchen, und gleichzeitig elektrophysiologische Aufzeichnungen von Rhesusaffen gesammelt. Wir nutzen diese Daten, um neue Modelle zu entwerfen, zu testen und zu validieren. In diesem Rahmen haben wir eine neuartige Architektur entwickelt, die auf der Normreferenzcodierung basiert. Diese wurde ursprünglich zur Modellierung von Gesichtsidentität vorgeschlagen und beruht auf der Annahme dass Gesichter als richtungsspezifische Abweichung eines Referenzgesichts repräsentiert sind. Zunächst zeigen wir, dass die Normreferenzcodierung ein gültiger Mechanismus zur Codierung von Gesichtsausdrücken ist. Wir erweitern dann die Normreferenzcodierung auf mehrere Referenzrahmen und erkennen, dass das Lernverfahren in diesem Fall in zwei Teile aufgeteilt werden kann. Diese Multi-Domänen- Normreferenzcodierung stellt einen dateneffizienten Mechanismus dar, der die Übertragung von Gesichtsausdrücken über grundlegend unterschiedliche Gesichtsformen mit nur einem Datenpunkt ermöglicht. Wir erweitern unser Modell, um es auf größeren Datensätzen zu testen, und zeigen, dass der Mechanismus auch dann robust ist und generalisiert. Schließlich stellen wir fest, dass die Normreferenzcodierung im Vergleich zu anderen Computer-Vision-Modellen eine größere Ähnlichkeit zur Wahrnehmung unserer menschlichen Probanden aufweist. Insgesamt beleuchtet unsere Arbeit einen in der Neurowissenschaft bekannten Mechanismus,der jedoch bisher in der Computer Vision wenig erforscht wurde, und zeigt, wie er in diesem Kontext genutzt werden kann. Die Normreferenzcodierung hat das Potenzial, Computersysteme zu entwickeln, deren Wahrnehmung von Gesichtsausdrücken der menschlichen Wahrnehmung ähnlicher ist als die aktueller Techniken.

Acknowledgments

Thank you, Martin, for accepting me into your lab. It has been a great journey, and I've learned a lot. We faced ups and downs, especially with the COVID-19 pandemic hitting right in the middle of our psychophysics experiments, but I am grateful for the opportunity and will always appreciate your fantastic support and personal guidance.

I am also thankful for meeting amazing people in the office, particularly Nick, who shared his passion for creating realistic avatars. Jesse, who delighted me with many discussions on machine learning and general intelligence, ans who was always quick to point out my English mistakes to win his argument. Alexander, who probably still struggles to understand how a bio-engineer like myself doesn't require a rigorous mathematical proof to start coding, but who helped me to frame the mathematical part of my work. I'm also grateful to the rest of the people I've worked with in this group: Jindra, Bjorn, Jens, Annika, Jana, Lucas, Prerana, Albert, Winfried, Tim, and Tahareh. All of you contributed to this thesis, making my working days more colorful.

A special thank you goes to Julius and Luigi, whose impressive insights allowed me to become a much better scientist. I also want to specifically say thank you to Ramona, my closest collaborator outside of the lab, whom I admire for the hard work she is doing on gathering the data with her monkeys.

I also express gratitude to Marie, who supported me throughout this time, without whom all this work would not have been possible. A big thank you to her for giving me the best present in the world, a beautiful son. Seeing him grow up only makes me feel that much has yet to be discovered to create human-like intelligence. Nonetheless, his smiles and laughter have cheered me up every day, and I can't wait to see what the future will bring.

Furthermore, I want to extend a big thank you to my friends and family; I believe they all played a part in this thesis as well. They enabled me to sharpen my descriptive skills to present my thesis in a minute and always asked the difficult question: Who is interested in that? Most of all, thanks to Floran, who even proposed to change my thesis subject to investigate "des arrosoirs pour plantes aquatiques" (watering cans for aquatic plants). Hopefully, my thesis would be more interesting to the readers.

Finally, I want to express my gratitude to all the funding agencies without whom this work would not have been possible.

Contents

Chapter 1

List Of Publications

1.1 Accepted

Manuscript 1 (shared first author):
Taubert, N., Stettler, M., Siebert, R., Spadacenta, S., Sting, L., Dicke, P., Thier, P. & Giese, M. A. (2021). Shape-invariant encoding of dynamic primate facial expressions in human perception. Elife, 10, e61197.

Manuscript 2:
Stettler, M., Taubert, N., Azizpour, T., Siebert, R., Spadacenta, S., Dicke, P., Thier, P. & Giese, M. A. (2020, September). Physiologically-Inspired Neural Circuits for the Recognition of Dynamic Faces. In International Conference on Artificial Neural Networks (pp. 168-179). Springer, Cham.

1.2 Submitted

Manuscript 3:
Stettler, M., Lappe. A., Taubert, N., Giese, M. A. Multi-Domain Norm-referenced Encoding Enables Data Efficient Transfer Learning of Facial Expressions. WACV 2025.

1.3 Manuscripts ready for submission

Manuscript 4:
Stettler, M., Lappe, A., Siebert, R., Thier, P. & Giese. Norm-reference Encoding Explains Intra-Cortical Recording and Behavioural Results of Dynamic Facial Expression Perception. ELife.

Chapter 2

Introduction

2.1 Motivation

How does the brain encode dynamic facial expression perception?

Facial expressions play a crucial role in human social interactions. We are one of the most socially sophisticated species on the planet Wilson *et al.* (2012). Our face is an essential tool to communicate our emotions, allowing others to adapt their behaviour. Since the pioneering work of Darwin Darwin and Prodger (1998), the question of facial expression recognition (FER) has remained central to fields such as psychology, philosophy, anthropology, computer vision, and social robotics Ekman (1992); Matsumoto *et al.* (2008); Jack *et al.* (2016). Many scientists have developed models to explain and derive new assumptions that advance our understanding of facial-expression perception.

Traditionally, neuroscientists have developed neural networks that often use controllable and semantically meaningful multidimensional space, e.g. Reed (1972); Goldman and Homa (1977); Medin and Schaffer (1978); Neumann (1977); Nosofsky (1991). These models assume that we can decompose the brain into pieces and manipulate these pieces in isolation. This assumption has one significant benefit: Reducing the task complexity makes interpreting and testing the model's hypothesis easier. However, many studies regard these models as too simple Valentine (2005); Felsen and Dan (2005); Olshausen and Field (2005); Hasson and Honey (2012).

On the other side of the spectrum, computer scientists developing FER models have embraced complexity. These recent techniques reach human-level behavioural performance using real-life stimuli. These highly engineered and optimized models show higher robustness to real-life examples trained on big data Mollahosseini *et al.* (2017); Li and Deng (2020); Ngo and Yoon (2020). However, recent FER models are typically developed as a sub-task from the more general object classification task in computer vision. Therefore, recent FER models based on artificial neural networks (ANN) follow the broader machine-learning advances. As such, FER models aim at getting better performance scores over several benchmarks with only marginal interest in the interpretability of the feature space. (See Serre (2019) for a review on deep learning.) Nonetheless, due to the impressive results achieved by these recent architectures, many questioned how similar these models are to the brain Lake *et al.* (2017); Marcus (2018); Chang

et al. (2021) with some showing stark differences between deep learning-based model and human perception Geirhos *et al.* (2018b,a); Baker *et al.* (2020).

Therefore, after decades of studies in facial expression recognition we have two schools of thought: on the one hand we have many theories of neurological frameworks that are not applicable to large datasets. On the other hand, we have many (over-)parameterized models. While both approaches have contributed valuable insights, they each have limitations that hinder our understanding of how the brain encodes dynamic facial expression perception. As a result, the underlying mechanism in the brain for perceiving dynamic facial expressions remains largely unclear. Moving forward, a more integrated and interdisciplinary approach that combines the strengths of both traditional and modern techniques may be necessary to address this fundamental question. This thesis is an attempt to bridge this gap.

2.2 Objectives of the Doctoral Research

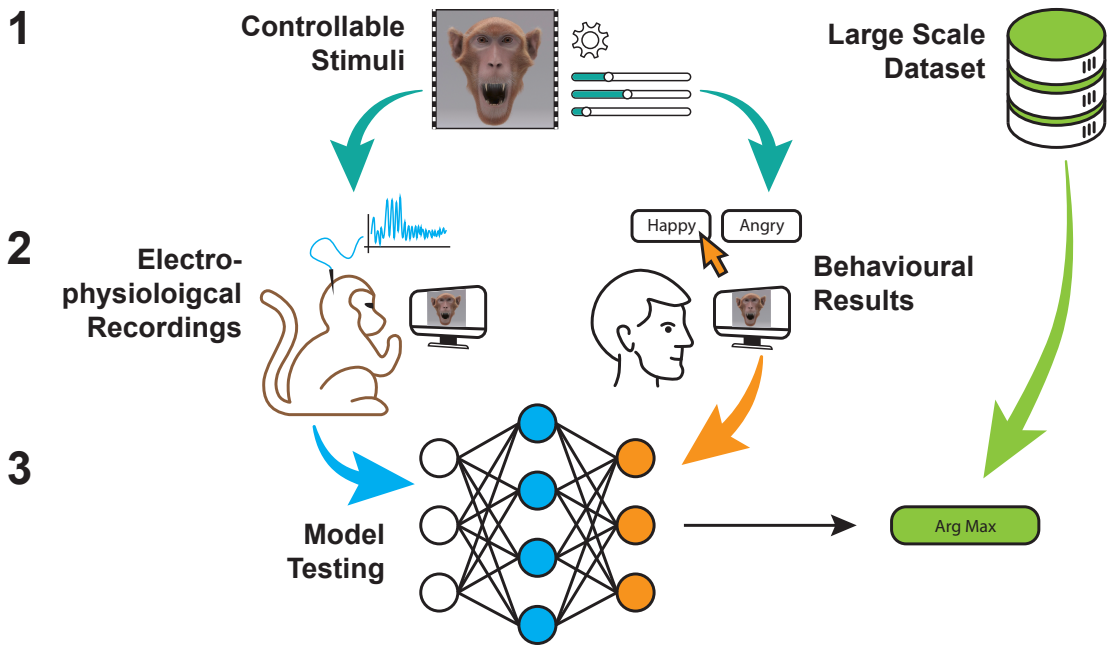

Figure 2.1: Objective of the project. 1) Stimuli level: We develop and validate controllable dynamic stimuli to investigate the perception of dynamic facial expression, and, we selected a large scale publicly available dataset. 2) Data acquisition level: We record electrophysiological data from Rhesus Macaque and behavioural data from Humans from our developed stimuli. 3) Model level: We develop a model that correlate electrophysiological recordings, reproduce behavioural data and accomplish data efficient image classification on a publicly available facial expression dataset.

Our visual system has a remarkable property: We can innately recognize facial expressions on non-human faces, such as those of cartoon characters. There exists a significant body of work on facial perception within a variety of domains that illustrates this property. We can recognise facial expression from caricature drawings Rhodes *et al.* (1987); Gao *et al.* (2003), emoticons/emojis Kaye *et al.* (2017), cartoon characters Zhao *et al.* (2019); Zhang *et al.* (2021), as well as studies on the phenomenon of pareidolia (the perception of faces on arbitrary objects) Wardle *et al.* (2020).

We echo the viewpoints of others Lake *et al.* (2017); Hassabis *et al.* (2017); Zador *et al.* (2022) who assert the importance of integrating principles from computer science, cognitive science, and neuroscience to advance towards achieving human-like intelligence. We hold the belief that the brain, with its many yet-to-be-clarified key characteristics, serves as a rich source of inspiration. It is our firm conviction that understanding these mechanisms is crucial for the development of more human-like intelligence. Consequently, this dissertation aims to investigate a specific mechanism—the norm-referenced encoding (see §3)—and to develop biologically plausible models to explore the benefits of this mechanism. Our objectives are threefold: replicating human behavioral results, excelling in computer vision tasks, and providing a framework for comparison with neuronal activities.

Behaviour Our primary objective is to ensure that our model accurately replicates behavioral results. To achieve this, we have meticulously designed psychophysics experiments focused on testing human visual perception. The outcomes of these experiments serve as the foundational framework for our model's design. Our ultimate aim is to provide a model that perceives facial expressions with greater fidelity compared to existing Facial Expression Recognition (FER) models.

Computer Vision One significant concern with neuro-scientific models lies in their tendency to oversimplify real-world scenarios, often leading to shortcomings when put to the test. To address this concern, we followed closely advances in machine learning to ensuring that our model can scale up effectively. To accomplish this, we use deep learning framework to enable our model to perform well on publicly available datasets. Our goal is to demonstrate the model's scalability and robustness beyond controlled settings.

Neuroscience Our approach is firmly rooted in staying closely aligned with experimentalists in the field of neuroscience. To achieve this alignment, we used only well-studied components from neuroscience into our model's architecture. This deliberate choice allows for direct comparisons between our model and electrophysiological recordings, which represent the most reliable method for confirming or refuting potential brain mechanisms.

These principles are the main motivation behind our models' design and our contribution to connect more closely neuro-scientific and machine learning approaches. By succeeding is those three goals, we provide a robust and biologically plausible model of facial expression recognition that can serve as a valuable tool for advancing our understanding of the neural basis of this fundamental aspect of human communication. Figure 2.1 shows an overview of the objectives.

2.3 Overview of My Thesis

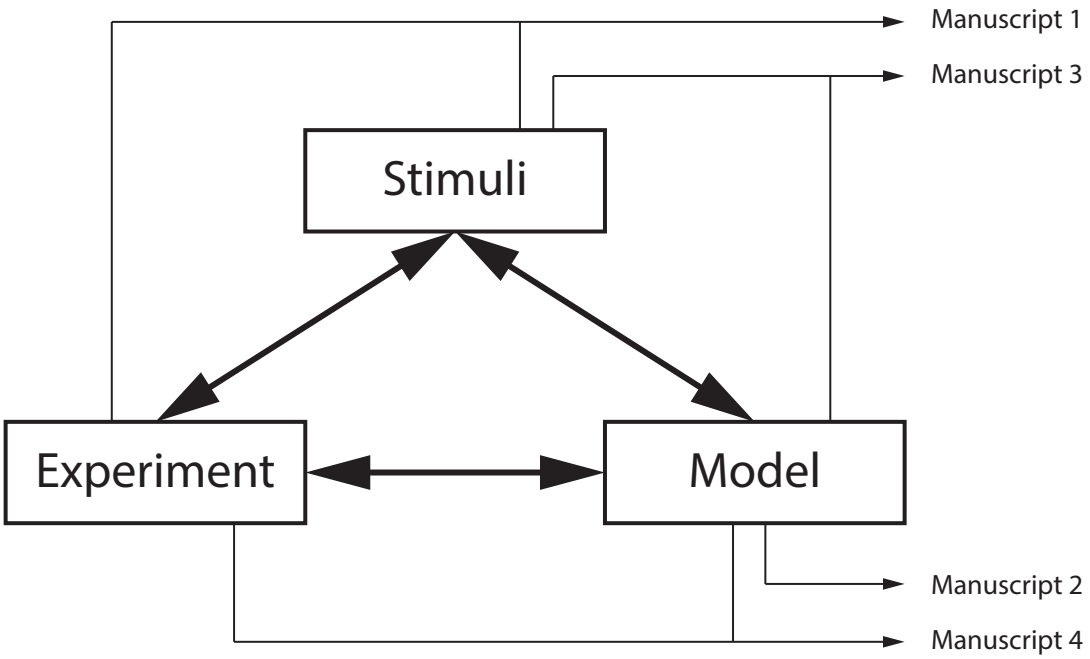

Figure 2.2: Overview of the different part of the project and their related manuscripts.

The first step of this project was the creation of highly realistic and controllable set of dynamic facial expressions to gather data. These stimuli are currently used by our collaborator to record intra-cortical neuronal responses in the STS (superior temporal sulcus) of Rhesus Macaque. Our collaborators showed that our stimuli were on the right side of the uncanny valley Siebert *et al.* (2020). This close collaboration with neuro-recording promises to test the assumption of our models. Nonetheless, it also make my progress dependent of their success, which is a relatively high risk as it involves monkey recording. Therefore, to lower my dependency on others, I designed a psychophysics experiment to investigate humans perception on cross-species facial expression perception Taubert *et al.* (2021) (Manuscript 1 in Figure 2.2). Having my own psychophysic experiment allowed me gain control over my own work. Equipped with my behavioural results, I started to develop my first model to investigate the difference between common computer vision models and our model. The main assumption behind our model is the use of the norm-referenced encoding Leopold *et al.* (2006); Giese and Leopold (2005) which I will introduce in more details §3. My first success was to show that this encoding, first proposed for facial identity recognition, is also a valid mechanism for facial expression recognition (FER) Stettler *et al.* (2020) (Manuscript 2 in Figure 2.2). Then, I updated my model to scale it up towards publicly available FER dataset and showed how

the mechanism shows strong robustness and transfer learning capabilities Stettler *et al.* (2023a) (Manuscript 3 in Figure 2.2). Finally, I used my model to reproduced the behavioural results from my psychophysic experiment and discuss its architecture in light of neuroscience frameworks Stettler *et al.* (2023b) (Manuscript 4 in Figure 2.2).

2.4 Thesis Structure

The layout of this dissertation is set up to tell a connected story, not necessarily following the exact order of when the papers were published, but rather reflecting the flow of my research journey. Since there are multiple papers tied to the same project, I've organized the results in a way that fits the narrative of my progress. In rearranging the content from my papers, I only present the work that I personally did. For any parts that I introduce but didn't handle myself, I'll mention my collaborators.

The thesis will commence by introducing the norm-referenced mechanism 3. Subsequently, I will provide a comprehensive, detailed examination of the developed models 4, presenting each component's evolution step by step. This presentation approach serves to highlight the challenges I addressed throughout my thesis, challenges that may not be readily apparent in the published papers. Moreover, I'll explore the link between my models within recent neuroscience frameworks 5 and computer vision models 6

After explaining my models, I'll discuss the datasets used and created during my project 7. These datasets provides a rational on the experiments I conducted. Next, I'll present the results 8. In this section, I'll follow my manuscripts, providing comments and a brief discussion of each result and its connection to my thesis objectives and model designs. I will finish my results section with the current status and limitation of my models 8.5.

Finally, the thesis will conclude with a comprehensive general discussion 9. I hope that you will find this dissertation an enlightening read.

Chapter 3

Norm-Reference Encoding

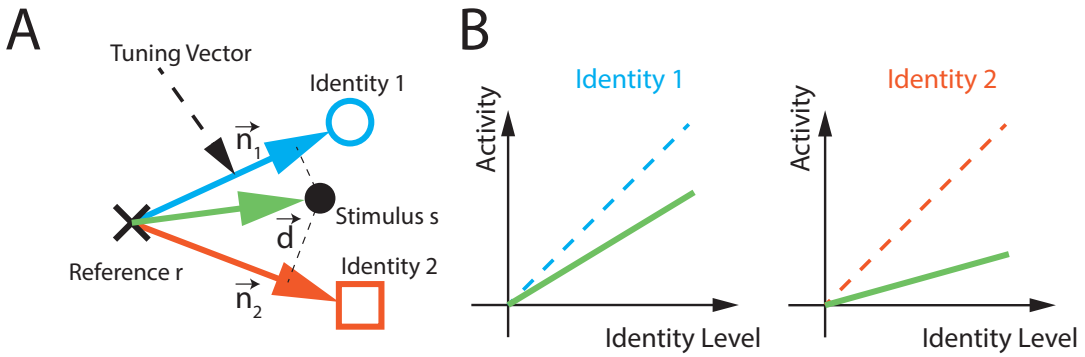

Figure 3.1: **A)** A schematic representation of norm-based encoding. The sketch displays two identities in feature space and their respective tuning vectors $\mathbf{n}_1$ and $textbfn_2$. The stimulus $\mathbf{s}$ is encoded by its position relative to the reference $\mathbf{r}$ through the difference vector $\mathbf{d} = \mathbf{s} - \mathbf{r}$. **B)** Read-out activity v_i for the input stimulus $\mathbf{s}$ for the two identities. The activity is given by $v_i = \mathbf{d}^T \mathbf{n}_i$ and thus grows linearly in the identity level $\|\mathbf{d}\|$. The slope is given by $\mathbf{d}^T \mathbf{n}_i / \|\mathbf{d}\|$, meaning $\mathbf{d} = \mathbf{n}_i$ yields the identity function.

In this chapter, I will explain the key mechanism investigated during my thesis, the so-called norm-referenced encoding (NRE). Norm-referenced encoding is a classical principle in neuroscience for face identity representation. The main idea of NRE is that face identities are encoded by difference vectors relative to a norm stimulus, typically the average face Valentine *et al.* (2001). The direction of this difference encodes facial identity, while its length represents the distinctiveness of the face relative to the average face (Figure 3.1A).

The encoding principle under consideration is supported by electrophysiological evidence Leopold *et al.* (2006); Koyano *et al.* (2021), and a related neural model has been proposed to replicate these findings Giese and Leopold (2005). Electrophysiological recordings have identified *face-encoding* neurons specifically tuned to encode facial identity relative to a norm (average) face. Notably, these *face-encoding* neurons exhibit a delayed response, emerging approximately 100 ms after stimulus presentation Koyano

et al. (2021). These findings suggest a dual process: a rapid one employing encoding in absolute space Chang and Tsao (2017), and a slower one realizing relative encoding to the average face, potentially facilitated by recurrent feedback Freiwald and Hosoya (2021).

Mathematically, we can model this mechanism as follow: Let $\mathbf{s}$ be a vector that represents a face stimulus in an appropriate feature space. Then the difference vector is defined by $\mathbf{d} = \mathbf{s} - \mathbf{r}$, where $\mathbf{r}$ is a norm or reference vector. Classically, the reference vector is the *average face* computed by averaging the feature vectors of a large number of faces.

The response of a *face-encoding* neuron v is computed as the scalar product of the difference vector $\mathbf{d}$ of the actual stimulus and a unit vector $\mathbf{n}_i$ (tuning vector) that determines the preference of the neuron in terms of face identity:

$$v_i = \mathbf{d}^T \mathbf{n}_i. \tag{1}$$

The neuron would respond maximally to face stimuli for which the associated difference vector $\mathbf{d}$ is collinear with the vector $\mathbf{n}_i$. The length of the difference vector $\|\mathbf{d}\|$ encodes how characteristic the face is for a specific facial identity (called 'identity level' in psychophysics) (See Fig. 3.1B). As such, NRE is an intuitive model to encode facial identities, where each *tuning vector* represents an identity in face space, and moving along their directions increase the strength of the distinctiveness of the identity.

While the mechanism is convenient for psychologist, it remains unclear whether this mechanism can be used to encode other task and what might be its benefit compared to more traditional approach (i.e using an absolute encoding in a face space instead of a relative).

These lead us to the two main questions behind my thesis: Could we transfer this mechanism to other tasks (i.e object recognition, facial expression recognition, etc.)? And, what are the benefit of using a relative encoding?

I answer the first question in my initial paper Stettler *et al.* (2020), where I demonstrate that the norm-referenced mechanism can effectively encode facial expressions recognition (FER). In its simplest form, this is achievable by swapping the tuning vectors from encoding identities to expressions types (i.e. angry, fear etc.) and changing the reference frame from the average face in FR to the neutral expression in FER. I will delve deeper into these findings in section, §8.2. However, to maintain the clarity of the argument, I will assume that facial expression recognition using NRE is a given result and continue this chapter by presenting my contribution to the theoretical framework of norm-referenced encoding, which highlights a key advantages of relative encoding.

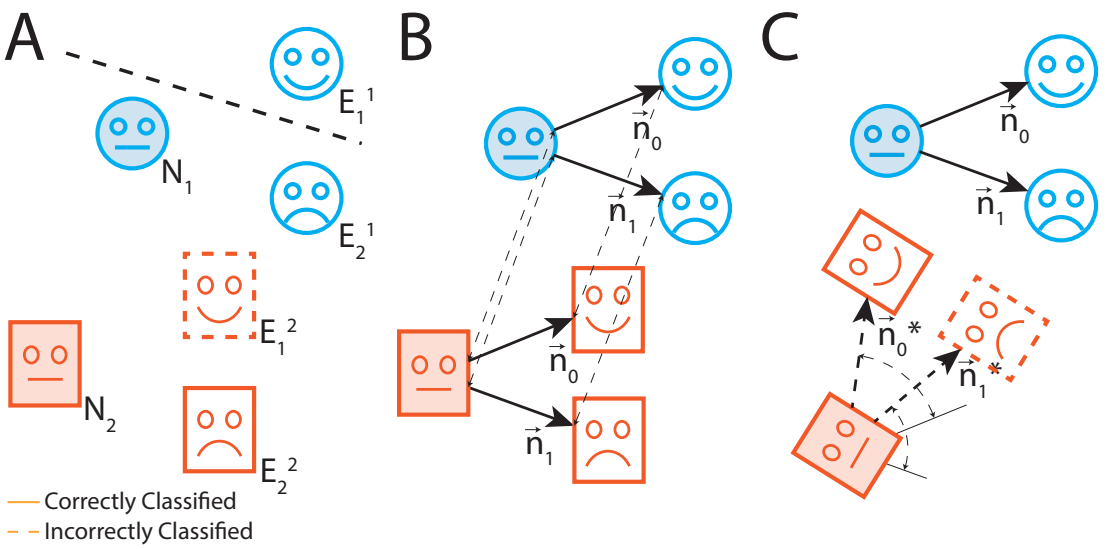

Figure 3.2: Schematic representation of the classification of two expressions (E_1 and E_2) in a 2D face space, presented on two different basic head shapes. N_1 and N_2 represent the *norm faces* (neutral expressions) relative to which the individual expressions are encoded. **A)** A linear classifier class boundary indicated by the dashed line results in the correct classification of the expressions of the first head shape (N_1), but fails on the second head shape (N_2). **B)** A norm-referenced classifier, which transfers the tuning vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ from the first head shape N_1 to the second head shape N_2 accomplishes correct classification without retraining the classifying neurons. **C)** An example of a norm-referenced classifier in a poorly-chosen feature space, resulting in misclassification for the second head shape.

3.1 Multi-Domain Norm-Referenced Encoding

As previously mentioned, Norm-referenced encoding (NRE) is an intuitive model to encode facial recognition (FR) and facial expression recognition (FER). Yet as the mechanism depends on a learned norm-referenced, it makes it highly task-specific.

To address this shortcoming, we propose to extend NRE to multiple reference frames, referring to it as *multi-domain norm-referenced encoding* (MD-NRE). We hypothesize that utilizing multiple norm-references might be a highly data-efficient transfer-learning approach for multi-domain FER. In a FER task, we consider domains to be different species (humans, monkeys, cartoons) and we will refer to identities as a within domain difference. As such, within different domain not only the face shape might be highly different (humans vs. cartoons) but also the skin textures display higher variance (i.e. skin vs. fur).

Consider the problem of classifying facial expressions E that correspond to feature vectors $\mathbf{s}$. Let N denote the domain (head shape) from which $\mathbf{s}$ was drawn. For each

domain, we consider a domain-specific reference vector $\mathbf{r}_N$. Finding $\mathbf{r}_N$ is part of the learning procedure. The MD-NRE assumes that the distribution of E only depends on the domain N through the difference vector between the feature vector $\mathbf{s}$ and the domain-specific reference. More specifically,

$$\begin{aligned} p(E|\mathbf{s}) &= \sum_N p(E|\mathbf{s},N)p(N|\mathbf{s}) \\ &= \sum_N p(E|\mathbf{s}-\mathbf{r_N})p(N|\mathbf{s}). \end{aligned} \tag{2}$$

This assumption breaks the learning process down into two parts: learning to infer the domain from the feature vector, and learning to infer the expression from the difference vector $\mathbf{d} = \mathbf{s} - \mathbf{r}_N$. This decomposition allows us to learn the distribution of E for all domains while only training on data from one domain.

Fig. 3.2 illustrates the situation. In a two-dimensional feature space, we aim to classify the expressions E_1 and E_2 for two different head shapes. The neutral expressions for the two head shapes are denoted as N_1 and N_2. Panel A shows the situation using a linear classifier that separates the two expressions correctly for only one head shape. Suppose the effect of changing the basic head shape is a collinear translation of all feature vectors in the feature space. Even if the translation vector is the same for all tested expressions, the classifier fails to classify the two expressions correctly for the second head shape. However, this is different when MD-NRE is applied. As illustrated in panel B, assume that the directions of the difference vectors between the expressions (E_1^1, E_2^1 and E_1^2, E_2^2) and the corresponding neutral reference face vectors (N_1 or N_2) are the same for the two head shapes. Then MD-NRE, which uses the same tuning vectors $\mathbf{n}_1$ and $\mathbf{n}_2$ for the two head shapes, would show perfect transfer of expression recognition from the first head shape to the second. Therefore, training the system on the first head shape and knowing the feature vector of the second neutral (norm) face (N_2) allows a correct classification of the expressions on the second head shape without the need for explicit training of all expressions for the second head shape. Therefore, if done correctly, the MD-NRE promises great data efficiency in generalizing to new domains. In addition, the length of the difference vectors covaries directly with the expression strength, which is useful for many technical applications that exploit expressiveness.

Of course, the previous assumption that the tuning vectors $\mathbf{n}_0$ and $\mathbf{n}_1$ remain similar for different head shapes is far from trivial to achieve. It depends critically on the embedding of images in the 2D feature space. Panel C shows a hypothetical example where the tuning vectors change direction. Here the expressions for the first head shape do not help the classification of the expressions for the second. Such misalignment of the corresponding tuning vectors will be particularly apparent if the embedding feature space contains many dimensions unrelated to the expression classification task. Using, for example, texture-sensitive features, one would expect strong and unsystematic changes in the tuning vector moving from a human to a non-human head, which likely outweighs

the expression-specific changes.

3.1.1 NRE Classifier

The original NRE classifier is inspired by results on face-selective neurons in the visual cortex Giese and Leopold (2005); Leopold *et al.* (2006). It assumes that a norm vector $\mathbf{r}$ has been learned for each individual types of head shapes. The activity v_m of the detector neuron for expression m. Originally, the norm-referenced was proposed to encode facial identity using multidimensional vectors. But, as I will present in 4.4, I decided to encode facial expression using a set of 2D landmarks l. As such, I propose to sum the displacement over the different landmarks. Our NRE classifier is given by the function

$$v_m = \sum_l \left[\mathbf{d}_l^T \mathbf{n}_{l,m}\right]_+ . \tag{3}$$

Here $\mathbf{d}_l = \mathbf{s}_l - \mathbf{r}_l$ is the two-dimensional difference vector for the l-th face fragment between the input $\mathbf{s}$ feature vector and the reference $\mathbf{r}$ vector. The term in the bracket measures its similarity to the learned direction-tuning unit vector $\mathbf{n}_{l,m}$. The output v_m then depends linearly on the length of the difference vector, which varies monotonically with expression strength. This type of tuning matches the tuning function of neurons in the IT cortex for identity coding Leopold *et al.* (2006). We obtain the final classification result by determining the norm-referenced neural unit with the maximum output $\hat{m} = \arg\max_m v_m$.

Chapter 4

Models

In this chapter, I will describe how I developed the models using an iterative, step-by-step approach that emphasizes the different development phase of my models. The goal of this chapter is to provide a detailed explanation of the development process for the final model by explaining the rationale behind each component and how it solved a particular problem.

Figure 4.1 describes the main iteration process of the developed model trough time. The Figure review the first published model Stettler *et al.* (2020) (A) to the latest submitted model (E). Each model have a similar core architecture which consists of three main components: A feature extraction module, a feature reduction module that selects task-relevant features, and lastly, the mechanism of interest in this thesis, a readout level consisting of our norm-referenced encoding highlighted by the blue, green and red color respectively. All the presented models uses static images and I will discuss the dynamic part of the model in the section 4.7.

4.1 Model Motivation

In our study on dynamic cross-species facial expression categorization Taubert *et al.* (2021), two pivotal findings have emerged, which I will succinctly summarize here and delve into further detail in Section §8.1. Firstly, our investigation revealed that human perception of dynamic facial expressions is significantly independent of facial shapes. Secondly, participants showcased a remarkable capacity for rapid learning. These insights, crucial to the first objective of my thesis, serve as the guiding principles for the design of my models.

I elaborate on how we can harness norm-referenced encoding with multi-domain norm-referenced encoding (MD-NRE) in Section §3.1 to achieve rapid learning and facial expression recognition, irrespective of facial shape, through strategic transfer learning. Therefore, the primary challenge for my model lies in constructing a latent space that facilitates the transfer of tuning vectors responsible for encoding facial expressions. Consequently, my journey in developing my model architectures led to the construction of an efficient representation of facial features implementing MD-NRE.

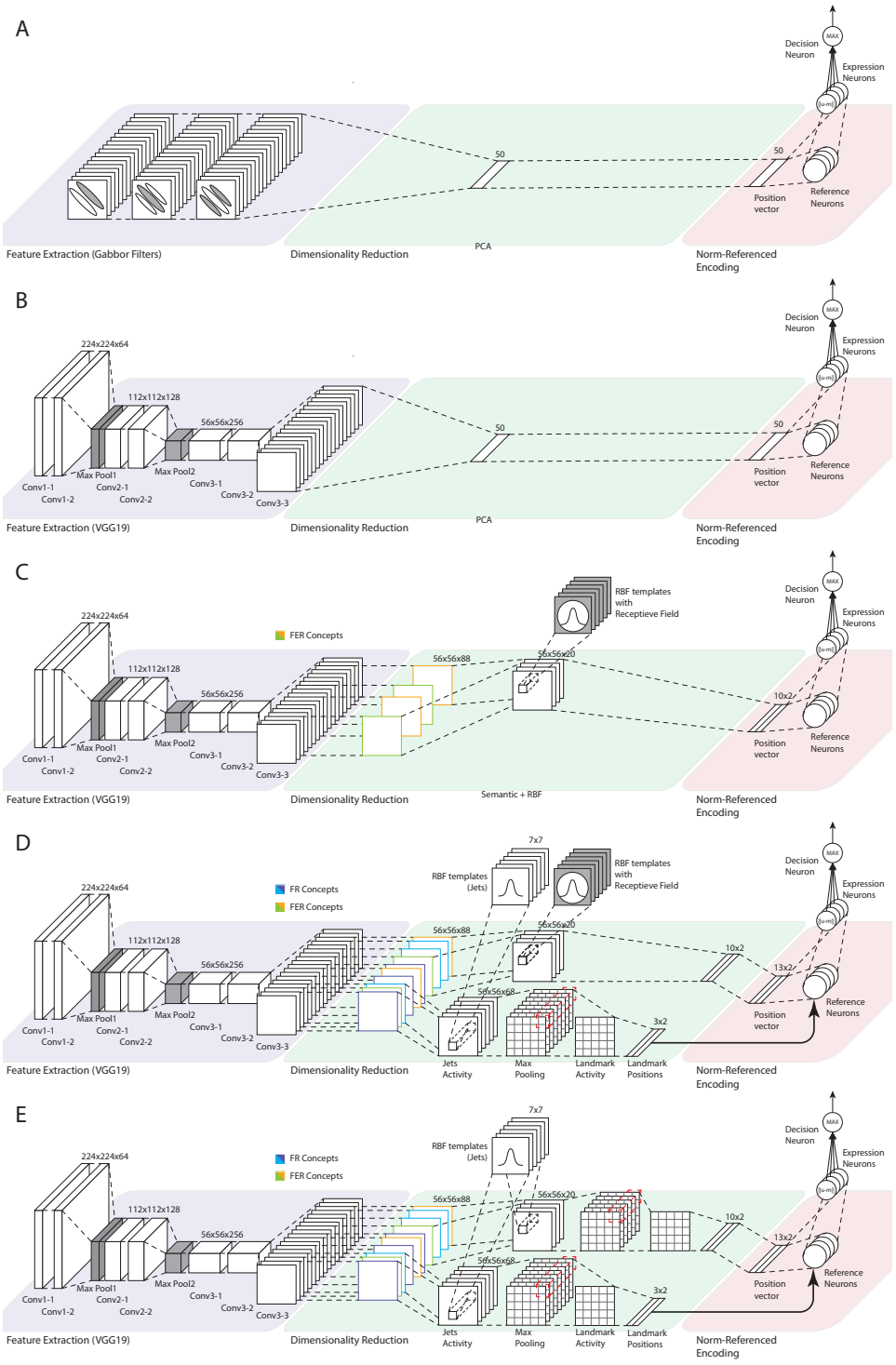

Figure 4.1: Development progress of the norm-referenced encoding-based model throughout my thesis. Panels A to E highlight the main changes in each version, from the original model based on Giese and Leopold (2005) to the final iteration in E.

However, this process involved several iterations, which I will now elucidate. All my models consist of a general architecture made from three main components:

- A feature extraction component, which serves as a general and robust facial feature extractor. Specifically, this component aims at providing an invariant encoding in terms of scale, lighting and textures.
- A feature reduction component, which is implemented to encode facial features that facilitate norm-referenced encoding through knowledge transfer.
- A read out layer, which is the norm-referenced encoding mechanism.

In the following sections, I will explain how these components have been modified from the original model (see Figure 4.1A), inspired by the work of Giese & Leopold Giese and Leopold (2005), to my latest proposed model E (Figure 4.1E).

4.2 Model A (Original)

Model A draws significant inspiration from the original norm-referenced encoding model Giese and Leopold (2005) and serves as the foundation for my initial publication Stettler *et al.* (2020) (Manuscript 2). As my simplest model, it laid the groundwork for subsequent models. Initially, it was unclear if NRE could be used to encode facial expression recognition. Therefore, to achieve my thesis objective, I had to explore and validate the feasibility of norm-referenced encoding within a facial expression recognition (FER) classification task. Results from this model are put in Section §8.2.

This model already consists of three main components outlined in the general architecture. We use three layers of Gabor filters in our feature extraction component to extract mid-level features from input images. These mid-level features bear similarity to neurons in area V4 of the macaque visual cortex Giese and Leopold (2005). These simplified layers proved sufficient for our first set of stimuli (see Section §7.1.2) as this initial implementation aimed to concentrate on higher-level facial expression-selective circuits.

Specifically, the face region in our stimulus movies was down-sampled to 200 x 200 pixels and converted to grayscale for further processing. The first layer includes even and uneven Gabor filters with eight orientations and three spatial scales differing by a factor of two. A constant was subtracted from these filter functions to ensure they were mean-free. Filter responses were computed on rectangular grids with 49, 69, and 97 points evenly spaced along the sides of the image region. This layer models orientation-selective neurons, akin to V1 simple cells Jones and Palmer (1987).

After these three layers of Gabor filters, the next layer models V1 complex cells and introduces partial position and spatial phase invariance. Using a maximum operation, we pool the responses of even and uneven Gabor filters with the same orientation preference

within a spatial region (receptive field). The receptive fields of these pooling neurons comprised three neurons in the previous layer (one for the largest spatial scale). The receptive field centres of the pooling neurons were positioned in quadratic grids with 15, 10, and 14 grid points for the three spatial scales.

Once the feature extraction is complete, we enter our feature reduction module and extract informative mid-level features by combining a simple heuristic feature selection algorithm with Principal Components Analysis (PCA). These steps are implemented through a sparsely connected, simple, linear feed-forward neural network. We compute the standard deviation over all input signals from the previous layer (after thresholding) for feature selection across our training set. Only features exceeding a certain variability threshold were retained, eliminating uninformative zero or constant features over the training set. In total, 17% of the original features were retained.

The resulting thresholded PCA features serve as input to the expression-selective neurons from our norm-referenced read-out layer.

4.3 Model B

Looking at the success of our previous model A 4.2, which demonstrated that norm-referenced encoding is a viable mechanism for the perception of dynamic facial expression, we wanted to leverage deep learning architecture to access more robust early-layer features. We aim to feed our model with more challenging stimuli. We thus selected a CNN architecture expecting to obtain colour, shape and texture-invariant facial features. Therefore, the main difference between models A and B is the removal of the Gabor filters that are too simplistic to extract feature from more advance dataset as to succeed in the second objective of my thesis. I decided to add a more advanced feature extraction technique using a CNN architecture.

We chose VGG19 architecture Simonyan and Zisserman (2014) to replace our Gabor filters for this step for two reasons. First, it is a plain network structure which facilitates the interpretation of its feature maps. Second, detailed comparisons with neural data suggest that this network provides relatively good fits of neural data from the area V4 according to the *brain-score metric* Schrimpf *et al.* (2018, 2020) (objective three of my thesis). Thus, this model makes a biologically plausible input that provides a similar dictionary of mid-level features as the brain. We used a model implemented in Tensorflow Abadi *et al.* (2015) that was trained on the ImageNet dataset Deng *et al.* (2009). As features entered our architecture's higher levels, I selected the max-pool3 layer which had the highest IT score from the *Brain-score* metric, resulting in a spatial grid resolution of 28×28 points over the whole image.

Using this model, I replicate results from model A 8.2, but did not publish any results as we consider this step as a incremental step.

4.4 Model C

By incorporating a Convolutional Neural Network (CNN) architecture into my previous model (model B, see Section §4.3), I gained access to a large dictionary of features that closely resemble those found in area V4 of the brain. However, utilising feature maps from a mid-layer of VGG19 introduced many complex features that are irrelevant for facial expression recognition. Thus, the next step involves identifying and filtrating these non-task-specific features.

Furthermore, the norm-referenced encoding framework necessitates a specific latent space that facilitates the transfer of its tuning vectors across different facial shapes. I constructed such a latent space in this iteration by leveraging the 2D displacements within the CNN feature maps. The core concept behind this approach is that the direction of the movements associated with expressive features, such as raising an eyebrow, remain consistent across various facial shapes in the specific case where all faces are oriented in the same direction. Moreover, the model does not yet deal with translation or scale shift, which I will resolve in my next iteration.

I use the convolutional property of the CNN architecture and assumed that each feature map corresponds to a distinct semantic feature. Consequently, I treated the activity within a feature map as the strength and spatial location of the feature of interest on the face. This assumption enabled the tracking of spatial features, similar to a facial landmark detector, within the latent space of the VGG19 architecture.

Concretely, the main difference between the previous model B and this new model C, is the elimination of the principal component analysis (PCA) component and the introduction of a network dissection technique. This technique facilitated the creation of a set of 2D facial landmark detectors using Radial Basis Function (RBF) templates. With these adjustments, I developed ten facial landmarks that independently track facial expression displacement features, irrespective of face shape, colour, or texture.

4.4.1 Network Dissection

To filter all irrelevant facial expression features from the VGG19 readout layer, I exploit a network dissection approach for CNNs from Bau *et al.* (2017) that extracts only useful feature maps. The benefit of this approach is that the method works on any CNN. However, it requires constructing a novel set of label images to meet our needs (details of the dataset is given in Section §7.2.3). As such, I built a dataset defining 11 facial parts. The annotation was accomplished by defining a binary mask $L_c(\mathbf{x})$ for each face part c, where $\mathbf{x}$ signifies the pixel coordinate. The activity (a) of the output units that extract feature type k from the image with receptive field centre $\mathbf{x}$ of the CNN are thresholded, resulting in a binary mask $M_k(\mathbf{x})$, where M_k takes the value one if the corresponding neuron activity exceeds the top quantile level T_k, such that $P(a_k > T_k) = 0.005$, and the value 0 otherwise. An index that quantifies the accuracy of the identification of a face part c by neuron k is given by the quantity

$$\mathrm{IoU}_{k,c} = \frac{\sum_{\mathbf{x}} |M_k(\mathbf{x}) \bigcap L_c(\mathbf{x})|}{\sum'_{\mathbf{x}} |M_k(\mathbf{x}') \bigcup L_c(\mathbf{x}')|}. \tag{5}$$

I keep the feature of type k in the set of features relevant for detecting face part c if the corresponding index exceeds 0.04. To construct the latent space, I select only four face parts corresponding to the *eyes, nose, eyebrows, and lips*. I found seven activation feature maps matching the *eyes*, 61 matching the *nose*, 2 for the *eyebrows*, and 18 for the *lips*. Therefore, I keep these 88 matching feature map types k from the VGG19 Conv3-3 layer.

4.4.2 Landmark Detectors (RBF Templates)

The second step is to create facial expression landmarks. I create 10 landmarks by training radial basis function (RBF) templates. The ten landmarks are left and right exterior eyebrows, left and right interior eyebrows, left and right lower eyelids, left and right lips corner and top and down middle lips (see Figure 4.2). I used the *eyebrows* and *lips* semantic feature maps from the network dissection technique.

I construct my landmarks by saving templates within the remaining (filtered) feature maps at the spatial position corresponding to the different local face fragments of interest. The RBF centres are specified by the activations of the relevant features $F_k(\mathbf{x})$, concatenated into a feature vector $\mathbf{f}$. The RBF outputs are given (dropping indices) by the function $a = \exp\left(\frac{\|\mathbf{f}\text{ - }\mathbf{b}\|^2}{2\sigma^2}\right)$, where the vector $\mathbf{b}$ signifies the RBF centre, and σ determines the tuning width of the template detector. RBF responses are computed for each grid point $\mathbf{x}$, defining an output map. I threshold the RBF outputs with a ReLU function (threshold 0.1).

This RBF approach has two advantages: RBF patterns have support from electrophysiological recordings from the inferior temporal (IT) cortex in Rhesus Macaque Logothetis *et al.* (1995), and a single example is sufficient to learn a template. To create templates from a single image and to minimize the cross-talk between different landmarks, we define a spatial region (patch) around each landmark centre, which was chosen to include all relevant parts of the landmark while minimizing the cross-talk with all other landmarks. The tuning parameter σ was then increased as much as possible, trying to obtain RBF outputs that detect the positions of the landmarks robustly across all expressions. While simple, this method produces satisfying results. However, it is not applicable to larger datasets due to the need to optimize the patch size and the parameter σ for each detector. This limitation will be removed latter in model E 4.6.

4.4.3 Estimation of Landmark Positions

We base our architecture on the assumption that position shifts of landmarks are similar for expressions presented on different head shapes. The position of an individual

landmark can be easily derived from the corresponding RBF output map by a population vector approach. A position estimate of landmark (face fragment) l is given by the two-dimensional vector

$$\mathbf{p}_l = \frac{\sum_{\mathbf{x}} \mathbf{x}\, a_l(\mathbf{x})}{\eta + \sum_{\mathbf{x}} a_l(\mathbf{x})} + \mathbf{p}_0. \tag{6}$$

If the RBF output activity map $a_l(\mathbf{x})$ has a single peak, this expression estimates the center of gravity of the activation distribution in image coordinates. The small positive parameter η prevents division by zero if the activation map is zero. The constant $\mathbf{p}_0$ is a default position that becomes relevant if the activity is zero and was set to the center of the image. (This offset value drops out in the computation of the difference vectors relative to the norm vectors.) We concatenated the individual estimated landmark positions into a single feature vector, which defines the vector space for norm-referenced encoding.

At this stage, I achieved the capability to track the displacement of facial landmarks from a single training example across various head shapes. This accomplishment marks a crucial milestone in translating our psychophysical findings into practical model guidelines. The objective is to construct a model that not only demonstrates data efficiency but also exhibits invariance across different head shapes. The model, showcasing these attributes, was formally introduced during a conference talk at VSS2022 Stettler *et al.* (2022).

4.5 Model D

Model C validates our design by demonstrating how norm-referenced encoding can encode diverse facial shapes by constructing landmarks that can independently track facial expression displacement features, irrespective of variations in colour and texture. However, a significant limitation persists within Model C. It relies heavily on a specific reference frame, necessitating the learning of a unique reference for each face shape, which causes the model to be highly impaired with translations or scale shifts.

To address this limitation, I introduce a second stream that encodes the facial identity feature and updates the reference frame accordingly. This innovation is the fundamental principle of multi-domain norm-referenced encoding, demonstrating the benefits of splitting the learning process into two distinct phases: learning to deduce the domain from the feature vector and learning to infer the expression from the difference vector (see Section 3.1). This constitutes one of the key contributions presented in my Manuscript 4.

This new model implements the creation of a two-pathway architecture. It comprises the Facial Expression Recognition (FER) pathway inherited from Model C, and an entirely new Facial Recognition (FR) pathway designed to infer facial identity, face size, and position.

4.5.1 Face Recognition (FR) Pathway

The landmark detectors in the FR pathway serve the purpose of determining the position of the face, its size, and the type (identity) of the head, such as human, monkey, or cartoon. Therefore, the FR pathway acts as an identity detector and should not be invariant to different face shapes and textures, as we want to utilize these features for face recognition within the image. However, this poses a new challenge as our previous *masked* RBF templates cannot be used. These templates were designed with the assumption that the model knows the exact position of facial features in the image, which is provided when we give the facial identity to the model. The limitation of the mask becomes evident when a face translates within an image, as it may hide the wrong part of the face.

To address this issue, we decided to remove the mask altogether. However, this introduces a strong cross-talk between the RBF templates. The high symmetry of the face means that a template designed for the left eye may also recognize the right eye. To mitigate this, we employ a technique found in Vision Transformer (ViT) architecture (see Dosovitskiy *et al.* (2020)), which involves dividing the images into patches.

To maximize the spatial accuracy of the detected landmark positions, we split the spatial grid into 16 smaller patches (each with a size of $14 \times 14 \times l$ grid points). Moreover, we constructed a set of three new facial landmarks for the nose and the two eyes. Within each patch, the position of these individual landmarks are determined by a population vector approach using formula (6). In addition, the face type of the winning RBF template is determined. This is done separately for each landmark type and patch. At the end, only the position of the most activated patch is retained for each landmark type, assuming that there is only one valid landmark position per landmark type. Based on this representation, the face type is determined by counting the winning template detectors for each face type. The face type with the maximum count determines the detected head type or identity (e.g. human, monkey, etc.).

The horizontal position of the face is estimated as the centre between the positions of the two eye landmarks, and the vertical position is the average between the vertical position of the two eyes and the nose. The position and size of the face are used to rescale and translate the previously learned reference vectors for the actual face size and position. The reference vectors are learned from a neutral face for each face type at a standard size.

The previously described procedure assumes that only a single landmark of each type is detected per patch. This constrains the input face to be sufficiently large. Therefore, if two RBF templates are activated within the same patch, the landmark position estimation (see 4.4.3) will be inaccurate. However, we found that this simple approach resulted in acceptable accuracy of the landmark position estimates.

As such, model D now creates 13 landmarks, defined by training of RBF templates, split into two sets for the FR and FER pathway (Figure 4.2). The first set of 3 landmarks is from the FR pathway and consists of the left and right eyes and the nose, exploiting the feature maps from the semantic concept filtering for *eyes* and *nose*. The second set

(10 landmarks) provides the input for the FER pathway.

4.6 Model E

Figure 4.2: Landmark positions. Red: 10 landmarks used for the FER pathway. Green: 3 landmarks used in the FR pathway.

Model D confirms our hypothesis that multi-domain norm-referenced encoding (MD-NRE) can facilitate data-efficient transfer learning in a Facial Expression Recognition (FER) task, as discussed in Section 8.3.1. However, one notable limitation inherited from Model C is the use of ”masked” Radial Basis Function (RBF) templates within the FER pathway. While this limitation was acceptable for testing Model D on a small dataset (see Section 7.3), the forthcoming iteration, referred to as the MD-NRE model in our paper Stettler *et al.* (2023a), addresses this issue to accommodate larger datasets effectively.

In essence, MD-NRE updates the FER pathway from Model D to align with the FR pathway introduced earlier. Furthermore, it uses a semi-supervised algorithm for training the landmark detector to optimize the RBF σ parameters. A significant distinction in this approach is the possibility of the same landmark being represented with multiple RBF templates. Figure 4.3 depicts this change. The incorporation of semi-supervised optimization for the σ parameter minimizes interference between different landmarks while preserving robust generalization across various images of the same facial type. This iterative enhancement positions our model to scale effectively to significantly larger datasets.

4.6.1 RBF Optimization

Using a larger dataset allows us to train RBF patterns with multiple images of the same category. By combining the output responses from the same landmark, our model can be trained with various images. We decided to combine each output response using a simple addition operation. This means we use a set of RBF templates for each landmark instead of just one, enabling our model to capture more complex patterns in the feature space.

The tuning width σ and the use of multiple RBF templates per landmark are optimized through an iterative algorithm that minimizes interference between different landmark detectors (Figure 4.3, right).

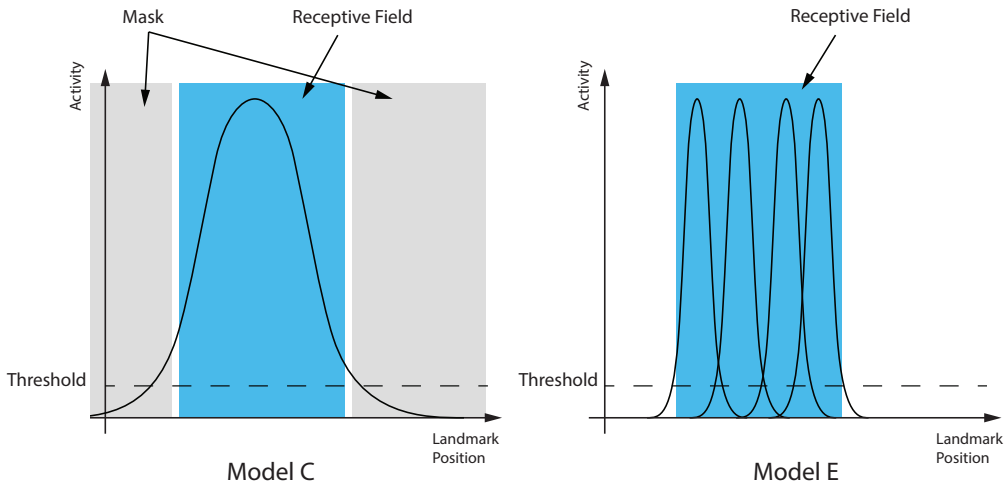

Figure 4.3: Illustration of the constructed receptive fields for a hypothetical one-dimensional feature space for model C (left) and model E (right). Templates for the BFS dataset consist of a single RBF with optimized width. For the FERG dataset we use a semi-supervied algorithm to construct templates that can be composed from mutliple RBFs and whos tuning widths is jointly optimized.

We take advantage of our model design architecture to obtain an initial set of landmark detectors. We implement a semi-supervised approach to determine generalization and cross-talk between landmark detectors in the same class, or different classes of expressions and head shapes. For each landmark we start with one image and label the landmark position. To avoid cross talk between mirror symmetric parts of the face (e.g. the two eyes). We use the 16 patches as a 4×4 grid, with each patch having a size of 14×14 grid points (see 4.5.1). This constrains the landmark search to individual patches. Starting with a low tuning width σ, we then increase this value until we obtain a spurious detection of the same landmark in another patch. This is possible with the assumption that only one face is present within an image during training. This initial value of the tuning width is further optimized using images from the same avatar type. For the next image from this class we maintain the tuning parameters if this result in the detection of a single landmark. If no landmark is detected in this subsequent image, we add a new RBF template for this landmark by labelling its position in the new image. If however the new image result in the detection of more than one landmark (meaning cross-talk activity), we reduce σ until only one landmark is detected. We summarize the steps in the box Algorithm 1.

Algorithm 1: Semi-supervised learning procedure for the RBF templates and optimization of σ

```
foreach image do
    set initial σ new_σ ← σ;
    find all landmarks;
    if no landmark found then
        label new image;
        create novel template from new labelled image;
        add it to the RBF templates for this landmark;
        find a maximum new_σ that does not lead to cross talks between
         patches;
    else if number of landmarks found > 1 then
        reduce new_σ until no more cross-talk occurs;
    end
    if new_σ < σ then
        σ ← new_σ;
    end
end
```

4.7 Dynamic Module

Architectures A to E were initially developed as object classification models for facial expression recognition. These models follow a feed-forward architecture designed to process static images. However, our experimental approach involved dynamic stimuli in psychophysics and electrophysiological experiments.

Previous research by Krumhuber et al. Krumhuber *et al.* (2013) has demonstrated the advantages of employing dynamic stimuli in facial expression recognition tasks. Dynamic stimuli have been shown to enhance the ecological validity of such tasks, particularly in improving the identification of affect for subtle expressions and distinguishing between genuine and fake expressions. It's worth noting that some studies have reported no discernible difference between static and dynamic perceptions when examining natural faces Kätsyri and Sams (2008); Kamachi *et al.* (2013).

Given the ongoing debate about the advantages of dynamic stimuli, the primary objective of our dynamic module is to compare the predictions generated by our Multi-Domain Norm-Referenced Encoding (MD-NRE) with our behavioural results. To ensure minimal interference with our static categorization performance, we included an output network composed of "differentiator" neurons, a concept first introduced in Stettler *et al.* (2020). Differentiator neurons respond phasically to changes in activity (v), and we aggregate the output signals from these neurons to obtain the decision neuron's output

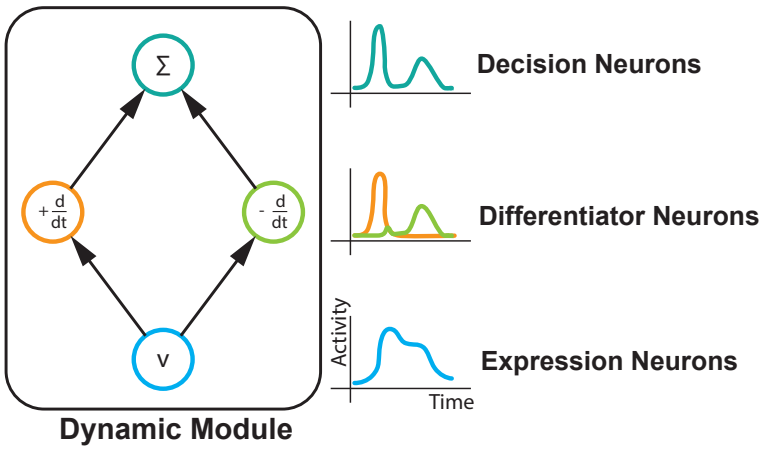

Figure 4.4: Representation of a dynamic module. The activity of the expression neurons from our norm-referenced encoding is fed into two differentiator neurons. The differentiator neurons respond phasically to increase and decrease in activity. Then, the activity of these differentiator neurons are summed up witin the decision neuron.

$$\varphi(t) = |v| = \left[\frac{\partial}{\partial t}v\right]_+ + \left[-\frac{\partial}{\partial t}v\right]_+ \tag{7}$$

Figure 4.4 illustrates this network architecture and the predictions of the neuronal activities. To create a dynamic model, we add one of this sub-network on top of each expression neuron from our base model. As showed in Stettler *et al.* (2020), such a dynamic module makes two strong behavioural predictions. First, the model cannot differentiate between a stimulus played forward or backward. Second, the perception of affect (expression strength) would be encoded by the norm-referenced encoding rather than by a dynamic process. These predictions will be further presented in §8.2 and §8.4.2 respectively.

Chapter 5

Neuroscience

One of my thesis objectives, detailed in Section 2 (refer to §2.1), centers on closely aligning with experimentalists in neuroscience. To accomplish this, I meticulously selected each component during the architecture design. My intentional decision was to exclusively employ components with a biological equivalent. While this choice imposes an additional constraint on my models, it significantly enhances transparency compared to training an end-to-end 'black box' architecture. The fundamental concept behind this approach is to facilitate the comparison of my model's predictions with electrophysiological recordings. We argue that this comparison is crucial for assisting experimentalists in confirming or dismissing potential brain mechanisms. This chapter delves into how our components and model design harmonize with recent neuroscience frameworks, divided into two main parts. The first part summarizes the links between neuroscience findings and the components used in my models. The second part relates my interpretation of theoretical neuroscience notions to my model, aiming to spark discussions on the future of model design and the norm-referenced encoding principle.

5.1 Biologically Plausible Components

Feature extraction. For the feature extraction component, we selected a VGG19 model Simonyan and Zisserman (2014) as detailed comparisons with neural data suggest that this network provides relatively good fits of neural data from the area V4 according to the brain-score metric Schrimpf *et al.* (2018, 2020). Thus, this model provides outputs that resemble features extracted at the middle level of the visual processing stream. The VGG architecture has further advantages compared to more recent architectures. It is well-known and is a *plain network* which enables a better understanding of the features space within the layers.

5.1.1 Landmark Detectors

We used Gaussian Radial Basis Functions (GRBFs) as a core component to build our landmark detector. Poggio and Edelman Poggio and Edelman (1990) proposed a neural network designed for object recognition based on such GRBFs. There is evidence in the

Rhesus Macaque IT cortex that such mechanisms are plausible in the brain Logothetis *et al.* (1995). We use these GRBFs to learn patterns of face fragments. These patterns are replicated, and their output forms a 2D activity map that constructs landmark detectors using simple max-pooling operations.

5.1.2 Norm-referenced Encoding

Giese and Leopold Giese and Leopold (2005); Leopold *et al.* (2006) demonstrated that the norm-referenced encoding readout layer is biologically plausible in Rhesus Macaque IT cortex. Moreover, several other studies support this encoding Loffler *et al.* (2005); Panis *et al.* (2011); Latinus *et al.* (2013); Koyano *et al.* (2021).

5.1.3 Dynamic Module

The same is true for the dynamic part of the mechanism. We implemented the dynamics with the so-called: *differentiator* neurons. Such differentiating neurons have been observed, and the dependence of this behaviour on channel dynamics has been analyzed in detail (e.g. Ratté *et al.* (2015)).

5.2 Neuroscience Frameworks

We believe that the brain remains a source of inspiration to pursue human-like intelligence. Therefore, in the following, we discuss how our model (MD-NRE) relates to several recent neuroscience frameworks.

5.2.1 A Thousand Brains Theory

It is argued that in the evolution of the brain, the development of an entirely new mechanism is more complex, and thus a rarer evolutionary event, than re-purposing an existing mechanism to a new task Hawkins (2021). In this framework, it is believed that the cortical column is the unit of the brain, which by duplication has helped the human brain to evolve and form the *new* neocortex structure. The thousand brains framework emphasizes the learning of reference frames within these cortical columns. The core idea of the thousand brains theory is that each cortical unit has the same neuronal structure but encodes different concepts through reference frames. Therefore, the framework assumes that if a neuronal mechanism is found in a cortical column, it should also be present in all the other cortical columns of the brain.

MD-NRE partly fits this framework. On the one hand, multiple studies have found evidence of norm-referenced encoding in facial identity encoding Leopold *et al.* (2006); Loffler *et al.* (2005); Koyano *et al.* (2021), Moreover, we show how the norm-referenced mechanism is also a valid mechanism for facial expression recognition (FER). I will

present these results in §8.2 which have been published in Stettler *et al.* (2020). Finally, norm-referenced encoding has been found for the recognition of human voices Latinus *et al.* (2013) and learning shapes Panis *et al.* (2011). These results align with the framework's re-purposing argument (assuming that FR and FER take place in different cortical columns). Secondly, the norm-referenced, as its name implies, uses reference frames: The average-norm face in FR and the neutral expression in FER. Therefore, norm-referenced encoding gives a concrete use of reference frames in FR and FER tasks.

5.2.2 Absolute vs. Relative Face Space Encoding

The MD-NRE model is primarily based on findings from Giese and Leopold (2005); Leopold *et al.* (2006) describing a V-shape encoding of face neurons relative to the average-norm face. Nonetheless, Chang and Tsao (2017) in their *active appearance model* describes a linear encoding of the face space with no particular role given to the average face in facial identity encoding. While these studies were once considered discrepant results, they have recently been reconciled Koyano *et al.* (2021); Freiwald and Hosoya (2021). These results suggest a fast process that exploits encoding in absolute space and a slower one that realizes an encoding relative to the average face.

MD-NRE combines both the active appearance and V-shape encoding. MD-NRE constructs shape-invariant face fragments in its early layer within its FER pathway, leading to specialized facial landmark detectors. These landmarks relate to the shape dimensions of the active appearance model Chang and Tsao (2017). In the active appearance model, the shape dimension is constructed from the hand-labelled landmark position of the FEI database Thomaz and Giraldi (2010). The difference between MD-NRE and the active appearance model is that our model constructs a 20-dimension shape space from trained RBF patterns. In contrast, the active appearance model uses the 25 highest principal components as the face space axis from the 68 hand-labelled landmarks of the FEI database. Finally, MD-NRE implements the relative encoding as an inhibitory connection between its FR and FER pathway while the vector projections (Equation (1)) construct the linear increase of activity described in the V-shape encoding.

5.2.3 Direct Fit Model

When assessing the performance of a model, a pivotal consideration lies in its capacity to generalize effectively across diverse and novel contexts, regardless of its type or complexity. In this context, we align our perspective with Hasson *et al.* (2020), who categorizes both biological neural networks (BNNs) and artificial neural networks (ANNs) as direct-fit models. According to their viewpoint, these models engage in *mindless* optimization, using local interpolation rules within a multi-dimensional embedding space to determine the values of new examples based on their proximity to past examples. Essentially, Hasson *et al.* (2020) suggests that only one type of models exists—direct-fit models—that achieve generalization through interpolation.

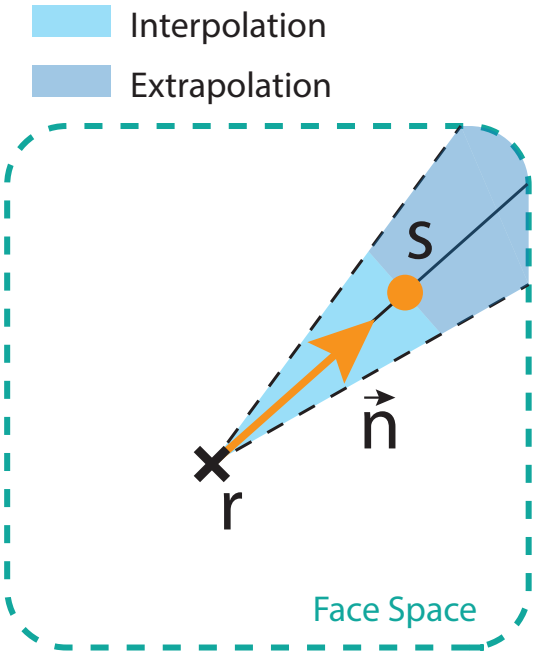

Figure 5.1: Schematic representation of a norm-referenced encoding classification boundaries from a tuning vector n (orange vector) with its reference r (black cross) and one training sample s (orange dot). The light blue depict the interpolation regime, while the dark blue depict the extrapolation regime of the mechanism.

This perspective sharply contrasts with the conventional belief regarding generalization in BNNs. BNN models are often described as engaging in extrapolation, transferring learned features from their subspace to other spaces. Conversely, ANN models are thought to perform interpolation, requiring an abundance of data to cover the entire space. The norm-referenced encoding (NRE) mechanism might offer insights into how our brains interpolate or extrapolate to derive generalizations from a limited number of training samples.

In its simplest form, norm-referenced encoding involves a scalar projection (as shown in Equation (1)). The input is projected onto *tuning* vectors, necessitating only a single data point to learn a tuning direction corresponding to the expression of interest. Let's consider the simplest case using a single training data point, as additional samples lying in the tuning direction are redundant. NRE has two regimes: an interpolation regime between the training sample and the reference frame and an extrapolation regime for the space further away from the training sample. Figure 5.1 illustrates this example.

Despite these nuances, a parallel with Hasson *et al.* (2020) becomes apparent. NRE applies the same categorization rule to new samples, irrespective of the regime in which the sample exists. Consequently, since it is agnostic to the specific regime during category inference, NRE can be regarded as a direct-fit model always performing interpola-

tion. Effectively, NRE, by learning a tuning direction, always perform an interpolation between an infinitely distant sample and the reference frame, elucidating how the brain generalizes in the extrapolation regime. However, it is interesting to note that NRE can disclose the regime. A value between 0 and 1 indicates interpolation, while values surpassing 1 signify extrapolation.

While it's important to note that I lack any definitive evidence to support the claim that norm-referenced encoding could explain generalization in the brain, it is worth considering that the norm-referenced mechanism effectively generalizes using a *mindless* local rule, as Hasson *et al.* (2020) suggest. Formally, our local rule can be expressed as $m = \arg\max_m v_m$, with v_m representing the neuronal activity of an expression m from Equation (1), with the proximity computed by a simple dot product.

I acknowledge that my argument requires further scrutiny, I wish to underscore that NRE provides insights into how BNNs might generalize employing an interpolation rule. The perspective that the brain generalise by performing *mindless* interpolation, which is also not entirely clear to me, is surely prone to contentious discussion within the neuroscience community.

5.2.4 Compositionality

The classic idea of compositionality is that new representations can be constructed by combining primitive elements. Compositionality is one of the three elements argued by Lake *et al.* (2017) to pursue human like-intelligence. (A second element is learning-to-learn, also referred as transfer learning, which is the main topic of MD-NRE.) MD-NRE constructs *primitive* landmark detectors and track their displacement. MD-NRE assumes that the landmarks' displacements for the recognition of facial expression are invariant across facial shapes and texture. This approach is similar to the famous FACS system from Ekman and Friesen (1978). The FACS quantification can be performed on any human head regardless of its shape, gender etc. As such, our model not only outputs a probability distribution of the class category but gives a precise quantification of the displacement of each element. Therefore, as the FACS system, our model perceives a smile as a combination of elements. Thus, MD-fits the compositionality argument.

Chapter 6

Computer Science

Consistent with the perspective shared by other researchers Lake *et al.* (2017); Hassabis *et al.* (2017); Zador *et al.* (2022), we align ourselves with the belief that gaining a deeper understanding of neural computation will unveil the fundamental components of intelligence. This incentive had a huge impact on the choice of the selected component when designing our architecture. Nonetheless, generalisation in machine learning is a large topic and a large amount of effort is dedicated to it. In this chapter, I will establish connections to the literature in computer vision and computer graphics, highlighting how my work relate to these fields.

To clarify, I want to emphasize that I am not opposed to training models on large datasets. On the contrary, I find recent achievements in zero-shot generalization, showcased by models such as CLIP Radford *et al.* (2021) or OFA Wang *et al.* (2022b), to be engineering marvels. However, the focus of my thesis is to explore potential mechanisms employed by the human brain to encode information. Thus, I draw a distinction between a data-efficient system, defined by its ability to generalise from a small training dataset, and a large multi-modal model.

Contrary to the large training sample used in CLIP or OFA, there are individuals who, despite limited exposure to books, showcase remarkable reasoning skills and the capacity to generalize knowledge to new domains. In this thesis, our interest lies in gaining a comprehensive understanding of possible mechanisms constituting human intelligence, with a specific focus on dynamic facial perception. It's essential to clarify that this work should not be directly compared to recent large multi-modal models. I acknowledge that such models, with appropriate training data, could outperform my models significantly. However, the goal of my research is to assist experimentalists in unraveling the brain's mechanisms rather than achieving mere performance metrics.

6.1 Parallel to Computer Vision

Despite the recent advances in machine learning, tested state-of-the-art FER models do not generalize (transfer) to, *e.g.*, cartoon faces (see results on our developed BFS dataset 8.3.1). There are several possible explanation for these results. CNN architectures are know to have strong bias towards textures Geirhos *et al.* (2018a); Baker *et al.* (2020). We

argue that CNNs rely too heavily on local features which outweigh global information to achieve our task. More recent computer vision models, such as vision transformers (ViT) Dosovitskiy *et al.* (2020), partially solve the local issue able to access any part of an image at any given layer. Nonetheless, they require large amounts of data to learn the spatial embedding.

Facial Expression Recognition: There are numerous publicly available human facial expression recognition (FER) datasets in computer vision (*e.g.*AffectNet Mollahosseini *et al.* (2017), FER2013 Courville *et al.* (2013), CK+ Lucey *et al.* (2010), KDEF Calvo and Lundqvist (2008)) as well as datasets of cartoon characters (FERG dataset Aneja *et al.* (2016)). FER studies typically aim to reach the best classification on one or several of these datasets Ionescu *et al.* (2013); Georgescu *et al.* (2019); Giannopoulos *et al.* (2018); Niu *et al.* (2021); Ashir *et al.* (2020). Others investigate transfer learning (domain adaptation) Zhao *et al.* (2018); Wang *et al.* (2018) from a source (human) to a target dataset (another human). Transfer learning is of special interest to us as it is a a core feature of our MD-NRE model (3.1). As I will introduce latter in Section §8.1), we have developed out stimuli to specifically investigate cross-species percetpion. To the best of our knowledge, only one study experiments with transfer learning in a cross-species FER experiment between humans (JAFFE, KDEF) and cartoon faces (FERG) Zhao *et al.* (2018) in a computer vision setting. However, this method did not investigate data efficiency and employs an alternative method, learning a mapping function from the source and target datasets to a common feature space. Constructing such a mapping function bears greater resemblance to the task of facial retargeting Ribera *et al.* (2017); Chaudhuri *et al.* (2019). Facial retargeting aims at learning a mapping function which links the same facial expressions across different head shapes (domains) in the context of facial animation. As such, data-efficient, cross-species FER generalization is not well studied.

Data Efficient Transfer Learning: Developing machine learning models that transfer knowledge is a critical goal in the field. There are many generalization-related research topics such as transfer learning, zero-shot learning, meta-learning, and causal representational learning. In our study, we define our task as falling between domain adaptation (DA) Kouw and Loog (2019) and domain generalization (DG) Wang *et al.* (2022a). The goal of domain generalization is a good performance on novel domains by models trained on one or several distinct, but related domains. As in DG tasks, we assume that our training and testing sets are not identically and independently distributed (*i.i.d.*).

A key difference between our approach and DG is that, as DA, we do not consider the target domain to be entirely unknown. Instead, we allow our model to see one training example per target domain. We achieve high data efficiency by introducing domain-specific inductive biases, which we refer to as our reference vectors. As the reference vectors are updatable parameters of our MD-NRE model (Model D 4.5 and E 4.6), our approach falls under the *algorithm strategy* in few-shot learning Wang *et al.* (2020), which means that existing parameters are refined during domain transfer.

6.2 Parallel to Computer Graphic

Early computational vision work proposed encoding shapes in vector spaces relative to a reference or norm pattern Vetter and Troje (1995); Vetter *et al.* (1997). The construction of vector spaces that parameterize properties of image classes is standard in computer graphics, particularly for faces Blanz and Vetter (1999); Li *et al.* (2017); Egger *et al.* (2020). 3D morphable models (3DMM) are powerful tools to construct, animate, and transfer facial expressions between avatars. 3DMMs learn to separate shape from appearance variation and construct a low-dimensional prior of a basic head shape. As such, they are intuitive and help in a natural way to disentangle shape changes that carry semantically critical information, such as expression or identity from the basic shape of the head. The intuitive representation of a 3DMM that disantangle shape and expression features is the closest to our landmark systems.

Chapter 7

Stimuli & Dataset

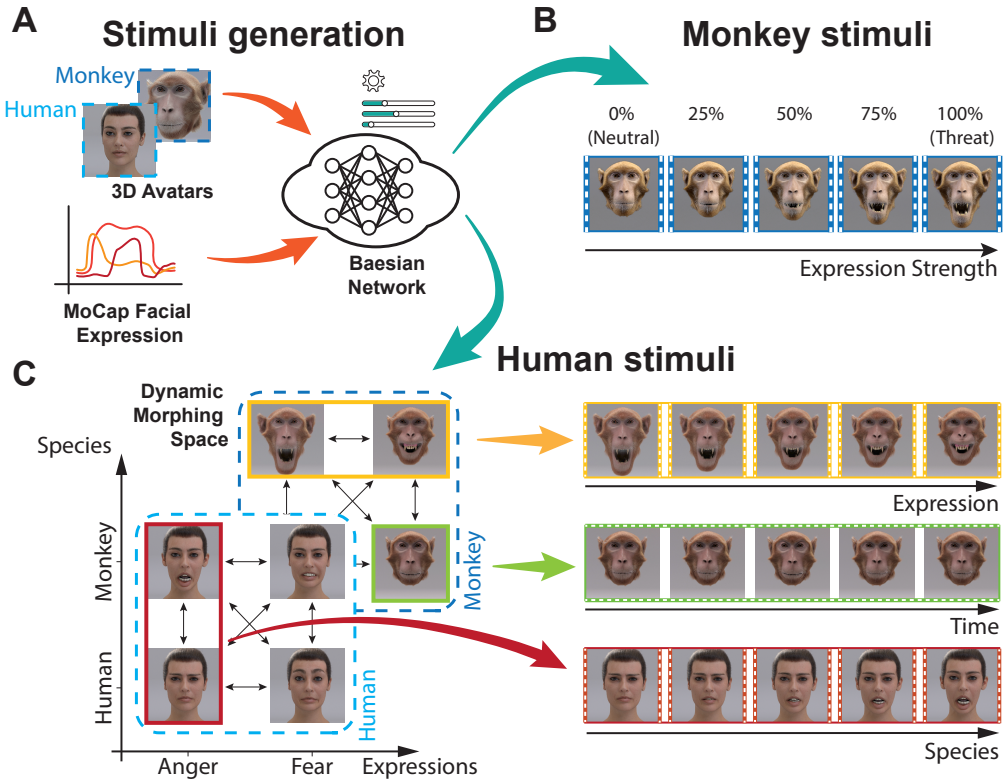

Figure 7.1: A) Schematic representation of the stimulus generation. The Bayesian network is fed with motion capture (MoCap) data of dynamic facial expression from Human and Rhesus macaque recording. The inferred dynamic expression is then map onto a human and a monkey 3D avatar. B) Example of a monkey dynamic stimuli displaying the 5 level dynamic morphing of the expression strength level. C) Representation of the 2D dynamical morphing space used in the psychophysic experiment. with the two axis for the control of the expression and species level. The colored rectangles display example of several stimuli over the expression axis (orange), species axis (red) or the expression over time (green).

In this chapter, I will introduce the stimuli and dataset developed and used during my PhD. The developed stimuli are primarily used to record psychophysical behavioural data from Humans and electrophysiological recordings from Rhesus Macaque. Nonetheless, I have also utilized them to train and test the assumptions of my models. Specifically, I will focus on discussing the design and characteristics of the stimuli. For more detailed information about the generative Bayesian model used for generating motion and the creation of the avatar heads, I encourage readers to refer to our publications Taubert *et al.* (2012, 2021) as it has been mainly developed by my collaborator Nick Taubert. Therefore, my role within the stimuli creation part of the project involved designing the experiment and using the generative model to create the dynamic stimuli and datasets. The generated static dataset (7.3 and 7.4) were developed entirely by myself. Figure 7.1 shows an overview of the stimulus generation process. The key characteristic of these stimuli is to display highly realistic sets with full parametric control of motion style and face shape.

7.1 Stimuli

Our stimuli can be divided into two main sets: one to record electrophysiological recordings from Rhesus macaque, and one to record psychophysics responses from Human participants. I will start with the background of the stimuli that lead to the electrophysiological recording and follow with the stimuli used for the human participants.

7.1.1 Background: Avatar Creation and Motion Recording

As to create photo-realistic dynamic facial expression stimuli, we first required realistic avatars and dynamic motion of facial expressions. We use two avatars, a Rhesus Macaque, for a better ecological validity with the electrophysiological recording of Rhesus Macaque, and a Human avatar. We also recorded motion capture (MoCap) data from both Humans and Rhesus Macaque to drive our avatars with realistic dynamic facial expressions.

Monkey Avatar. The monkey head model is derived from a structural magnetic resonance scan of a Rhesus Macaque (9 years old, male). The surface of the face is modeled by an elastic mesh structure (Figure 7.3C) that imitates the deformations induced by the major face muscles of macaque monkeys Parr *et al.* (2010). The resulting surface mesh model is regularized and optimized for animation. A sophisticated multi-channel texture model for the skin and fur animation were added. To our knowledge, we present the only highly realistic monkey avatar that is animated with motion capture data from real animals used in physiology so far. It has been demonstrated by a recent study of our lab that our dynamic monkey avatar induces behavioral reactions of macaque monkeys that are very similar to ones elicited by real movies, reaching the *good side* of the uncanny valley Siebert *et al.* (2020).

Human Avatar. The human head is based on a scan-based commercial avatar with blend-shape animation (EISKO Digital Louise), exploiting a multi-channel texture simulation software. Mesh deformations compatible with the human face muscle structure are computed from motion capture data in the same way as for the monkey face.

MoCap data. The face motion is based on motion capture (VICON), exploiting 43 markers that were placed on the face of a rhesus monkey (7.3B) that executed different facial expressions. By interaction with an experimenter, three expressions (prototypes) were recorded: Fear, Lip Smacking and a Threat/Angry expression (Figure 7.2). Exploiting a muscle-like deformable ribbon structure, the motion capture data is transferred to the face mesh, resulting in highly realistic face motion. The human marker set correspond to the one of the monkey, except that it lacks markers on the ears. The human actor was instructed to show two facial expressions 'anger' and 'fear'. Processing is identical to the marker trajectories of the monkey expressions.

Bayesian Network. In order to create continuous parameterized spaces of facial movements, we exploit a motion morphing method that is based on a probabilistic hierarchical generative Gaussian process model. The architecture comprises three layers. Two Gaussian process latent variable models (GP-LVMs) Lawrence and Moore (2007) form the lower two layers, and a Gaussian process dynamical model (GPDM) Wang *et al.* (2007) form the highest layer.

7.1.2 Expression Strength Level

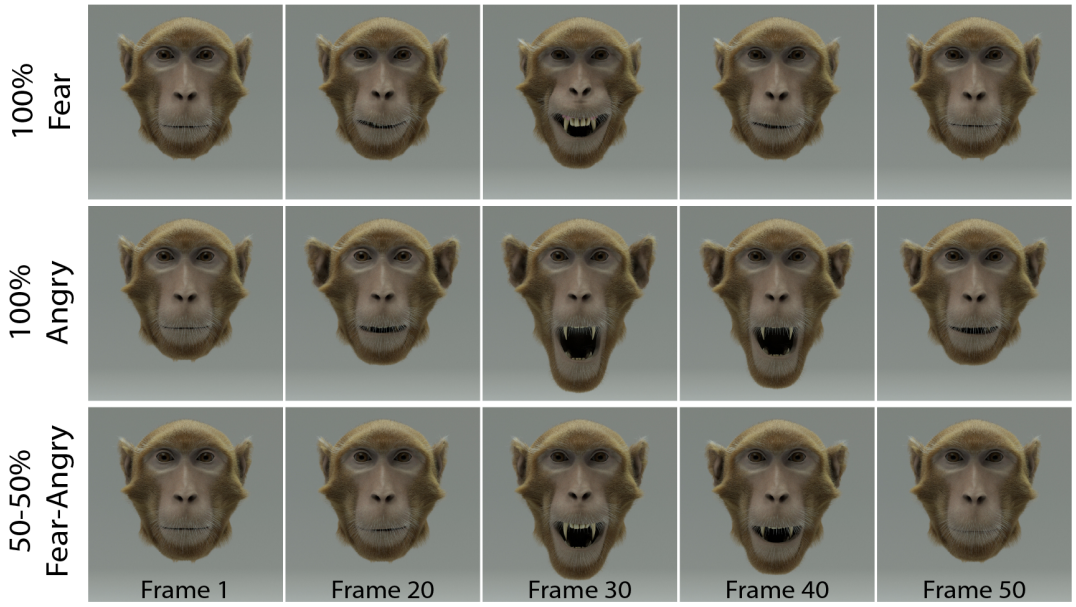

Figure 7.2: Monkey Avatar Head. Movies of the three training expressions Fear, Lip Smacking and Angry, and morph (50% - 50%) between Fear and Angry.

The first developed dataset serves the purpose of gathering electrophysiological data in Rhesus Macaque to explore the potential presence of norm-referenced encoding in dynamic facial expression perception. Simultaneously, I utilize this dataset to investigate whether norm-referenced encoding could be a suitable encoding for facial expression recognition (FER) tasks. This dataset is used with model A (4.2) and published in Manuscript 2 Stettler *et al.* (2020).

To create this dataset, we started with the three recorded monkey expressions (prototype): Fear, Lip Smacking and a Threat/Angry expression (Figure 7.2). In addition to the original expressions (prototypes), we generate morphs between them exploiting a Bayesian motion morphing algorithm Taubert *et al.* (2012). These reduced expression strength are generated by morphing the original expression with a neutral facial expression. The result is a set of controlled facial expression with the following expression strength: 25-50-75-100 % for each prototype. This allowed us to investigate the activity of STS neurons and study the behavior of the models for gradual variations of expression strength for each individual prototype, as well as for the gradual morph between prototypes. Expressions always start with neutral, evolved to a maximally expressive frame, and goes back to the neutral face. Natural duration of the expressions are between 2 and 3s, nonetheless, for analysis of our models, all expressions are time-normalized to 50 frames. In addition, morphs between the expressions Fear and Angry with contributions of 0-25-50-75-100% of the Fear prototype are generated. Figure 7.2 illustrate this dataset.

7.1.3 2D Dynamical Morphable Space

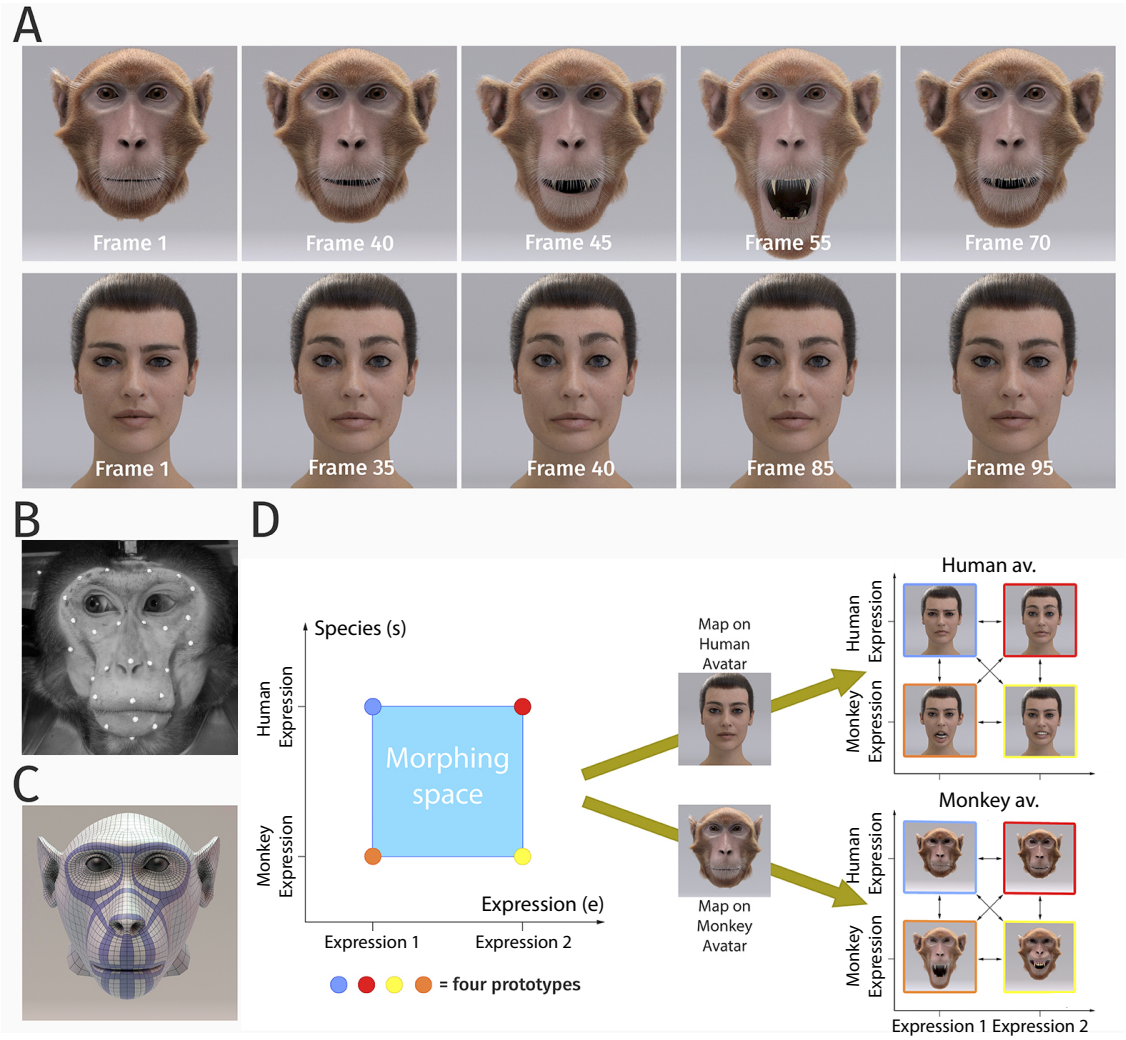

Figure 7.3: Stimulus generation and paradigm. (A) Frame sequence of a monkey and a human facial expression. (B) Monkey motion capture with 43 reflecting facial markers. (C) Regularized face mesh, whose deformation is controlled by an embedded elastic ribbon-like control structure that is optimized for animation. (D) Stimulus consisting of 25 motion patterns, spanning up a two-dimensional style space with the dimensions 'expression' and 'species', generated by interpolation between two expressions ('anger/threat' and 'fear') and the two species ('monkey' and 'human'). Each motion pattern was used to animate a monkey and a human avatar model.

The second dataset developed is used to investigate cross-species expression categorization perception across different head shapes. This is the dataset published in Manuscript 1 Taubert *et al.* (2021). The specificity of the dataset is the control of the information

content and the expression strength of the dynamic face stimuli to investigate the perceived categorization boundary between expressions. For this purpose, we exploit our generative Bayesian model on the trajectories of the control points of the face. The algorithm allows to create linear combinations in space-time between prototypical motions, controlling smoothly the expressiveness and the style of the generated facial motion.

This dataset construct a three factor design within a 2D dynamic morph space using different basic face shape (*e.g.*avatar type). The goal is to investigate whether a difference of perception might exists between known expressions (humans) and naive expressions (monkey) and if the face shape influence the perceived expression. Human observers categorized these dynamic expressions, presented on the human or the monkey head model, in terms of the perceived expression type and species-specificity of the motion (human vs. monkey expression).

To realize the full parametric control of motion style, we exploit a Bayesian motion morphing technique to create a continuous expression space that smoothly interpolates human and monkey expressions. We used two human expressions and two monkey expressions as basic patterns, which represented corresponding emotional states (*fear* and *anger/threat*). Interpolating between these four prototypical motions in five equidistant steps, we generated a set of 25 facial movements that vary in five steps along two dimensions, the expression type and the species, as illustrated in Figure 7.3D. Each generated motion pattern can be parameterized by a two-dimensional style vector (e,s), where the first component e specifies the expression type ($e = 0$ expression 1 (*fear*) and $e = 1$ expression 2 (*anger/threat*)), and where the second variable s defines the species-specificity of the motion ($s = 0$: monkey and $s = 1$: human). The dynamic expressions were used to animate a highly realistic monkey and a human avatar model. To vary the two-dimensional stimulus features, we rendered the avatars from two different view angles: from the front view and from the view that was rotated by 30 degrees about the vertical axis. This rotated view maximized the differences in the two-dimensional appearance relative to the front view while avoiding strong salient changes, such as occlusions of face parts.

Equilibrated Morphing Space

In order to control the amount of expressive low-level information, that is, the total amount of motion or shape deformation, we generate sets of equilibrated stimuli. For this purpose, we first define different measures for the low-level information content and balanced the stimuli by equilibrating these measures. Tested measures 7.4A include optic flow (computed with an optic flow algorithm) (OF), the maximum amount of deformation (projected to the plane) of the polygon mesh relative to the neutral pose (DF), and the (two-dimensional) motion flow of the polygon mesh integrated over time (MF). To control the information content of the stimuli, we generate morphs between the original motion and the trajectories of a neutral expression using our motion morphing technique. In these morphs, the original expression is weighted with the morph level l and the neu-

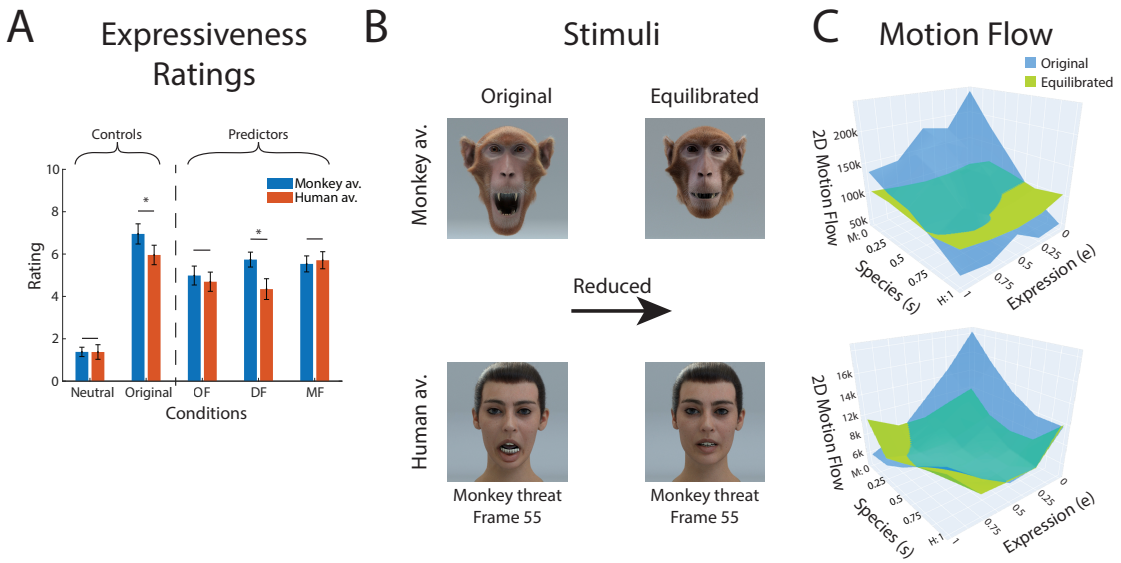

Figure 7.4: Equilibration of low-level expressive information. (A) Mean perceived expressivity ratings for stimulus sets that were equilibrated using different types of measures for the amount of expressive low-level information: OF: optic flow computed with an optic flow algorithm; DF: shape difference compared to the neutral face (measured by the 2D distance in polygon vertex space); MF: two-dimensional motion of the polygons on the surface of the face. In addition, the ratings for a static neutral face are shown as reference point for the rating (neutral). (B) Extreme frames of the monkey threat prototype before and after equilibration using the MF measure (C) 2D polygon motion flow (MF) computed for the 25 stimuli in our expression style space for the monkey avatar for the front view (similar results were obtained for the other stimulus types).

tral expression with the weight $1-\lambda$. The parameter l is chosen to equate the low-level measures of all four prototypical stimuli, separately for the two avatar models (for the front view). For this purpose, we fit the relationship between the individual measures M for the low-level information and the morphing parameter l by a logistic function of the form (a_i signifying constants)

$$M(\lambda) = a_0 + \frac{a_1}{1 + e^{a_2\lambda + a_3}}. \tag{8}$$

We use the inverse of this function to determine the values of the morph parameter l that match the value M of the most expressive prototype motion. The MF measure result in the least variability of the perceived expressiveness of the equilibrated stimuli, and thus is used to equilibrate the stimuli for all experimental conditions.

7.2 Dataset

I developed several dataset during this project. In the following, I will cover their development and introduce the publicly available dataset I used.

7.2.1 Behavioural Data

The behavioural data is the set used to investigate our participants' perception and classfication performance on our 2D morphing space. To create our behavioural dataset from the 2D morphing space stimuli 7.1.3, we first modeled our participant answers using a multinomial logistic regression analysis, the relative frequencies of the four classes $\hat{C}_j(e,s)$ were approximated by class probabilities $P_j(e,s)$ for the four classes that were modeled by a generalized linear model (GLM) of the form

$$P_i(e,s) = \frac{e^{y_i}}{\sum_{j'=1}^{4} e^{y'_j}}. \tag{9}$$

The variables y_j were given by linear combinations of predictor with coefficients β_i and variables X_i in the form

$$y_j = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + ... + \beta_{Nj}X_N. \tag{10}$$

We compared a multitude of models, including different sets of predictors. The most compact model was linear in the style space variables e and s and was given by the equation

$$y_j = \beta_{0j} + \beta_{1j}e + \beta_{2j}s. \tag{11}$$

We also tested variants of linear models that included the predictor variable $e * s$ and a predictor variable that is proportional to the total amount of optical flow, computed using a Horn-Schunck algorithm (CV Toolbox) from the stimulus movies. The different versions of the model were compared exploiting their prediction accuracy and the BIC. We discarded the models if, after addition of a new predictor, either their accuracy was decreasing or the BIC showed a decrease. See Section §8.1 for more details.

7.2.2 Computation of the tuning functions

The species-tuning functions were computed by marginalization of the discriminant functions belonging to the same species category along the variable e. The tuning function to monkey expressions as a function of the species parameter s was defined as

$$D_M(s) = \int_0^1 (P_1(e,s) + P_2(e,s))de. \tag{12}$$

Similarly, the tuning function for human expressions was given by

$$D_H(s) = \int_0^1 (P_3(e, 1-s) + P_4(e, 1-s))de. \tag{13}$$

For this function, the direction of the s-axis was flipped, so that the category center also appears for $s = 0$, just as for the function $D_M(s)$.

7.2.3 Face Semantic Dataset

The face semantic dataset is used to implement the network dissection technique from Bau *et al.* (2017) 4.4.1 introduced in model C 4.4. The goal of the dataset is to create a set of mask around the part of interest within the images as to compute the IoU score index (5).

To create this dataset, we manually labelled a custom set of 40 images following the Microsoft COCO framework Lin *et al.* (2014). The dataset comprises 20 randomly picked non-face images from ImageNet Deng *et al.* (2009) and 20 face images from AffectNet Mollahosseini *et al.* (2017). We created 23 classes (concepts) following the work of Bau *et al.*Bau *et al.* (2017), ranging from general concepts such as colours (*e.g.*blue, red, etc.), textures (*e.g.*background, clothes etc.) to the 11 face concepts of interest: eyes, eyelids, eyebrows, ears, mouth, lips, teeth, nose, face, head and hair. We used the website Hasty.ai to label our dataset.

7.3 Corresponding Expressions - BFS Dataset

The purpose of creating this dataset is to evaluate the architecture of the developed model D (4.5). The goal of this dataset is to assess our two-part learning process 3.1 hypothesis, which involves updating the reference frame of the norm-referenced encoding.

To investigate our hypothesis, we had to create our own dataset in order to control exactly the deformation of face expression across different face shape. Therefore, the specificity of the dataset is that each avatar display a set of matching facial expression. We call our new dataset the BFS (Basic Face Shapes) dataset. We use it to test the generalisation of expression recognition across multiple very different head shapes. For this purpose, we used three basic face-shape avatar models, a human, a monkey and a caricature head. We extended the dataset by constructing 15 different identities (5 identities per basic face shape) displaying 7 expressions (see Figure 7.5), resulting in a total of 105 images. All face shapes were animated using the same facial expressions.

To establish correspondence between the meshes of different avatars types, we retarget the meshes of the monkey and cartoon avatars onto the human head base mesh. We interpolate the avatar mesh to have the same number of vertices using the program R3DS Wrap (https://www.russian3dscanner.com). The correspondence between the meshes made it easy to compute offsets between the neutral poses of the neutral faces for the

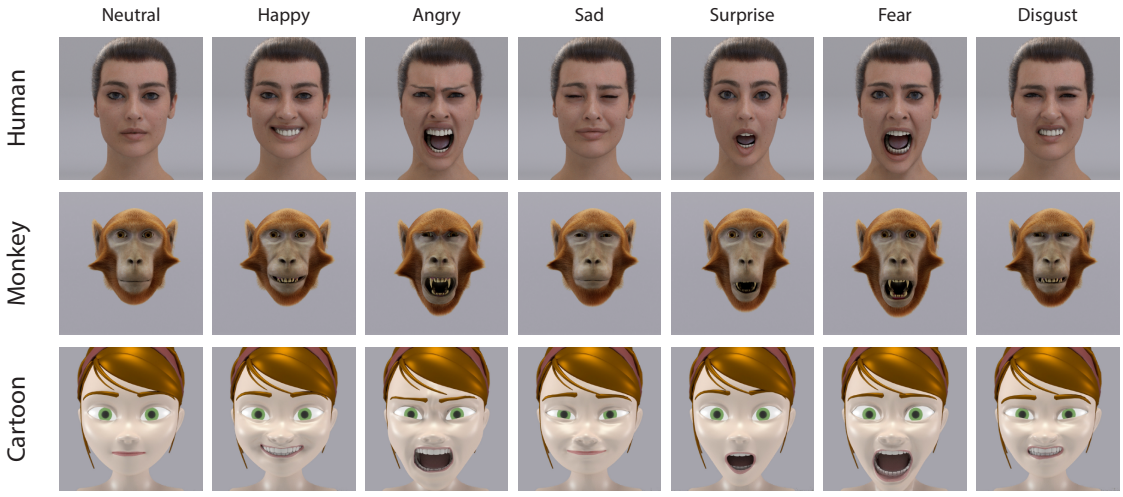

Figure 7.5: Examples of face pictures form our BFS dataset showing 7 different expressions on the three basic head shapes.

different avatar types. The human facial expressions are based on a blendshape representation provided by EISKO. We increase the number of identities by scaling each avatar head in the horizontal direction with a factor ranging from 0.8 to 1.2 with steps of 0.1, leading to 15 different identities. The dataset comprises a training set: 21 images in total constructed from each neutral face (15 images) and one image of each expression displayed on the human head avatar, and a test set (84 images). The Human and Monkey head shapes with textures and hair (fur) are taken from our stimuli 7.1.1. The cartoon character (Mery) was courtesy provided by Mark Pauly and Jason Osipa (meryproject.com).

7.4 BFS-L Dataset

This dataset is an extension of the BFS dataset, following the previous idea of the expression strength level 7.1.2. The idea is to be able to reproduce the stimuli from the rhesus macaque 7.1.2 and test the prediction of our models. Here we simply used the prototype of each expression and created reduced expression strength by manually decreasing the blenshape weights of the corresponding expressions to generate steps of 25-50-75% expression level. The BFS-L dataset comprise a total of 435 images.

Figure 7.6 shows image examples for the Human avatar from each facial expression category, varying the expression strength from 0% (corresponding to Neutral) to 100% expression strength.

7.5 AffectNet Dataset

An important dataset used during this project is the AffectNet dataset Mollahosseini *et al.* (2017), used to train differnt CNN architecture for comparison with our model. This set includes more than 1'000'000 images of human facial expressions in the wild. We also use a subset of the AffectNet dataset as Ngo *et al.*Ngo and Yoon (2020) to train our models. This subset includes only the manually labelled images for 8 facial expressions, which reduces the training set to 280k+ images.

7.6 FERG Dataset

Finally, on of our main obejective is to show that the norm-referenced encoding is capable to work is larger dataset. Therefore, to test our MD-NRE model (model E 4.6), we use a publicly available large-scale dataset including, the FERG Aneja *et al.* (2016) dataset. The dataset consists of stylized comic characters with annotated facial expressions. It contains 55'767 images across 6 different head shapes and 7 expressions. The dataset comprises a training set (43'767), a validation set (6k) and a test set (6K). Furthermore, the pose and illumination are fixed, making the FERG dataset suitable for our landmark detector.

7.7 DFEW Dataset

Dynamic Facial Expression in-the-Wild (DFEW) is a comprehensive facial expression database comprising 16,372 challenging video clips sourced from movies. These clips feature various difficulties, such as extreme illumination, occlusions, and erratic pose changes. Utilizing crowdsourcing, 12 expert annotators independently labeled each clip ten times. The DFEW database offers significant diversity, a large volume of data, and detailed annotations, including a 7-dimensional expression distribution vector for each clip, single-labeled expression annotations for seven classic discrete emotions, and baseline classifier outputs based on single-labeled annotations.

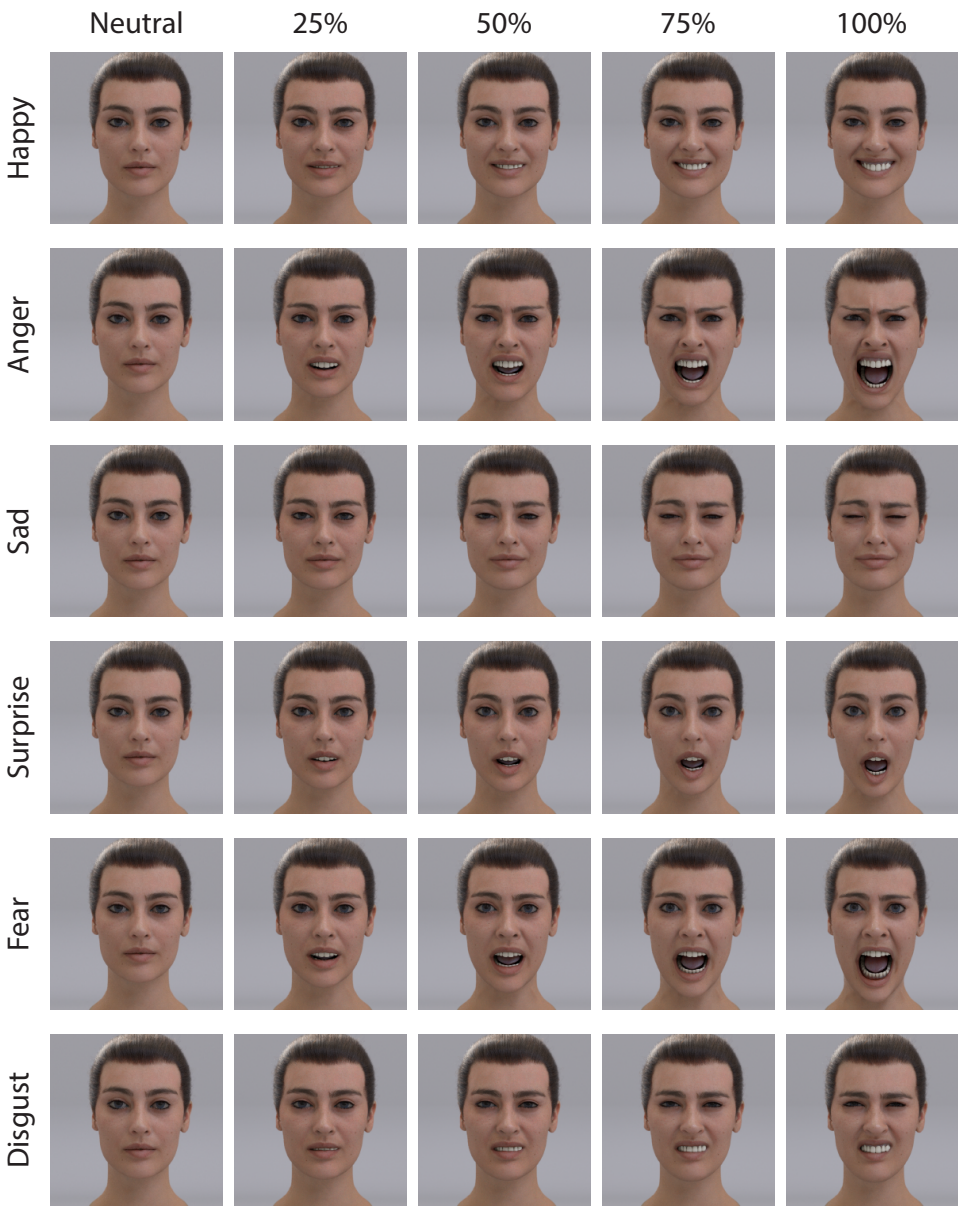

Figure 7.6: Example images taken from the BFS-L dataset on the Human Avatar.

Chapter 8

Results

Up until now, I have presented the primary hypothesis of our mechanism of interest: norm-referenced encoding (NRE) and how we can extend it to become a data-efficient mechanism by leveraging transfer learning across multi domain 3. I have contextualized this mechanism within the frameworks of neuroscience 5 and computer vision 6, and explained the setup of my PhD thesis to investigate this mechanism with behavioural and electrophysiological stimuli 7.1 as well as using facial expression recognition (FER) datasets 7.2 to benchmark my work with other computer vision models. As a result, in this chapter, I am in a position to provide all the findings obtained during my PhD research by exploiting the NRE mechanism.

The presentation of results will unfold in a systematic sequence. First, I will delve into the outcomes of my psychophysical experiments, serving as the foundational basis for the specifications guiding the design of my models. Subsequently, I will showcase how my initial model successfully encoded facial expression recognition. Building upon this, I will demonstrate the data-efficient advantages gained through transfer learning from my latest model. Finally, I will illustrate how our most recent model aligns with preliminary findings from electrophysiological recordings on Rhesus Macaque and replicates our psychophysical results. Each section will begin with a brief introduction to guide readers into the extracted results and discussion from the manuscript. This format is designed to adhere to a cumulative thesis structure, allowing a concentrated focus on the results given that the perquisite background has been laid out beforehand.

8.1 Human Perception

In this section, I report the results from my psychophysics experiments published in Taubert *et al.* (2021). Here, we use our created 2D morphing space 7.1.3 to investigate cross-species perception. This was the first step of my thesis, and these results became the foundation of my model design 4 and the development of the norm-referenced encoding framework 3. The initial motivation behind conducting this psychophysic experiment was the following: In primate phylogenesis, the visual processing of dynamic facial expressions has co-evolved with the neuromuscular control of faces Schmidt and Cohn (2001). Remarkably, the structure and arrangement of facial muscles is highly

similar across different primate species Vick *et al.* (2007); Parr *et al.* (2010), while face shapes differ considerably, for example, between humans, apes, and monkeys. This motivates the following two hypotheses: (1) The phylogenetic continuity in motor control should facilitate fast learning of dynamic expressions across primate species and (2) the different speeds of the phylogenetic development of the facial shape and its motor control should potentially imply a separate visual encoding of expression dynamics and basic face shape. The second hypothesis seems consistent with a variety of data in functional imaging, which suggests a partial separation of the anatomical structures processing changeable and non-changeable aspects of faces Haxby *et al.* (2000); Bernstein and Yovel (2015).

In the following, I will present how humans participants perceived our 2D morphing space to test these two assumptions. Details of the experiment stimuli presentation A.2 and of our pool of participants are left in the Appendix A.1.

8.1.1 Dynamic expression perception is largely independent of facial shape

In our first experiment, we use the original dynamic expressions of humans and monkeys as prototypes and present morphs between them, separately, on the human and the monkey avatar faces, with two different view angles (0 and 30 degrees rotation about the vertical axis). Facial movements of humans and monkeys are quite different Vick *et al.* (2007), so that our participants, all of whom have no prior experience with macaque monkeys, need to be familiarized briefly with the monkey expressions prior to the main experiment. During the familiarization, participants learn to recognize the four prototypical expressions perfectly, always with maximally four stimulus repetitions. During the main experiment, motions are presented in a block-randomized order, and in separate blocks for the two avatars and for the two tested views. The expression movies, lasting 5 seconds each, depict the face transitioning from a neutral expression to an extreme one, and then returning back to a neutral expression 7.3. Participants observed 10 repetitions of each stimulus. They had to decide whether the observed stimulus is looking more like a human or a monkey expression (independent of the avatar shape and view), and whether the expression is rather ‘anger/threat’ or ‘fear’. The resulting two binary responses in each trial is interpreted as assignment of one out of four classes to the perceived expression of the stimulus, independent of avatar type and view (1: human-angry, 2: human-fear, 3: monkey-threat, and 4: monkey-fear).

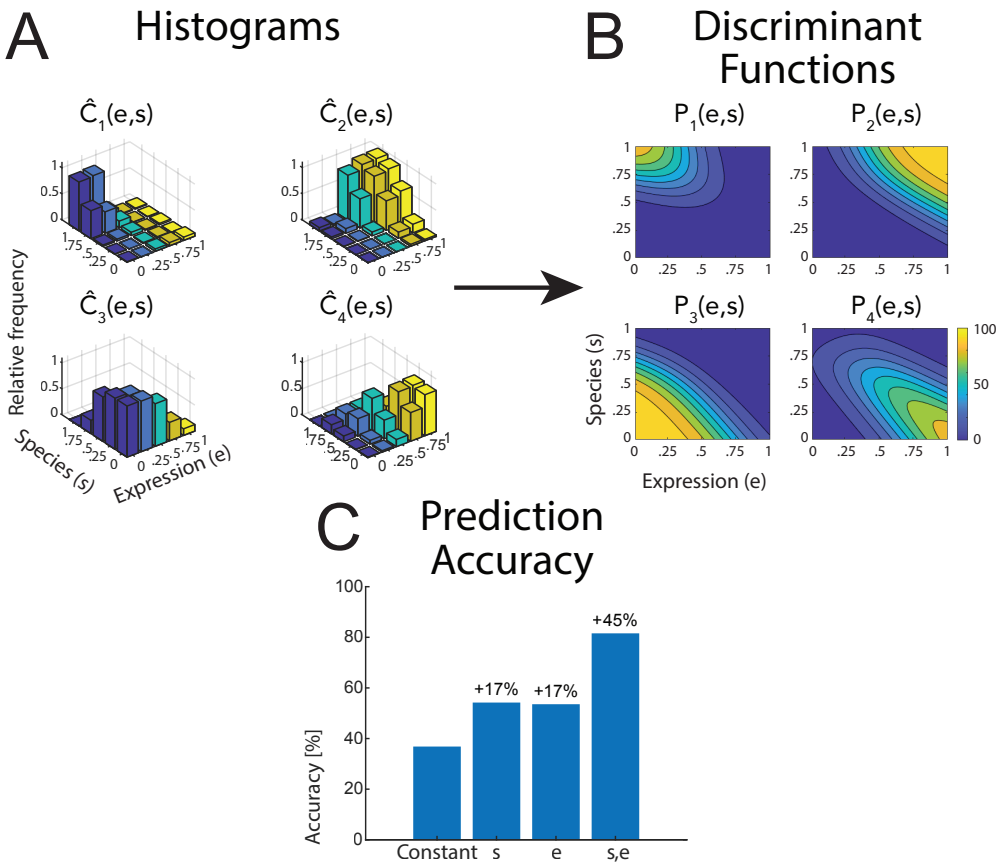

Figure 8.1: Raw data and statistical analysis. (A) Histograms of the classification data for the four classes (see text) as functions of the style parameters e (expression-axis) and s (species-axis). Data is shown for the human avatar, front view, using the original motion-captured expressions as prototypes. (B) Fitted discriminant functions using a logistic multinomial regression model (see 7.2.1). Data is shown for the human avatar, front view, using the original motion- captured expressions as prototypes. (C) Prediction accuracy of the multinomial regression models with different numbers of predictors (constant predictor, only style variable e or s, and both of them).

Figure 8.1A shows the raw classification data as histograms of the relative frequencies of the four classes $\hat{C}(e,s)$, as a function of the style parameters e (expression) and s (species) for the four tested classes. The class probabilities $P_i(e,s)$ are modeled by a logistic multinomial regression model 7.2.1, resulting in the fitted discriminant functions shown in Figure 8.1B for the different classes. Comparing regression models with different sets of predictor variables, we find that in almost all cases, a model of the form that contains the two style variables for expression (e) and the species (s) as predictors (in addition to a constant predictor) was the simplest model that provided good fits of the data (more details in Appendix A.3). Figure 8.1C shows the prediction accuracy of

regression models with different sets of predictors for the monkey avatar stimulus (data from the other conditions are presented in A.3). The different models are compared quantitatively using prediction accuracy and the Bayesian Information Criterion (BIC). Specifically, a model that also include the product $e \times s$ does not provide significantly better prediction results, except for a very small improvement of the prediction accuracy for the rotated view conditions. Models only including one of the predictors, e or s, provide significantly worse fits. Likewise, models that contain the average amount of optic flow as the additional predictor does not result in higher prediction accuracy. This implies that simple motion features, such as the amount of optic flow, do not provide a trivial explanation of our results. Summarizing, both style variables e and s are necessary as predictors, and there is no strong interaction between them. This motivate us to use the model with these two predictor variables for our further analyses.

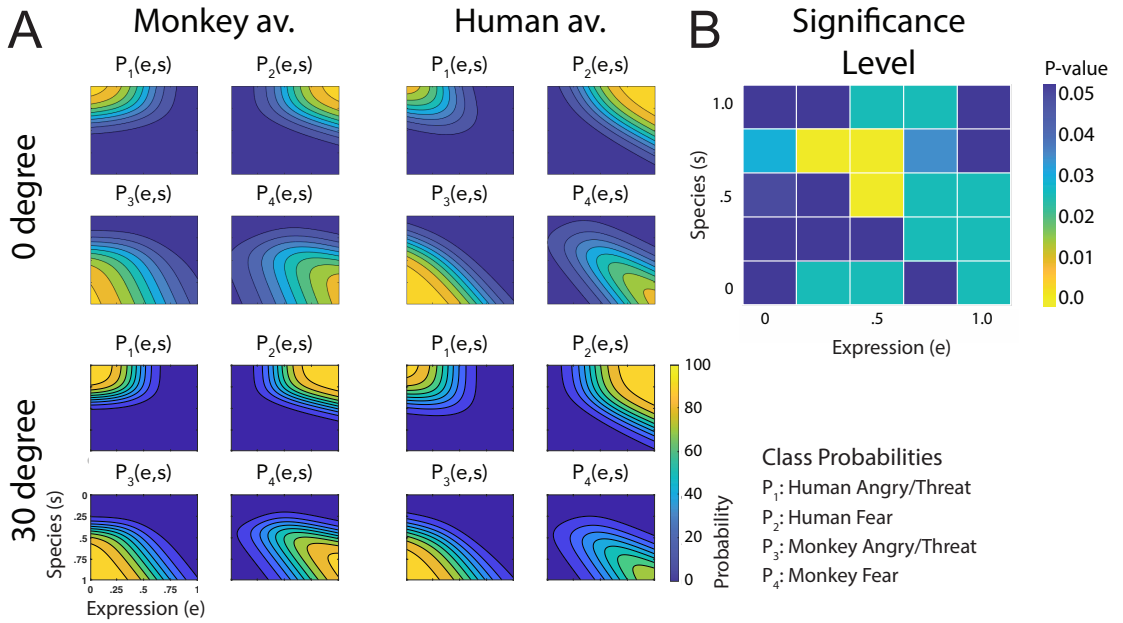

Figure 8.2: Fitted discriminant functions $P_i(e,s)$ for the original stimuli. Classes correspond to the four prototype motions, as specified in Figure 7.3D (i = 1: human-angry, 2: human-fear, 3: monkey-threat, 4: monkey-fear). (A) Results for the stimuli generated using original motion-captured expressions of humans and monkeys as prototypes, for presentation on a monkey and a human avatar. (B) Significance levels (Bonferroni-corrected) of the differences between the multinomially distributed classification responses for the 25 motion patterns, presented on the monkey and human avatar.

Figure 8.2A presents a comparison of all fitted discriminant functions, which are displayed separately for the two avatar types and the two tested view conditions. These functions illustrate the predicted class probabilities as functions of the two style parameters e and s. The form of these discriminant functions exhibits a high level of similarity

between the two avatar types as well as between the view conditions. This similarity is confirmed by the observation that the fraction of the variance that differs among these functions, relative to the shared variance, does not exceed 3% ($q = 2.75\%$; see A.4). Furthermore, a comparison of the multinomially distributed classification responses using a contingency table analysis (see A.5) across the four conditions (avatar types and views) supports the same conclusion. Specifically, the analysis reveals that only three stimuli (12%) in the style space exhibit significantly different classification responses ($p = 0.02$, Bonferroni-corrected). Notably, these differences tend to be more pronounced for stimuli with intermediate values of the style space coordinates e and s, indicating higher perceptual ambiguity (Figure 8.2B). These findings suggest that primate facial expressions are predominantly perceptually encoded independent of the head shape (human vs. monkey) and stimulus view. Furthermore, this independence extends to the two-dimensional image features, which vary considerably across view conditions and between the human and monkey avatar models. The observed independence may also account for why many of our subjects are able to recognize human facial expressions on the monkey avatar face without any prior familiarization. This finding aligns with the common experience that humans possess innate recognition abilities for dynamic facial expressions, even when presented with non-human cartoon characters that often display highly unnatural features.

8.1.2 Tuning is narrower for human-specific than for monkey-specific dynamic expressions

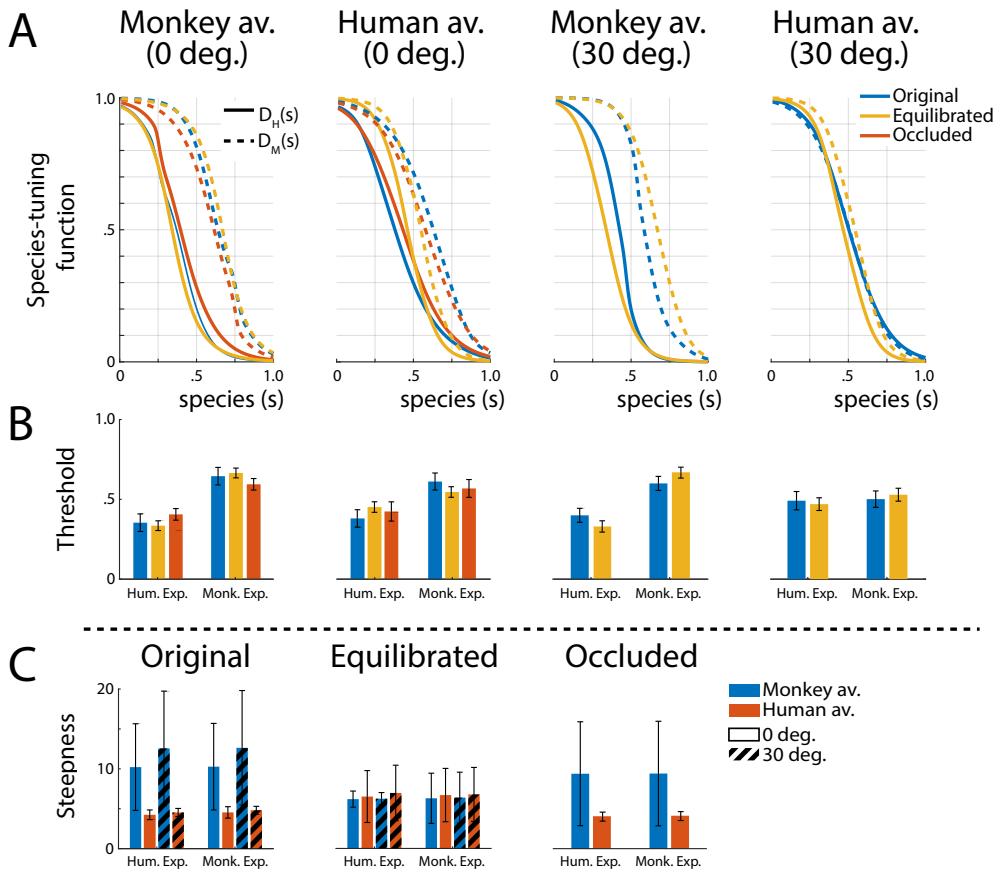

Figure 8.3: Tuning functions. (A) Fitted species-tuning functions DH(s) (solid lines) and DM(s) (dashed lines) for the categorization of patterns as monkey vs. human expressions, separately for the two avatar types (human and monkey) and the two view conditions. Different line styles indicate the experiments using original motion-captured motion, stimuli with occluded ears, and the experiment using prototype motions that were equilibrated for the amount of motion/deformation across prototypes. (B) Thresholds of the tuning functions for the three experiments for presentation on the two avatar types and the two view angles. (C) Steepness of the tuning functions at the threshold points for the experiments with and without equilibration of the prototype motions, and with occlusion of the ears.

A biologically important question concerns whether expressions of one's own species are processed differently from those of other primate species, potentially supporting an own-species advantage in the processing of dynamic facial expressions Dahl *et al.* (2014). To characterize the tuning of the perceptual representation for monkey and human ex-

pressions, we compute tuning functions by marginalizing the discriminant functions belonging to the same species category (human: P_1 and P_2, monkey: P_3 and P_4) over the expression dimension e (see 7.2.2 for details). Figure 8.3A displays the resulting two species-tuning functions, $D_H(s)$ and $D_M(s)$, revealing a narrower tuning width for human expressions compared to monkey expressions for all stimulus types, except for the 30 degrees rotated human condition.

The fitted threshold values, $D_M(s_{th})$ and $D_H(s_{th}) = 0.5$, are shown in Figure 8.3B for monkey-specific and human-specific motion (solid and dashed lines, respectively). This observation is confirmed by computing the threshold values of the tuning functions through fitting them with a sigmoidal function (see 7.2.1). When comparing the threshold values using separate ANOVAs for the four stimulus types (monkey and human front view, monkey and human rotated view), significantly narrower tuning is found for human expressions compared to monkey expressions in all tested conditions, except for the human avatar in the 30 degrees condition. These two-way mixed-model ANOVAs include the expression type (human vs. monkey motion) as a within-subject factor and the stimulus type (original motion, stimuli with occluded ears, or animated with equilibrated motion) as a between-subject factor. The ANOVAs reveal a strong effect of the expression type ($F(1,4) = 188.82$, $F(1,66) = 46.39$, and $F(1,40) = 127.35; p < 0.001$, respectively), except for the human 30 degrees condition, where the influence of this factor did not reach significance ($F(1,40) = 1.43; p > 0.23$). In all cases, we failed to find a significant influence of the stimulus type ($F(2,66) = 0.0$, $F(2,66) = 0.01$, $F(1,40) = 0.002$, and $F(1,40) = 0.014; p > 0.91$, respectively). Interactions between stimulus type and expression type were found for all conditions ($F(2,66) = 4.51; p < 0.015$, $F(2,66) = 3.15; p = 0.049$, $F(1,40) = 8.31; p < 0.007$, respectively), but not for the human 30 degrees condition ($F(1,40) = 0.735; p > 0.39$).

In summary, there is a strong tendency for species-specific expression tuning to be narrower for human *own-species* expressions, while this tendency is less prominent in rotated views.

8.1.3 Robustness of results against variations of species-specific features

One might question the robustness of the previous observations when considering variations in the chosen stimuli. For instance, monkey facial movements encompass species-specific features, such as ear motion, which are absent in human expressions. Do the observed differences in the recognition of human and monkey expressions depend on these features? To address this question, we conducted a follow-up experiment where we presented only the front view of the stimuli to a new set of participants, while occluding the ear region. Figure 8.4A illustrates the corresponding fitted discriminant functions, which exhibit remarkable similarity to the functions without occlusion. Once again, we observe a high resemblance in shape between the human and monkey avatars (ratio of

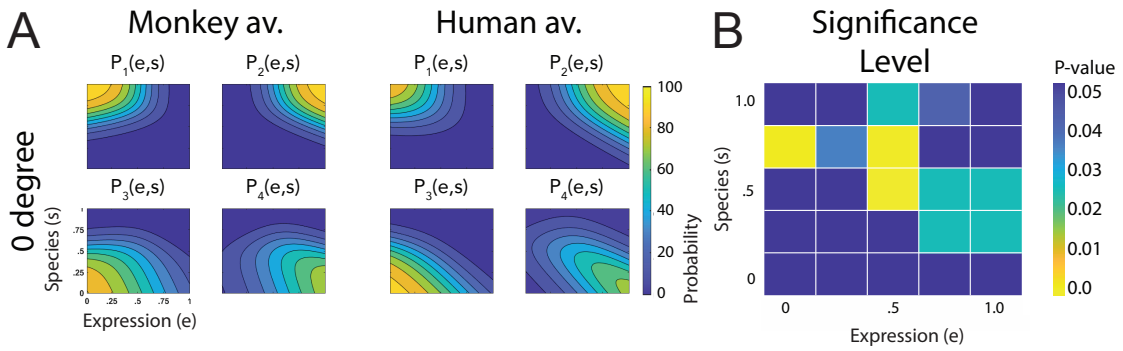

Figure 8.4: Fitted discriminant functions $P_i(e,s)$ for the condition with occlusions of the ears. Classes correspond to the four prototype motions, as specified in Figure 7.3D (i = 1: human-angry, 2: human-fear, 3: monkey-threat, 4: monkey-fear). (A) Results for the stimuli generated using original motion-captured expressions of humans and monkeys as prototypes but with occluded ears, for presentation on a monkey and a human avatar (only using the front view). (B) Significance levels (Bonferroni-corrected) of the differences between the multinomially distributed classification responses for the 25 motion patterns, presented on the monkey and human avatar.

different vs. shared variance: $q = 1.44\%$). Only 12% of the categorization responses among the 25 points in morphing space significantly differed between the two avatar types ($p = 0.02$; Figure 8.4B). Furthermore, Figure 8.3A demonstrates that the corresponding tuning functions, D_M and D_H, closely resemble those for the non-occluded stimuli, and the associated threshold values (Figure 8.3B) exhibit no significant difference compared to the non-occluded stimuli (see A.4). In summary, the removal of ear motion as a monkey-specific feature did not significantly impact the main findings of the original experiment.

8.1.4 Robustness against variations of expression strength

A further possible concern arises regarding whether the chosen prototypical expressions specify varying amounts of salient low-level features, possibly due to species differences in motion or anatomical disparities between the human and monkey face. To address this concern and control for the influence of expressive low-level information, we conducted a reiteration of the main experiment using stimuli that were equilibrated (balanced) in terms of the amount of such expressive information 7.1.3.

The equilibration procedure was based on a pilot experiment that compared different equilibration methods utilizing various measures for low-level information. These measures included the total amount of optic flow (OF), the maximum deformation of the polygon mesh during the expression (DF), and the total motion flow of the polygon mesh during the expression (MF). In the control experiment, nine participants rated these

equilibrated stimulus sets based on the perceived expression strength of their motion, disregarding the avatar type. Ratings of perceived expression strength were obtained using a nine-point Likert scale (1: non-expressive, 9: very expressive), with each stimulus presented in a block-randomized manner four times.

The average ratings comparing the different low-level measures are presented in Figure 7.4A. Additionally, this figure displays the ratings for the neutral expression (which were very low) and the ratings for the original non-equilibrated expressions. Balancing the amount of polygon motion (*MF*) resulted in the lowest standard deviation of expression strength ratings after equilibration (except for the neutral condition $F(1,479); p < 0.021$).

Specifically, for the human avatar, the *MF* condition exhibit smaller variance in perceived expression strength compared to the *DF* conditions ($F(1,142) = 1.479; p < 0.021$). Similarly, for the monkey avatar, the variance is smaller compared to all other conditions ($F(1.403); p < 0.045$), except for the *DF* condition ($F(1,142) = 0.869; p > 0.407$). Moreover, the difference in perceived expressiveness between the two avatars is not significant ($t(283) = 0.937; p > 0:349$) when equilibrated using the *DF* measure. Therefore, considering these reasons, we chose *MF* as the measure for equilibrating the prototype motion in our main experiment (a more extensive analysis of these data and additional tested measures for low-level expressive information are discussed in A.6).

Equilibration involved creating morphs between the original motion-captured expressions and a neutral expression, adjusting the morphing weight of the neutral expression to match the amount of motion flow (7.1.3). Equilibration is performed separately for the two avatars and for different view conditions. Figure 7.4B illustrates an example of the effect of equilibration on the extreme frames of a monkey-threat expression. Equilibration also reduces the highly salient mouth opening motion of the monkey, which cannot be replicated by a real human face due to anatomical differences. The effectiveness of the procedure in balancing the amount of motion information is demonstrated in Figure 7.4C. It displays the motion flow before and after equilibration for various points in our motion style space for the front view.

The standard deviation of motion flow across the 25 conditions in style space is reduced by 83% for the monkey avatar and by 54% for the human avatar through equilibration. When focusing the flow analysis on the mouth region, we observed that the standard deviation of motion flow across conditions decreased by 79% for the monkey avatar and by 59% for the human avatar (results for other view conditions are similar).

The fitted discriminant functions for the data from the repetition of the experiment with equilibrated stimuli are shown in Figure 8.5A. These functions exhibit more symmetry along the axes of the morphing space compared to the original stimuli. For instance, confusions between human anger and monkey fear expressions, particularly for intermediate style parameters, are reduced, especially for the human avatar. This reduction may be due to the subtlety of the monkey fear expression. This observation is supported by the significant decrease in the asymmetry index (*AsI*), which measures the deviation from perfect symmetry with respect to the *e* and *s* axes (see A.5), for

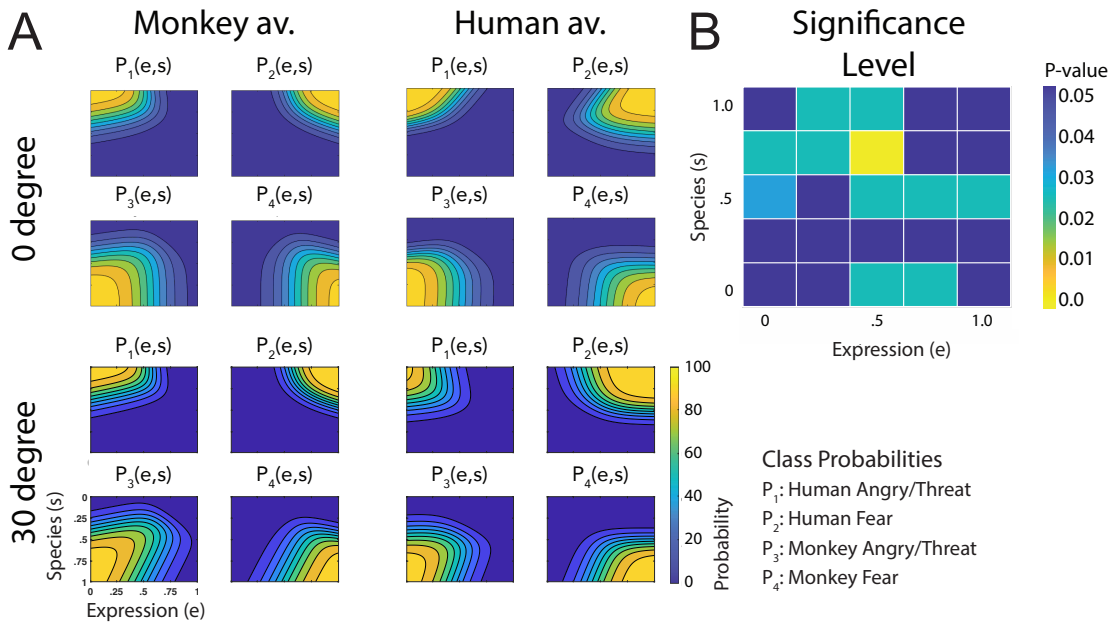

Figure 8.5: Fitted discriminant functions Pi(e,s) for the experiment with equilibration of expressive information. Classes correspond to the four prototype motions, as specified in Figure 7.3D (i = 1: human-angry, 2: human-fear, 3: monkey-threat, 4: monkey-fear). (A) Results for the stimuli set derived from prototype motions that were equilibrated with respect to the amount of local motion/deformation information, for presentation on a monkey and a human avatar. (B) Significance levels (Bonferroni-corrected) of the differences between the multinomially distributed classification responses for the 25 motion patterns, presented on the monkey and human avatar.

the equilibrated stimuli compared to the original motion prototypes ($AsI_{original} = 0.624$ vs. $AsI_{equilibrated} = 0.504$). The difference is statistically significant according to the Wilcoxon signed-rank test ($Z = 2.49; p < 0.013$). The discriminant functions also exhibit a higher similarity between the two avatar types and different view conditions compared to the original stimuli. The ratios of different vs. shared variance between the conditions are small ($q = 4.01\%$), with only 4% of the categorization responses across the 25 points in morphing space being significantly different between avatar types and view conditions, as determined by a contingency table analysis (Figure 8.5B).

Importantly, even for these equilibrated stimulus sets, we find a narrower tuning for human dynamic expressions compared to monkey expressions (Figure 8.3A). This is supported by the results of the ANOVA for the threshold points of the tuning functions $D_M(s)$ and $D_H(s)$ (Figure 8.3B), which do not show a significant influence of the stimulus type (original vs. occluded vs. equilibrated stimuli).

The steepness of the fitted threshold functions is analyzed in Figure 8.3C. This analysis reveals that the equilibration procedure effectively balances the steepness of the

tuning functions between human and monkey expressions, which is evident in the non-equilibrated stimuli. This observation is confirmed by two-way ANOVAs for the original motion stimuli and stimuli with occluded ears, showing significant effects of the avatar type/view factor ($F(3,83) = 12.76; p < 0:006$; and $F(1,39) = 3.33; p < 0.077$, respectively), but no significant effects of the expression type ($F(3,83) = 0.01$ and $F(1,39) = 0.01; p > 0.92$), and no interactions. In contrast, the ANOVA for stimuli with equilibrated motion does not reveal any significant effects of the avatar type/view ($F(3,87) = 1.27; p > 0.26$), the expression type ($F(3.87) = 0.03; p > 0.86$), or an interaction (full ANOVA results in A.7).

In summary, these results demonstrate that the high similarity of the classification data between the two avatar types and different view conditions remains unchanged when the expressiveness of the stimuli is controlled. Additionally, the tendency for a narrower tuning for human own-species expressions is robust against this manipulation. However, balancing the expressiveness eliminates the differences in the steepness of the computed species-tuning functions. Thus, the objection that the observed effects are merely a result of differences in the amount of low-level salient features in the chosen prototypical motion patterns is invalidated.

8.1.5 Discussion of behavioural results for model design

In these results, we employ advanced methods from computer animation, motion capture across species, and machine learning for motion interpolation to gain fundamental insights into the perceptual encoding of dynamic facial expressions across primate species.

Our first key finding is that human observers quickly learn facial expressions of macaque monkeys, despite the distinct differences between monkey and human expressions, which may not be readily interpretable by naive observers. This rapid learning could be attributed to the high similarity in the neuromuscular control of facial movements in humans and macaques Parr *et al.* (2010), resulting in comparable structural properties of expression dynamics that the visual system can exploit for efficient learning.

Surprisingly, contrary to shape-based accounts of dynamic expression recognition, we discover that categorization of dynamic facial expressions is remarkably independent of primate face type (human vs. monkey) and stimulus view (0 vs. 30 degrees rotation of the head about the vertical axis). This independence demonstrates a significant degree of invariance to changes in two-dimensional image features. Specifically, we observe no substantial differences in categorization responses dependent on these parameters, nor do we find a superior perceptual representation of species-specific dynamic expressions corresponding to the avatar's species (e.g., more accurate representation of human expressions on the human avatar or monkey expressions on the monkey avatar). Thus, facial expression dynamics appear to be represented independently of detailed shape features of the primate head and stimulus view.

However, we do find a clear and highly robust advantage for own-species expressions

in terms of the accuracy of tuning for expression dynamics Scott and Fava (2013); Pascalis *et al.* (2005): The tuning along the species axis of our motion style space is narrower for human expressions compared to monkey expressions. This observation remains consistent even for stimuli that eliminate species-specific features, such as ear motion, or stimuli that are carefully equilibrated in terms of low-level information.

Both key findings support our initial hypotheses: Perception can exploit the structural similarity of dynamic expressions across different primate species for rapid learning. Furthermore, consistent with the co-evolution of visual processing of dynamic facial expressions and their motor control, we observe a largely independent encoding of facial expression dynamics from basic facial shape in primate expressions. However, to confirm this independence and generalize our findings beyond primate faces, more extensive experiments including a broader range of facial shapes and potentially non-primate faces are required.

The observed independence between basic facial shape and expression encoding aligns with results from functional imaging studies that suggest a modular representation of different aspects of faces, including changeable and non-changeable features Haxby *et al.* (2000); Bernstein and Yovel (2015); Dobs *et al.* (2019). Nevertheless, our experiments are challenging to reconcile with popular (recurrent) neural network models that represent facial expressions as sequences of learned key shapes Curio *et al.* (2011); Li and Deng (2020). Given that shape differences between human and monkey faces are much larger than those between keyframes of the same expression, it is difficult for such models to account for the observed innate generalization of dynamic expressions to faces from a different primate species.

8.2 Norm-referenced encoding encode facial expression recognition

As mentioned earlier, our visual system possesses a remarkable capability: the innate recognition of facial expressions even on non-human faces, such as cartoon characters. In our behavioral experiment with different primate faces Taubert *et al.* (2021), we provided evidence for this ability. Based on these findings, we can identify two key requirements. First, the perception of facial expressions by humans is largely unaffected by variations in face shapes. Second, participants exhibit rapid learning. In section (3), I discussed the concept of norm-based reference encoding (NRE), a mechanism proposed for facial identity recognition, as a potential mechanism underlying this phenomenon. In the following results, I will establish the validity of NRE as a computer vision mechanism for facial expression recognition (FER).

In this result, I show how the norm-referenced encoding performs on our stimuli set developed for the monkey electrophysiological recording 7.1.2. I compare the NRE model (model A 4.2), referred as Norm-Based mechanisms, with a biologically-inspired stan-

dard RNN model, referred as *snapshot neurons* (describe in the Appendix A.8). These results are published in Stettler *et al.* (2020).

8.2.1 Norm-referenced encoding Simulation

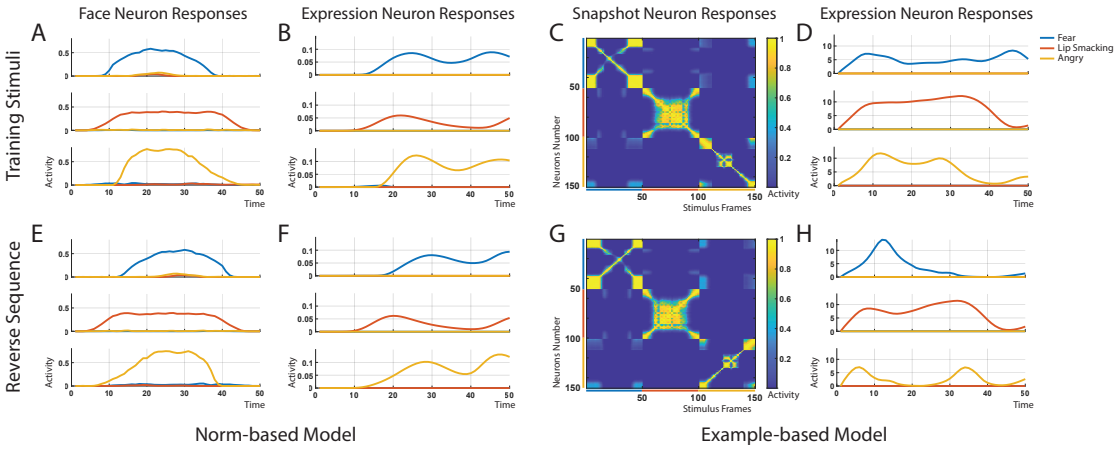

Figure 8.6: Activity of model neurons. Presentation of the prototypical expressions: Fear, Lip Smacking and Angry expressions. Upper panels shows data for the original, and the lower panels the reversely played expression movies. Norm-referenced model: A, E) Face neurons, and B, F) Expression Neurons. Example-referenced model: C, G) Snapshot neurons, and D,H) Expression Neurons.

We first test whether the models can accurately classify the prototypical expressions used during training. Figure 8.6A displays the responses of *face neurons* that exhibit selectivity for individual expressions. These neurons demonstrate a bell-shaped increase and decrease in activity that corresponds to the encoded expression. The response of the corresponding *expression neurons*, shown in Panel B, also exhibits selectivity for the individual expressions. Unlike the *face neurons*, these neurons remain inactive when presented with static images depicting the extreme frames of the expressions. Panels C and D illustrate the responses of the example-referenced model. Panel C presents the activity of the *snapshot neurons* for the three test expressions. Only the frames associated with the learned expression generate a traveling pulse solution within the corresponding part of the recurrent neural network (RNN), while the neurons remain silent for the other test patterns. Consequently, this induces a high level of selectivity in the responses of the corresponding *expression neurons* (Panel D).

We also conduct tests on the model using temporally reversed face sequences to examine sequence selectivity. The results are depicted in Figure 8.6E-H. Due to the high temporal symmetry of facial expressions (backward-played expressions appear very similar, though not identical, to forward-played expressions), the responses of the *face neurons* in the norm-referenced model, as well as those of the *expression neurons*, closely

resemble the responses shown in Panels A-D. Notably, the *snapshot neurons* now exhibit a traveling pulse that propagates in the opposite direction. The minor differences between forward and backward sequences result in slightly lower response amplitudes in the *expression neurons*, particularly noticeable in the example-based model (Panel H), but interestingly also present in the norm-referenced model.

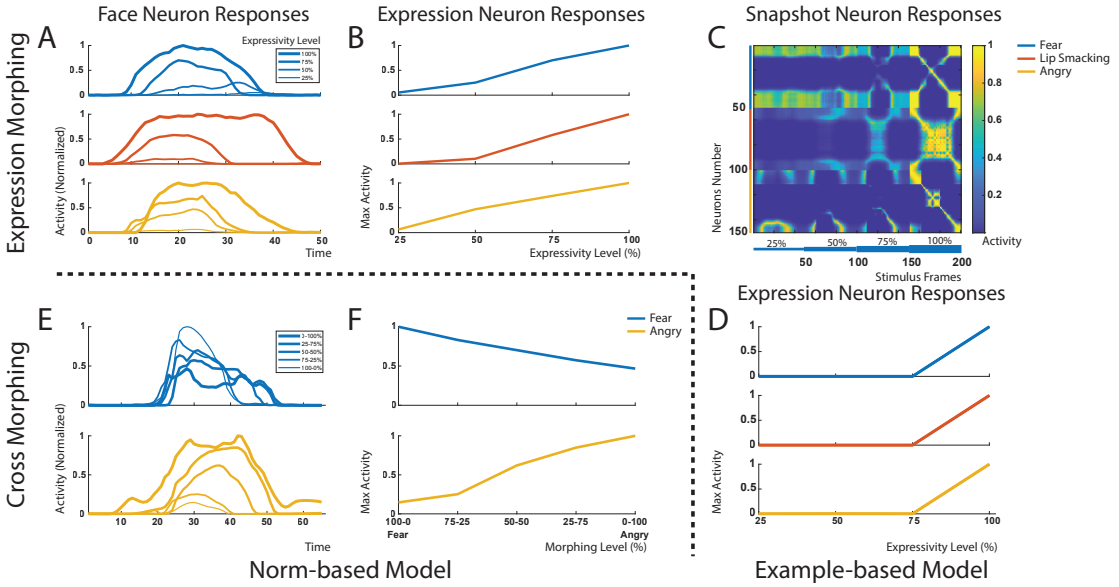

Figure 8.7: Upper panels: Neuron activities for stimuli with different expression strength levels (25-50-75-100%). Lower panels: Neuron activities for cross-expression morphs of the neurons of the norm-referenced model. A) Normalised activity of Face neurons, and B) maximum activity of Expression neurons as function of the expression strength level for the norm-referenced model. C) Normalised activities of snapshot neurons, and D) maximum activity of Expression neurons in example-based model. E) Normalised activity of Face neurons, and F) maximum activity of Expression neurons as function of the morph level form morphs between Fear and Angry expressions.

Interesting differential predictions emerge when testing the models with stimuli of varying expression strength, created by morphing between the prototypes and neutral facial expressions. In the norm-referenced model, both the *face neurons* and the *expression neurons* exhibit a gradual, nearly linear variation in their activation levels with the expression level (Figure 8.7A and B). However, the behavior of the *snapshot neurons* deviates significantly from this pattern. These neurons do not generate a traveling activity pulse for all stimuli with reduced expression strength levels. Only at the expression level of 75%, some activity emerges in the *snapshot neurons* representing frames that deviate from the neutral expression. As a result, the expression neurons do not show significant activity for conditions with reduced expression strength. Attempts to improve this behavior by reducing the selectivity of the snapshots to support generalization to more

dissimilar patterns were unsuccessful. Thus, it was not possible for this model to achieve both pattern- and sequence-selectivity while also generalizing to patterns with reduced expression strength.

The norm-referenced model also demonstrates very smooth and gradual generalization between different expressions. This is evident in Figure 8.7E-F, which displays the responses of the Face and Expression neurons for morphs between the *Fear* and *Angry* expressions. Both types of neurons display a smooth change in activity as the morph level progresses, with an antagonistic relationship between neurons selective for the two expressions. In this case as well, the example-based model fails to exhibit generalization to stimuli with intermediate morph levels (not shown).

8.2.2 Discussion on NRE as a valid FER model

Based on previous models that are grounded in electrophysiological data, we have proposed two alternative mechanisms for the processing of dynamic facial expressions. Both mechanisms are consistent with physiological data from other cortical structures that process social stimuli, static faces and dynamic bodies. Both models were able to recognize monkey expressions from movies. Also the recognition of reversed movies of facial expressions could be accounted for by both models. Testing the models with morphed expressions, and expressions with reduced expression strength, however, resulted in fundamentally different predictions. The norm-based model showed smooth and almost linear variation of the activity patterns with the expression strength and the morph level, while the example-based model had problems to generalize to such morphed patterns. In addition, the models make specific predictions about the activity dynamics of the different postulated neuron classes. For example, an example-based mechanism predicts a traveling pulse of activity, as observed e.g. in premotor cortex Caggiano *et al.* (2016). The norm-based mechanism predicts a linear tuning of the activity with the distance from the neutral reference pattern in morphing space for the Face neurons.

8.3 Multi-Domain Norm-Referenced Encoding

In 8.2, I have demonstrated the validity of NRE as a Facial Expression Recognition (FER) model. Subsequently, we delve into our innovative multi-domain norm-referenced encoding (MD-NRE) framework (refer to Section 3.1). The primary objective of these results is to assess the data efficiency of MD-NRE through the utilization of transfer learning. Our experimental setup closely resembles domain adaptation, wherein we initially train the model on one domain and then transfer the learned categories to other domains by introducing a single new reference for each domain. These results are drawn from Stettler *et al.* (2023a), employing Model E (refer to Section 4.6) on a publicly available dataset (refer to Chapter 7.2).

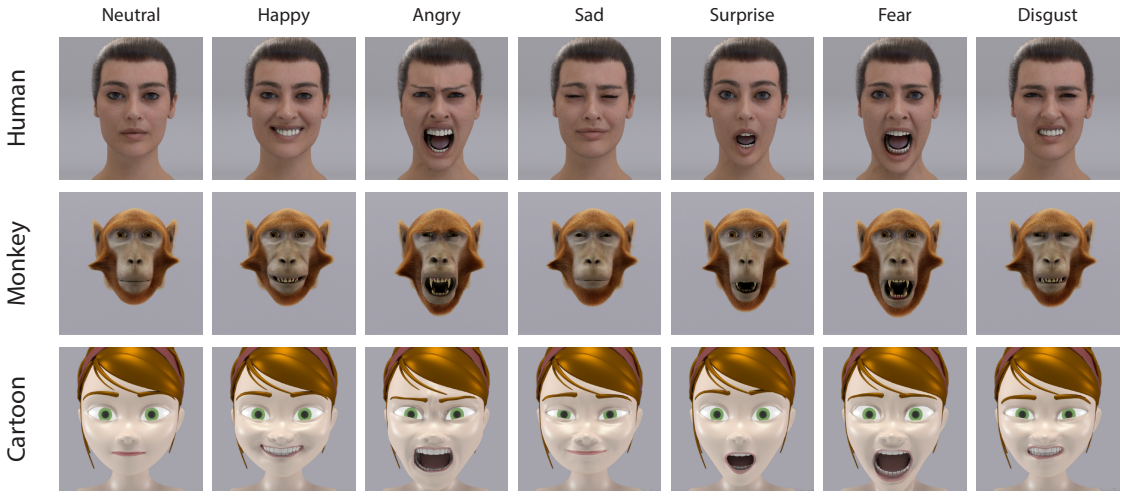

Figure 8.8: Examples of the portraits that form our basic face shape (BFS) dataset showing seven different expressions on the three basic head shapes-what we refer to as our source and target domains. Our task involves the classification of facial expressions across unseen target domains from a source domain (outlined in blue) using only a single image from the target domain as a reference (outlined in red and green).

The motivation for this work stems from a fundamental observation: existing state-of-the-art FER models often fail to generalize or transfer well to diverse domains, such as cartoon faces 8.3.1. Leveraging MD-NRE, we aim to showcase its advantages in computer vision. By extending NRE to multiple domains and developing a model that capitalizes on this encoding, we take a significant step towards establishing an end-to-end framework.

8.3.1 Domain Generalisation

Using our BFS dataset 7.1.3, we test the MD-NRE for *domain generalization*. In this test, we learn facial expressions from a source domain (human) and transfer the tuning vectors to the other domains by updating the reference frames. Here, domains refer to each basic head shape of the BFS dataset (human, monkey, and cartoon). As such, our model has never seen any expressions from the target domains. Nonetheless, the model has knowledge that the target domains exist from the training of the reference frame.

To conduct our test, we select one image of each expression displayed on the human avatar (blue rectangle from Figure 7.5). We use three images, one per domain, to train the reference vectors, and 7 images, one per category, to train the tuning vectors. However, as the neutral expression on the human identity is already selected for the training of the human reference vector, we effectively add only 6 images to the training set. Consequently, we train the classifier using a total of 9 images (training ratio: 8.57%).

Nonetheless, it is fair to wonder if our BFS cross-domain dataset is too similar to

Model	Type	Dataset	# im	Acc
SVM-LBP Luo *et al.* (2013)	SVM	BFS	9	14.3
ResNet50v2 Ngo and Yoon (2020)	CNN	Affectnet	280k+	52.4
CORNet-S	CNN	Affectnet	280k+	63.1
DAN Wen *et al.* (2021)	CNN	AffectNet	440k	37.4
EfficientFace Zhao *et al.* (2021)	CNN	AffectNet	440k	33.8
MD-NRE		BFS	9	**100**

Table 8.1: Classification accuracies (%) for our MD-NRE model and different standard CNN architectures.

human head shapes. The question emerges whether state-of-the-art FER models can accurately classify our non-human head shapes. Therefore, we test the performance of multiple available FER architectures and trained a CORNet-S model for comparisons (as it is argued to be a closer brain-like architecture) Kubilius *et al.* (2018). (See A.8.1 for implementation and training details.)

We obtain perfect classification accuracy (100%) for the test dataset. This implies a perfect transfer of the learned expressions from one human identity to all the other head shapes, without a need for training of the expressions on the non-human head shapes (Table 8.1). This result proves our assumption that facial expression tuning vectors can be transferred across different head shapes using only a single neutral face shape within an appropriate feature space. It also proves that our model reliably tracks the position of facial landmarks independently of face shapes and textures.

Table 8.1 shows that none of the other models performed as well as ours on the transfer learning task. The CORNet-S model reaches the best accuracy with a limited performance of 63.1%, and the SVM model performs the worst, only reaching chance level classification performance (14.3%). See A.10 for a detailed discussion of this poor result.

8.3.2 Analogous Encoding

The second test aims at assessing the second assumption of the model. Contrary to the output layers of typical DNN classification models, which return estimates for class probabilities (softmax readout), face-selective cortical neurons show an almost linear dependence of activity on the distance of the face from the average face Leopold *et al.* (2006); Koyano *et al.* (2021); Freiwald and Hosoya (2021). Likewise, in our architecture, the output variables v_m depend linearly on the length of the difference vectors $\mathbf{d}$, which covary with expression strength. Such *analogous encoding* is especially interesting for social interaction, where even small and subtle facial movements might be used to perceive a person's intent. This principle might underlie our ability to perceive complex and subtle facial movements such as a smirk. Furthermore, encoding of expression strength seems useful for different technical applications. One example would be a

Model	BFS-L	Human	Monkey	Cartoon
CORnet-S	46	62.5	44	32
MD-NRE	**78**	**92**	**68**	**76**

Table 8.2: Classification accuracy (%) for our MD-NRE and the CORnetS-A on the expression strength level test (BFS-L) for each head shapes.

human closed-loop interaction system, such as using facial expression recognition data to drive computer animations, where expression strength is often modeled by a superposition of action units or blendshapes. In addition, expression strength influences the reliability of expression categorization: weak expressions are more difficult to classify than extreme expressions.

We test the linearity of the encoding by leveraging our BFS dataset to simulate expression strength for each basic head shape. For this purpose, we extended the BFS dataset by adding versions of all expressions with three extra levels (25%, 50%, and 75%) of expression strength. These intermediate levels were created by blending (linearly scaling the displacements between the corresponding mesh polygons between the neutral and the 100 % expressive face for the same head shape). The resulting dataset (BFS-L) has a total of 75 testing images 7.4. We use the same 9 training images as in the previous validation experiments (training ratio: 2.07%).

Figure 8.9 compares the readout activity values of our MD-NRE and the CORNetS models (the best CNN model from the previous experiment). While both models show a monotonic increase of the output variables with the expression strength, our model shows greater linearity. Moreover, Table 8.2 shows that the classification accuracy (collapsed over all tested expression strength levels) of our model outperforms the CORNetS architecture, indicating higher robustness to variations of expression strength. Nonetheless, as expected, the overall performance for the MD-NRE model is reduced compared to using only the 100% expression strength test images (Table 8.1). This accuracy reduction is mainly caused by a reduced performance on the weaker expression strength level images (see A.9 for more details).

8.3.3 Efficient Learning

The third test aims to scale up the model and to conduct a first benchmark test of our approach with the literature. We select the FERG Aneja *et al.* (2016) dataset which consists of stylized comic characters with annotated facial expressions. The FERG dataset is suitable for our simple landmark detector as the pose and the illumination are fixed. The dataset contains 55,767 images across 6 different head shapes and 7 expressions. The dataset is split into a training set (43,767), a validation set (6k), and a test set (6K).

To apply our MD-NRE, we train a reference vector for each avatar identity and test our model's learning efficiency. This task is thus an assessment of the benefit of using

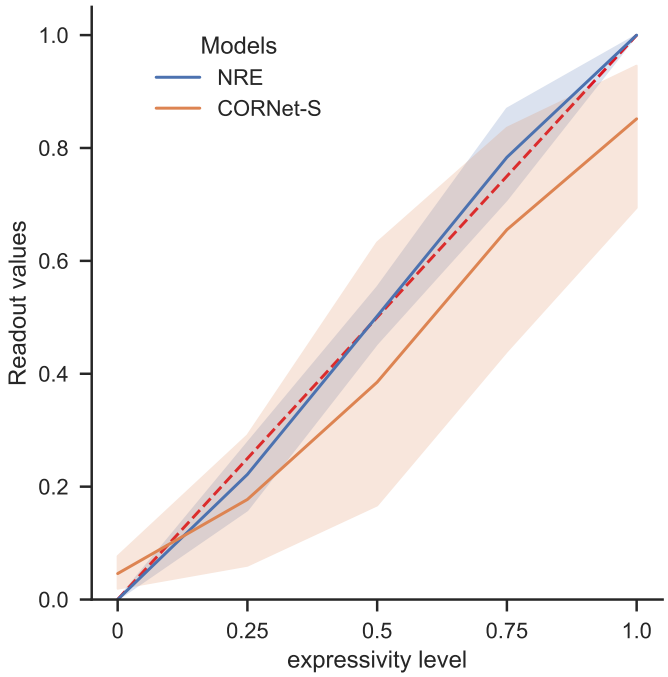

Figure 8.9: Readout values of the multi-domain norm-referenced encoding (MD-NRE) and the CORnetS-A model as a function of the expression strength level for each expression on the Human Avatar. The red dashed line corresponds to perfect linearity.

reference frames as inductive bias. Nonetheless, the variance of the head shapes among the avatars of the FERG dataset is arguably higher than among human identities. Therefore, we also test our method to transfer facial expressions from a single avatar head to the others.

We use 12 images to train our models (training ratio: 0.02%). One neutral pose for each avatar type (6 images) to train the reference vectors, and one for each expression (7 images) to train the tuning vectors. As before, the neutral expression is already included in the set used to train the reference vectors.

We test the efficiency of our model using two approaches. First, we train our model as a standard FER approach. Therefore, the tuning vectors are trained independently of the avatar type (referred to as MD-NRE-I). Here the model can select any expression that might help classification accuracy. Second, we train our classifier using a transfer learning task by using only a single avatar to train the tuning vectors, referred to by MD-NRE-<avatar_name>. Here, the models learn facial expressions from the subset of images belonging to the selected avatar.

In addition, we select the training expressions in two ways: In one condition, indicated by *First*, we use the first picture of the relevant expression class for training. This has the disadvantage that this first picture might not be the most typical one for that expression

Models	First		Optimized		# im
	Train	Test	Train	Test	
DeepExpr Aneja *et al.* (2016)	-	-	-	89	43k+
ACNN Minaee *et al.* (2021)	-	-	-	99.3	43k+
VGAN Chen *et al.* (2018)	-	-	-	100	43k+
MD-NRE-I	73.1	73.1	92.3	**92.2**	**12**
MD-NRE-Aia	76.7	76.4	84.6	84.2	12
MD-NRE-Bonnie	70.4	70.4	85	84.4	12
MD-NRE-Jules	80.6	79.9	80.5	79.7	12
MD-NRE-Malcolm	73.3	73.3	80.4	80	12
MD-NRE-Mery	71.6	71,1	81.9	82.2	12
MD-NRE-Ray	66.6	66.5	71.9	71.6	12

Table 8.3: Classification accuracy (%) for our MD-NRE models on the FERG dataset and the reported test classification accuracy from different FER models trained on the FERG dataset

Models	First		Optimized		# im
	Train	Test	Train	Test	
NRE-Aia	32.8	32.7	31.2	31.2	7
NRE-Bonnie	29.9	30.3	37.9	38.1	7
NRE-Jules	39.6	39.1	**42.4**	42.0	7
NRE-Malcolm	**39.7**	**39.4**	40.5	39.6	7
NRE-Mery	36.5	37.3	37.9	39.4	7
NRE-Ray	36.9	36.0	41.2	**42.1**	7

Table 8.4: Classification accuracy (%) for our NRE model on the FERG dataset and the reported test classification accuracy from different FER model trained on the FERG dataset

category. To remove this limitation, we perform a second set of tests where we choose for each class an optimized template stimulus (indicated by *Optimized*). This stimulus is determined by iterating through the training dataset and retaining the class-specific training picture that results in the highest training accuracy after a single training epoch.

Table 8.3 shows the accuracy of the classifications for the different training schemes. Using the first images in the dataset for each class (condition *First*), we obtain a testing accuracy of 73%. This was in the range of the accuracies for training with individual avatars, which range from 66% to 79%. While these results are not state-of-the-art (100% reported by Chen *et al.* (2018)), this test proves that our model can cope with a larger dataset. When the model have access to every domain (MD-NRE-I), the test accuracy reaches a value of 92.15%, which exceeds the accuracy (89.02%) reported in the original paper on the FERG dataset Aneja *et al.* (2016). Moreover, the best transfer learning model (MD-NRE-Bonnie) reaches 84.42% accuracy. This is an encouraging transfer learning result for a classifier trained with only 12 images, and where all the expressions are from a single avatar type.

Models	Eyes	Mouth	Left	Right	# im
MD-NRE-I	**72.6**	**55.3**	**84.6**	**81.6**	12
MD-NRE-Aia	64.6	40.7	73.9	68.4	12
MD-NRE-Bonnie	64.0	46.4	73.3	71.5	12
MD-NRE-Jules	67.6	48.7	73.3	74.1	12
MD-NRE-Malcolm	59.2	48.8	76.6	73.3	12
MD-NRE-Mery	54.5	47.6	64.3	64.3	12
MD-NRE-Ray	63.4	45.5	72.3	59.4	12

Table 8.5: Classification accuracy (%) for our MD-NRE model on the FERG dataset using subset of landmarks

Reference Frames

In our *Efficient Learning* experiment, we learn a reference frame for each avatar of the FERG dataset. This means that the model knows how many avatars exist and has an internal representation of each avatar identity. In this first ablation study, we investigate the effect of removing these reference vectors. As such, this test uses (single-domain) NRE models. We train our models in a transfer learning scheme using only the subset of the selected avatar. This is ecologically the most valid, since it mimics the observation of multiple expressions on the same agent, without observing other agents at all.

Table 8.4 shows the results of training NRE models (using a single reference frame) for all cartoon characters. As expected, this change reduced the classification performances of the models. The best model (NRE-Ray) achieves only 42.1% test accuracy (-54.3%). This result remains largely above chance level (14.3%), which is not terrible considering that the model is trained with only 7 images from a single avatar. Nonetheless, this test demonstrates the crucial role of domain-specific reference vectors for relative encoding.

Robustness to Occlusion

Our visual system perceives facial expression even if parts of the face are occluded. Therefore, in this second ablation study, we investigate how the models perform with a reduced set of facial landmarks. While 10 landmarks are already a sparse set, we crop it further into a subset displaying only the eyes (6 landmarks), mouth (4 landmarks), or left/right side of the face (6 landmarks). Due to the symmetric nature of the face, we expect moderate loss of accuracy on the left/right test, and stronger performance degradation for the removal of eyes and mouth as both convey important facial expression cues.

Validating our assumption, occluding the face's left or right side partially impairs our model's categorization accuracy. The best model (MD-NRE-I) achieves 84.6%, loosing 7.6% accuracy to our best model (See *Efficient Learning* 8.3.3). In contrast, occluding the eyes or the mouth results in stronger impairment (72.6% and 55.3& accuracy respectively). The results show that the eyes and eyebrows region accounts for more discriminating features in the recognition of facial expressions than the mouth region. These

Models	FAN Bulat and Tzimiropoulos (2017)	MediaPipe Kartynnik *et al.* (2019)
NRE-I	64.6	71.1
NRE-Aia	60.3	68.1
MD-NRE-Bonnie	69.8	66.8
MD-NRE-Jules	62.7	**73.0**
MD-NRE-Malcolm	70.6	48.9
MD-NRE-Mery	**76.5**	71.6
MD-NRE-Ray	64.5	67.7

Table 8.6: Classification accuracy (%) for our NRE model on the FERG dataset and the reported test classification accuracy from different FER model trained on the FERG dataset

results demonstrate that MD-NRE can still classify with high accuracy in the presence of partial occlusions. Good performance for occlusion of left/right facial landmarks can be explained by redundancy induced by the model's inductive bias. Occlusion of eyes or mouth impact performance more strongly, but the accuracy is far above chance (14.3%) even if more than half the landmarks are missing.

Landmark Detectors

Our landmark detector is an important module of our architecture. We demonstrate how our specialized architecture provides invariant facial expression features across various face shapes and textures (See *Domain Generalization* 8.3.1). In this third ablation study, we investigate the robustness of our model against different landmark detectors from the literature (*e.g.*, Bulat and Tzimiropoulos (2017); Kartynnik *et al.* (2019)) to infer the landmark positions on the FERG dataset. (See A.11 for an illustration of their position inference.)

In particular, we assess the robustness of the tuning vectors against being fed with sub-optimal landmarks. Relying on human facial landmark detectors biases the landmark positions to fit a human face shape. However, learning separate reference vectors for the different head shapes allows us to subtract the offset generated by the head shape in the feature space in order to determine the accuracy of the tracing of the individual features across different head shapes. To conduct this test, we keep the *FR pathway* from our pipeline to update the reference frames.

Table 8.6 shows the effect of using pre-trained landmark detectors as input to our NRE. The accuracy ranges from 48.9% (-43.3%) for the worst (NRE-Malcolm MediaPipe) to 76.5 (-15.7%) for the best model (NRE-Mery FAN). These results show that the NRE is impacted when operating in a less optimal feature space, but as previously shown with the occlusion study 8.3.3, the results demonstrate a substantial level of robustness since the classification performance remains well above chance (14.3%).

8.3.4 Discussion on Computer Vision

With these benchmarking results using a computer vision approach, I presented a proof-of-concept study of my Model E 4.1 that tests the use of the norm-referenced encoding

(NRE) for multi-domain transfer in facial expression classification. We hypothesize that this encoding principle explains how the human brain accomplishes high data-efficiency in transfer learning. We validate our hypothesis by illustrating how our model recognizes facial expressions on novel head shapes, an easy task for humans.

We demonstrate that this encoding principle can be implemented in the context of a deep network architecture that consists of two streams, one that identifies the type of the head shapes and the corresponding reference vectors, and a second one that computes the difference vectors and the activity of the NRE encoding neurons. The input of these two pathways is given by a landmark detector.

We also demonstrate its linear dependence on deviation from the reference frames makes this encoding especially interesting for tasks requiring a strong estimate of confidence, *e.g.*, human social interactions, and medical applications. Facial expression perception is a prime example: even subtle facial cues might reveal information about emotion. And in many social situations, the expressiveness about the displayed emotion has important behavioral consequences (e.g faked vs. real smiles, etc.). I acknowledge that substantial additional work is required to make these methods applicable to a wider spectrum and more complex machine-learning problems. Nonetheless, this work is a step towards a more generic end-to-end machine learning model exploiting NRE.

8.4 Model Validation

I have demonstrated that NRE serves as a data-efficient 8.3 and valid method for encoding facial expressions 8.2. Moreover, my model exhibits the capability to learn effectively even on a larger dataset and proficiently transfer facial expressions across different head shapes—a task that appears instinctive for humans-but not for standard current computer vision models. Therefore, to close the loop of my project 2.2, the next crucial step is to validate my model. The primary unanswered question revolves around the existence of a similar mechanism in the brain for perceiving facial expressions. To address this inquiry, I will present preliminary results from our experiment involving rhesus macaques. Additionally, I aim to explore whether the mechanism emulates human perception. To tackle this, I will replicate my psychophysical experiment 8.1 using my model E 4.6 and draw comparisons with traditional computer vision models.

8.4.1 Matching Electrophysiological Patterns in Rhesus Macaque

In this study, we conduct a sanity check to confirm the linearity of our latest NRE model E (see Section 4.6). We compare these results with electrophysiological data recorded from our monkey stimuli experiments (see Section 7.1.2).

Recent preliminary results suggest that the activity of face-responsive neurons in macaque superior temporal sulcus (STS) grows linearly with the intensity of a facial expression Siebert *et al.* (2022). This finding is well-aligned with the theory of norm-referenced

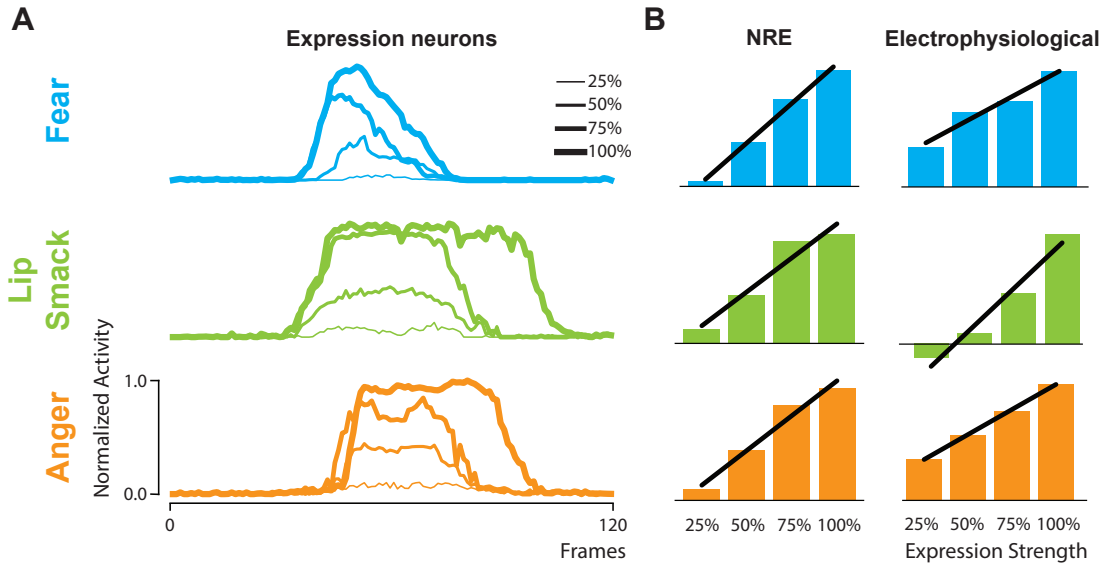

Figure 8.10: A) Predictions of Norm-referenced encoding (NRE) model on stimuli utilized to capture electrophysiological activity in rhesus macaques, for the three distinct facial expressions – Fear, Lip-smack, and Anger – across varying levels of expression intensity. B) A comparison between the peak activities generated by our NRE model and preliminary findings indicating a linear correlation between neural activity and the strength of facial expressions. The black line represents the linear regression over in function of the expression strength.

encoding which predicts linear tuning through a very simple mechanism: Since the activity of expression neuron i in the NRE model is given by $v_i = \mathbf{d}^T \mathbf{n}_i$, the output grows linearly with the deviation of the input from the neutral expression given by $\mathbf{d}$. To test this mechanism in practice and draw comparisons to neural data, we presented the stimuli used by Siebert *et al.* (2022) to the NRE model. These stimuli consisted of videos of macaque faces exhibiting expressions with an intensity of either 25%, 50%, 75% or 100% Results are shown in Fig. 8.10. Panel A displays the normalized output of the expression neuron v_i when confronted with the corresponding expression type. Panel B shows the maximal activity taken across time for each expression and intensity. Panel B is adapted from Siebert *et al.* (2022) which depicts population activity from recordings in macaque STS. Both are normalized to the response to the stimulus with full intensity. The figure reveals that norm-referenced encoding models the linear tuning properties of STS neurons well. One apparent difference is the activity for low expression values: For the 'Fear' and 'Anger' actions, recorded neural activity for the stimuli with 25% intensity was relatively higher than model activations. We hypothesize that this might be due to receptive field sizes in the VGG backbone of the NRE model. Larger receptive fields lead to higher spatial invariance yielding lower sensitivity for subtle facial expressions,

because landmarks are still perceived in the same location.

We acknowledge that more work is required to corroborate the electrophysiological predictions made by the NRE model. However, if these results are consolidated, the linear tuning of the NRE model would add to the growing amount of evidence for norm-referenced encoding in processing of facial expressions.

8.4.2 Matching Human Behavioural Perception

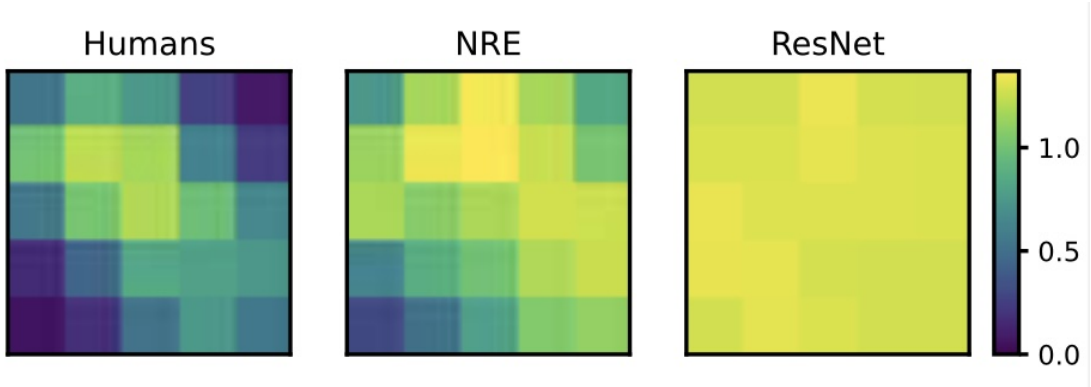

Figure 8.11: Graphical comparison of heat maps representing our 2D morphing space stimuli concatenated across all category prediction from 8.2 using the Human head shape. The color map represents the entropy value, where higher values indicate increased uncertainty. The heat maps shows the Human Avatar head shapes condition between our psychophysical experiment results, our norm-referenced encoding model, and a common computer vision model (ResNet50v2) predictions. As observed in our heat maps, these qualitative results demonstrate how our NRE model excels in capturing uncertainty by displaying patterns that are more similar to the Humans when compared with the ResNet model.

In this section, we test our latest model (Model E, see Section 4.6) using our 2D morphing stimuli, developed to investigate human perception across different face shapes (see Section 7.1.3). Our objective is to validate our model's ability to independently perceive facial expressions based on the basic face shapes identified in our human psychophysics results (see Section 8.1). Similar to our human psychophysics experiment, we train our model using only the four prototypes: human angry, human fear, monkey angry, and monkey fear. We then compare the model's predictions across the entire morphing space with our behavioral results (across every category), using the maximum activity over the length of our dynamic stimuli as the model's prediction.

Figure 8.11 shows the entropy for each location in the morphing space, comparing the predictions of human participants, our NRE model, and a ResNet50 model. The heat maps for human participants illustrate clear distinctions at the corners, where categorization is straightforward, and increased uncertainty at the center of the morphing

space. While the NRE model exhibits higher overall entropy than human participants, it demonstrates a similar pattern, unlike the ResNet50 model with its softmax function.

Building on this experiment, we presented the same stimuli to NRE and several standard computer vision models to test how similar their classification patterns are to those of humans.Tested models include both static and dynamic vision models. All static models models except for NRE were pretrained on either the Affectnet (Mollahosseini *et al.*, 2017) or ImageNet (Deng *et al.*, 2009) dataset and then finetuned on the stimuli used in the psychophysics experiment. Since static models do not directly accept dynamic input, we predicted the class evidence for each frame individually, and then computed the class evidence for the entire video by taking the maximum over the temporal dimension. Finally, we normalized the output to obtain a distribution over expression types. This procedure was used for both static NRE and the static CNNs. Dynamical face models were pretrained on the DFEW dataset (Jiang *et al.*, 2020) and then finetuned using one-shot learning. NRE models were trained only on the psychophysics stimuli. Importantly, only the non-morphed expressions (i.e. the corners of the morphing space) were presented during model training, as human participants were also unfamiliar with morphed expressions.

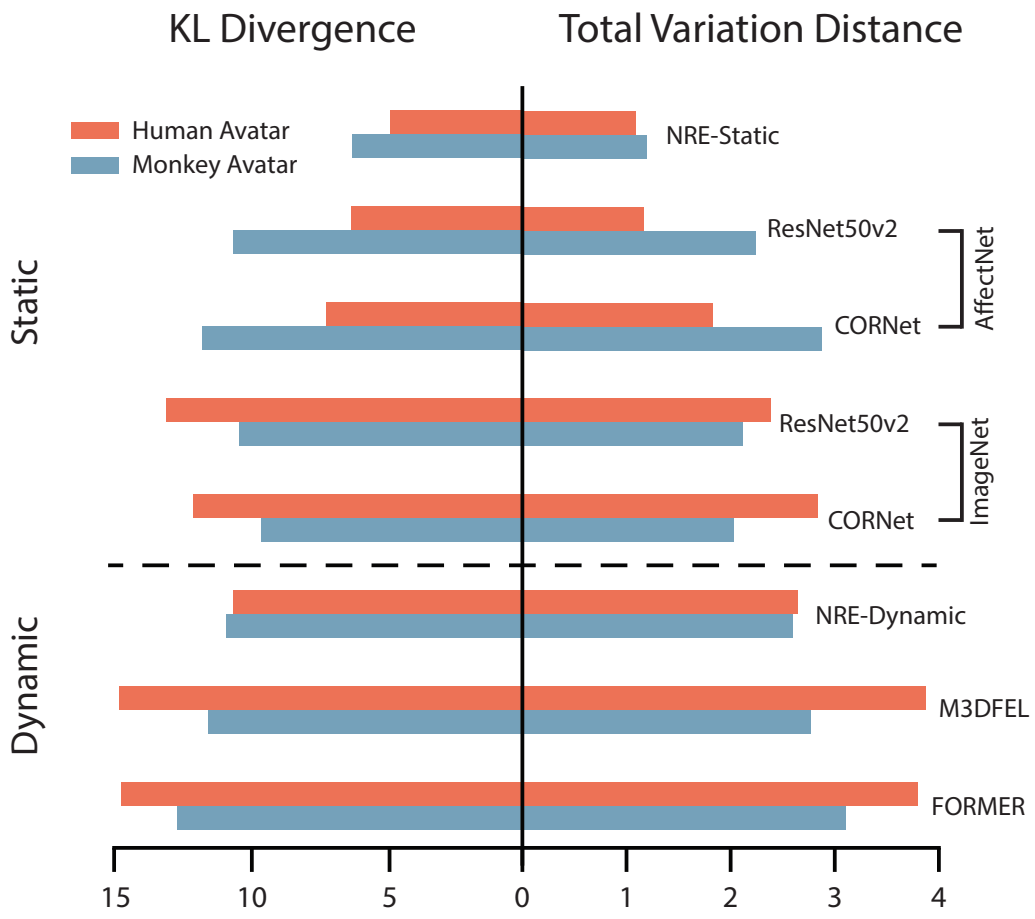

Figure 8.12: Comparison of classification of stimuli from the morphing space between models and humans. For each location in the morphing space, each model outputs a probability distribution over the expression types. We computed pairwise distance metrics between these distributions, and the empirical distribution obtained from humans in our previous experiment. The figure displays the sum of the distance metrics over the entire morphing space. Static NRE outperforms all other tested models.

Figure 8.12 displays a comparison of all tested models by plotting the sum of the KL divergence/total variation distance over the morphing space for both the human and monkey avatar. We observe that the NRE models exhibits lower divergence than all tested CNNs for both avatars. The static NRE (NRE-Static) obtains the closest prediction to human participants with the minimum sum of KL-divergences, obtaining 4.85 and 6.25 for the human and monkey avatars respectively. Affectnet-trained architectures perform similarly as well on the humans, but their performance drops substantially compared to NRE when confronted with the monkey avatar. Evidently, these models are biased towards human faces and fail to replicate the humans' generalization ability. Imagenet-

trained models actually perform slightly better on the monkey faces, even though they have not been trained on faces explicitly. We explain these results as ImageNet contains 10 monkey categories. Surprisingly, static models clearly outclassed all dynamic models we tested, even our dynamic NRE variant. Dynamic models from the computer vision literature showed highest overall deviation from our psychophysics experiment.

Overall, these results show that NRE yields more human-like facial-expression classification patterns than example-based CNNs. This finding reflects that norm-referenced tuning may be useful for interpolating between expressions, as well as facilitating more robust to unknown head shapes. Surprisingly, static models clearly matched psychophysics results better than their dynamic counterparts, indicating that the presented approaches for temporal modelling do not lead to better explanations of human perception on this stimulus set. Of course, we do not draw the conclusion that static models inherently fit human perception better than dynamic ones, as alternative approaches for temporal modelling might substantially improve the model fits. We simply observe that among the methods we tried, none matched human perception as well as the static models.

8.4.3 Discussion on Model Validation

With the presented results, I have demonstrated that my developed model more accurately mimics human perception, supported by compelling evidence pointing towards the existence of a similar mechanism in the brain. These findings are crucial in the endeavor to create models that closely resemble how humans perceive social interactions. NRE emerges as a strong candidate in elucidating our behavioral perceptions and how we navigate uncertainty.

These results carry significance in understanding the nuances of human social cognition, particularly in distinguishing between a mere face and a genuine expression. This capability holds substantial evolutionary importance, shedding light on how humans evolved into social species. The ability to differentiate between facial features and authentic emotions is a fundamental aspect of our social development, and NRE's potential in capturing and replicating this intricacy has promising implications for advancing our understanding of human social dynamics.

Applying these mechanism to future computer based vision system might help in many domains, ranging from robots to virtual world where humans will increasingly interact with computers.

8.5 Implications and Limitations

Given how norm-referenced encoding (NRE) matches human classification patterns on both human and monkey faces from our psychophysics experiments, the encouraging results from electrophysiological recordings, and the growing body of work supporting norm-referenced mechanisms Loffler *et al.* (2005); Panis *et al.* (2011); Koyano *et al.*

(2021); Freiwald and Hosoya (2021), I believe NRE offers valuable insights into how our brain perceives dynamic facial expressions.

NRE demonstrates a remarkable transfer learning capability, stemming from its two-stream processing pathway that leverages reference frames. Its simplicity provides a clear perspective on why the model closely matches human perception more accurately than other candidates. Transfer to novel face shapes is achieved by embedding expressions in a space where expression-induced deviations from the reference are consistent across face shapes. Furthermore, we argue that the human-like perception of morphed expressions is facilitated by the model's linearity with respect to expression strength. This feature enables an intuitive interpretation of the NRE's output activity. I find the linearity of NRE both elegant and a plausible brain mechanism, as I believe the brain employs simple and robust mechanism.

Considering these characteristics, NRE appears well-suited to offer fresh insights into the question of how we recognize faces. The robustness, rapid learning, and simple rules employed by NRE to categorize ambiguous facial expressions may elucidate the cornerstone of our visual system, which we rely on daily for social interaction. In neuroscience, my model could serve as a benchmark to validate or test new hypotheses. For computer scientists, my view is more nuanced. When I reflect on how convolutional neural networks (CNNs) came into existence, this research might inspire new model architectures that incorporate novel structural mechanisms for building latent spaces. However, looking at the recent results of models such as CLIPRadford *et al.* (2021) and OFAWang *et al.* (2022b), with their impressive zero-shot performance, and the current machine learning ecosystems to share model weights, I believe that the needs for extremely data-efficient models is largely reduced. Nonetheless, the linearity of the read-out activity in NRE-based models might better handle ambiguous inputs, such as perceiving subtle changes in facial expressions, compared to models trained to perform discrete decisions based solely on the statistical distribution found in their dataset.

Nonetheless, we acknowledge that some limitations remain. From a psychophysical perspective, a current limitation of our model is that it does not account for rotations larger than a dozen degrees, which is a trivial problem for humans. Future iterations of this work could investigate processing a three-dimensional reference representation. Recent work Alreja *et al.* (2023) has incorporated generative models from computer graphics for modeling face perception, utilizing representations in the model's latent space to construct deviations from a norm face. While slightly less interpretable than operating on facial landmarks, such approaches present a promising avenue for exploring view-independent, three-dimensional norm-based representations in forthcoming efforts.

From a machine learning perspective, our current NRE-model has two shortcomings, it does not have a differentiable loss function and the reference patterns are given. Solving for these two shortcomings would enable end-to-end training and make our model practical for large-scale computer vision tasks. The model could learn and optimize its reference frames from the training set. We have already worked to address this shortcoming. However, I will leave the derivation of our custom loss function out of my thesis

to avoid any conflict with publication, as this work was done in collaboration with my colleague Alexander Lappe.

Chapter 9

Conclusion

Throughout this thesis, I am gratified to have shed light to new insights into the mechanisms employed by the brain in perceiving dynamic facial expressions. In the course of this exploration, I have convincingly demonstrated the validity of the norm-referenced mechanism as an encoding for facial expression perception. This achievement is particularly rewarding, given the alignment with preliminary findings from electrophysiological data in rhesus macaques supporting the linear tuning posited by a norm-referenced encoding to neuroscience framework such as the thousand brain theory supporting the use of similar mechanism is each cortical columns.

Our contribution extends beyond mere validation, as it expands the repertoire of tasks where norm-referenced encoding proves effective. Furthermore, the proximity of perception encoded through norm-referenced mechanisms to human cognition is underscored by the minimal entropy observed between our model and other tested computer vision algorithms. These results offer a plausible explanation for our visual system's ability to discern subtle facial movements, distinguishing between genuine and fake dynamic facial expressions.

In addition, the robustness of our visual system, as evidenced in our psychophysics experiments, aligns with the framework and model developed in this thesis, affirming norm-referenced encoding as a data-efficient and robust mechanism. The demonstrated ability of this mechanism to transfer learned features across domains and handle larger datasets strengthens its utility.

Successfully achieving the objectives set forth in this thesis—replicating human behavioral results, excelling in computer vision tasks, and providing a framework aligning with neuronal activities—marks a significant accomplishment. This endeavor has seamlessly bridged the gap between experiments, data analysis, and modeling.

Nevertheless, certain aspects require further attention. The developed architecture currently operates optimally for frontal or moderately turned facial expressions. While the mechanism accommodates 3D space coordinates, the challenge lies in developing input, such as 3D morphable models, capable of generalizing across diverse face shapes. Additionally, the model's learning and selection of a reference frame remain unexplained, opening avenues for exploration using unsupervised methods.

Lastly, reflecting on dynamic facial expression versus static images, my perspective on

their impact on classification performance has become more nuanced. Although dynamic expressions may aid in distinguishing a liveable person, their contribution to categorizing specific expressions remains uncertain. The evidence gathered does not definitively establish dynamic facial expressions as superior in discriminating between different emotional expressions compared to static images.

Appendix A

Appendix

A.1 Human Participant

In total, 78 human participants (42 females) participated in the psychophysical studies. The age range was 21–53 years (mean 26.2, standard deviation 4.71). All participants had no prior experience with macaque monkeys and normal or to-normal corrected vision. Participants gave written informed consent and were reimbursed with 10 EUR per hour for the experiment. In total, 31 participants (16 females) were taking part in the first experiment using stimuli based on the original motion capture data and the experiment with occlusion of the ears. 22 participants (13 females) took part in the experiment with equilibrium motion of the prototypes. In addition, 16 participants (eight females) took part in a Turing test control experiment (see below), and nine (five females) participants took part in a control experiment to identify features that influence perceived expressiveness of the stimuli. All psychophysical experiments were approved by the Ethics Board of the University Clinic Tübingen and were consistent with the rules of the Declaration of Helsinki.

A.2 Stimulus Presentation

Subjects were presented the stimuli watching a computer screen at a distance of 70 cm in a dark room (view angle about 12 degrees), with a resolution of 720 * 720 pixels using MATLAB and the Psychotoolbox (3.0.15) library for stimulus presentation. Each stimulus was repeated for a maximum of three times before asking for the responses, but participants could skip after the first presentation if they were certain about their responses. Participants were first asked whether the perceived expression was rather from a human or a monkey, and whether it was rather the first or the second expression. Responses were given by key presses. Stimuli for the two different avatar types were presented in different blocks, with 10 repeated blocks per avatar type.

Monkey front view							
Model	Accuracy [%]	Accuracy increase [%]	BIC	# Parameters	df	χ^2	p
Model 1	38.29		7487	33	3		
Model 2	57.86	19.56 (rel. to 1)	5076	36	3	2411	<0.0001
Model 3	49.49	11.2 (rel. to 1)	6125	36	3	1362	<0.0001
Model 4	77.53	19.67 (rel. to 2)	3586	39	3	1490	<0.0001
Model 5	77.53	0 (rel. to 4)	3586	42	3	11.997	<0.0074
Model 6	77.42	-0.11 (rel. to 4)	3580	42	3	5.675	0.129

Table A.1: Results of the accuracy and the Bayesian Information Criterion (BIC) for the different logistic multinomial regression models for the stimuli derived from the original motion (no occlusions) monkey avatar front view. The models included the following predictors: Model 1: constant; Model 2: constant, e; Model 3: constant, s; Model 4: constant, s, e; Model 5: constant, s, e, product $s*e$; Model 5: constant, s, e, Optic Flow (OF).

A.3 Comparison of different classification models

Different multinomial regression models were compared in order to find the most compact model that explains our classification data. The models differed in terms of the predictor variables of the linear model for the approximation of the variables y_j. The six compared models were defined as

- Model 1: $y_j = \beta_{0j}$,
- Model 2: $y_j = \beta_{0j} + \beta_{1j}e$,
- Model 3: $y_j = \beta_{0j} + \beta_{1j}s$,
- Model 4: $y_j = \beta_{0j} + \beta_{1j}e + \beta_{1j}s$,
- Model 5: $y_j = \beta_{0j} + \beta_{1j}e + \beta_{1j}s + \beta_{3j}e*s$, and
- Model 6: $y_j = \beta_{0j} + \beta_{1j}e + \beta_{1j}s + \beta_{3j}OF$.

Apart from the style variables e and s, the variable OF signifies the optic flow computed from the image sequence with an optic flow algorithm. Models are compared based on two criteria. First, we require that the introduction of additional predictors does not result in a significantly higher prediction accuracy. According to this criterion, for almost all stimulus types, model four is the most compact model for the front view stimuli (Table A.1, Table A.2, Table A.3 and Table A.2). Only for the rotated views of the avatars, however, we find a slight significant increase of the prediction accuracy (by less than 1.57%). For this reason, we decided to use model four as the basis for our further analyses of all classification data in the main experiment.

Human front view

Model	Accuracy [%]	Accuracy increase [%]	BIC	# Parameters	df	χ^2	p
Model 1	36.84		7481	33	3		
Model 2	54.22	17.38 (rel. to 1)	5541	36	3	1940	<0.0001
Model 3	53.56	16.72 (rel. to 1)	5846	36	3	1633	<0.0001
Model 4	81.56	27.35 (rel. to 2)	3420	39	3	2120	<0.0001
Model 5	81.35	-0.22 (rel. to 4)	3309	42	3	112	<0.0001
Model 6	81.38	-0.18 (rel. to 4)	3389	42	3	31.66	<0.0001

Table A.2: Results of the accuracy and the Bayesian Information Criterion (BIC) for the different logistic multinomial regression models for the stimuli derived from the original motion (no occlusions) human avatar front view. The models included the following predictors: Model 1: constant; Model 2: constant, e; Model 3: constant, s; Model 4: constant, s, e; Model 5: constant, s, e, product $s * e$; Model 5: constant, s, e, Optic Flow (OF).

Monkey 30-degree

Model	Accuracy [%]	Accuracy increase [%]	BIC	# Parameters	df	χ^2	p
Model 1	35.32		6913	33	3		
Model 2	57.40	22.08 (rel. to 1)	4314	36	3	2622	<0.0001
Model 3	49.36	14.04 (rel. to 1)	5179	36	3	1757	<0.0001
Model 4	84.04	26.64 (rel. to 2)	2359	39	3	1977	<0.0001
Model 5	84.88	0.84 (rel. to 4)	2335	42	3	48	<0.0001
Model 6	84.08	0.04 (rel. to 4)	2331	42	3	28	<0.0001

Table A.3: Results of the accuracy and the Bayesian Information Criterion (BIC) for the different logistic multinomial regression models for the stimuli derived from the original motion (no occlusions) monkey avatar 30-degree view. The models included the following predictors: Model 1: constant; Model 2: constant, e; Model 3: constant, s; Model 4: constant, s, e; Model 5: constant, s, e, product $s * e$; Model 5: constant, s, e, Optic Flow (OF).

A.4 Statistical analysis

Statistical analyses are implemented using MATLAB and RStudio (3.6.2), using R and the package lme4 for the mixed models of ANOVA. We use G*Power 1.3 software to compute a prior rough estimate of the minimum required number of participants for medium effect size. Different GLMs for the modeling of the categorization data are fitted using the MATLAB Statistics Toolbox. Models for the discriminant functions, including different sets of predictors, are compared using a step-wise regression approach. Models of different complexity are compared based on the prediction accuracy and by exploiting the BIC. Two statistical measures are applied in order to compare the similarity of the categorization responses for the two avatar types. First, we compute the ratio of the different vs. shared variance between the fitted discriminant functions. For this purpose,

Human 30-degree							
Model	Accuracy [%]	Accuracy increase [%]	BIC	# Parameters	df	χ^2	p
Model 1	37.40		6819	33	3		
Model 2	54.72	18.32 (rel. to 1)	4843	36	3	1975	<0.0001
Model 3	54.36	16.96 (rel. to 1)	5217	36	3	1602	<0.0001
Model 4	81.32	25.6 (rel. to 2)	2910	39	3	1956	<0.0001
Model 5	82.88	1.56 (rel. to 4)	2809	42	3	101	<0.0001
Model 6	81.92	0.06 (rel. to 4)	2890	42	3	19	0.0002

Table A.4: Results of the accuracy and the Bayesian Information Criterion (BIC) for the different logistic multinomial regression models for the stimuli derived from the original motion (no occlusions) human avatar 30-degree view. The models included the following predictors: Model 1: constant; Model 2: constant, e; Model 3: constant, s; Model 4: constant, s, e; Model 5: constant, s, e, product $s*e$; Model 5: constant, s, e, Optic Flow (OF).

we first compute the average discriminant function across both the avatar types and the two view conditions, and separately for the different classes (the index k running over the avatar types and view conditions, and j indicating the class number):

$$\bar{P}_j(e,s) = \frac{1}{4}\sum_j^k (e,s). \tag{A1}$$

The ratio of the variance that is different and shared between the four conditions (avatars and views) is then given by the expression

$$q = \frac{\sum_k \sum_j \iint_0^1 (P_{kj}(e,s) - \bar{P}_j(e,s))^2 deds}{4\sum_{j'} \iint_0^1 \bar{P}_{j'}(e,s)^2 deds}. \tag{A2}$$

This ratio is zero if the discriminant functions across all four conditions are identical. As second statistical analysis, we compare the multinomially distributed four-class classification responses across the participants for the individual points in morphing space using a contingency table analysis that tested for the independence of the class probabilities from the avatar types and the two view conditions. Statistical differences are evaluated using a χ^2-test and, for cases for which predicted frequencies are lower than 5, we exploit a bootstrapping approach Wilson *et al.* (2020). The species-tuning functions, $D_H(s)$ and $D_M(s)$, are fitted by the sigmoidal function $D_{H,M} = (tanh(\omega(s-\theta))+1)/2$, with the parameter θ determining the threshold and ω, the steepness. Differences of the tuning parameters θ are tested using two-factor mixed-model ANOVAs species-specific of motion (monkey vs. human) as the within-subject factor and experiment (original motion, occlusion of the ears, and equilibrated motion) as the between-subject factor. Differences of the steepness parameters ω are tested using within-subject two-factor ANOVAs.

A.5 Asymmetry index

The deviation of the four discriminant functions $P_i(e,s)$ from the completely symmetrical case, where all four discriminant functions have the same basic shape (with their peaks centered on the different prototypes), are quantified by defining the asymmetry index AI. This index is exactly zero if the four discriminant functions are exactly symmetrical with respect to the axes $e = 0.5$ and $s = 0.5$. This implies the symmetry relationship $P_1(e,s) = P_2(1-2,s) = P_3(e,1-s) = P_4(1-e,1-s)$. In order to compute the index, we first computed a symmetrized average of all four discriminant functions according to the formula

$$P_{sym}(e,s) = \frac{P_1(e,s) + P_2(1-2,s) + P_3(e,1-s) + P_4(1-e,1-s)}{4}. \tag{A3}$$

Likewise, we defined a standard deviation relative to this symmetrized average by the expression $SD_{sym}(2,s) = \sqrt{Q_{sym}(e,s)/3}$ with the least square deviation sum

$$\begin{aligned} Q_{sym}(e,s) = {} & (P_1(e,s) - P_{sym}(e,s))^2 + (P_2(1-e,s) - P_{sym}(e,s))^2 \\ & + (P_3(e,1-s) - P_{sym}(e,s))^2 + (P_4(1-e,1-s) - P_{sym}(e,s))^2. \end{aligned} \tag{A4}$$

The asymmetry index was defined by the expression

$$AI = \frac{\iint_0^1 SD_{sym}(e,s)deds}{\iint_0^1 P_{sym}(e,s)deds} \tag{A5}$$

The AI increases with the deviation from the completely symmetric case, where all four categories are represented equally well.

A.6 Testing of low-level information that predicts expression strength

Since we find that a larger portion of the tested perceptual space is classified as monkey expressions compared to human expressions in natural dynamic expressions (Figure 8.2C), we suspect this result to be a potential consequence of monkey expressions specifying more salient low-level features, such as local motion or geometrical deformations. To control for this variable, we create a second stimulus set in which the amount of low-level information is balanced. Since it is unknown *a priori* which type of low-level information drives the expression strength of facial expressions, we test nine possible measures that quantify the amount of low-level features in a separate psychophysical experiment involving nine participants. These measures include two-dimensional optic flow computed from the movies using a Horn-Schunck algorithm (MATLAB implemen-

tation), the absolute spatial deformation relative to the neutral frame, and the motion flow computed either from the control point trajectories or from the regularized mesh points, either in three dimensions or after projection to the two-dimensional image plane. The spatial deformation relative to the neutral frame is quantified using the measures. The asymmetry index is defined by the expression.

$$DF = \sum_{t=1}^{N} \| X_t - X_0 \|_2, \tag{A6}$$

where X_t signifies a vector that contains the relevant control point or (two- or three-dimensional) mesh point coordinates and where N is the number of stimulus frames. Likewise, the motion flow was defined by the quantity

$$MF = \sum_{t=2}^{N} \| X_t - X_{t-1} \|_2 \, . \tag{A7}$$

For the true optic flow, the motion measure is computed by summing up the absolute values of all estimated local motion vectors across the image. The stimuli for this experiment are motion morphs between each of the four prototype expressions (two human and two monkey expressions) and a neutral expression. The original expression enters the motion morph with a weight of λ, and the neutral expression with a weight of $(1-\lambda)$, where the morphing weight is adjusted to obtain the same amount of low-level information in all adjusted prototypes.

In order to cut down the number of measures for the amount of low-level information in the first place, we generate a set of face motion stimuli with reduced and exaggerated expression strength, separately for the two face avatars, by choosing six different values for the morphing weight λ (values $0-25-50-75-100-125\%$ for the monkey expressions and the values $0-37.5-75-112.5-150\%$ for the human expressions). For all rendered movies, we compute the nine different measures for the low-level feature content and analyze their dependence on the morphing weight λ and their similarities. We find that the measures DF and MF, computed from the two-dimensional and three-dimensional mesh coordinates, and the control points are very highly correlated ($r > 86.24, r_{average} = 98.74; p < 0.0403$). The mesh point-based measures are monotonically increasing functions of the morphing level λ. This is not the case for the quantities computed from the control point trajectories, due to which we discard the measures derived from the control point trajectories from the balancing of the stimuli. Because of the high correlation between the measures computed from the two- and three-dimensional mesh-point trajectories, and the higher similarity of the two-dimensional trajectories with image motion, we keep only the measures computed from the two-dimensional mesh-point trajectories for the further analysis. In addition, we test the optic flow computed by the optic flow algorithm from pixel images as a third possible predictor of the low-level information. For each of these three predictors, we construct a balanced stimulus set by

Threshold monkey avatar

Source of variation	Sum of square	df	mean square	F	P
Stimulus type	0.00	2	0.00	0.00	0.999
Expression type	1.20	1	1.20	188.83	<0.0001
Stimulus * Expression	0.06	2	0.03	4.51	0.015
Error	0.42	60	0.01		
Total	1.72	65			

Table A.5: ANOVA for the threshold of the monkey avatar: two-way mixed model with expression type as within-subject factor and the stimulus type as between-subject factor for both the monkey and the human avatar. The mean square is defined as Mean Square = Sum of square/df; df = degree of freedom.

adjusting the morph levels of all prototypes, except for the one with the lowest low-level feature content, in order to match their low-level information contents. As a result, we obtain three balanced sets of stimuli, each with four dynamic expressions, separately for each avatar type.

All stimuli are shown in a block-wise randomized order to the participants who have to rate their expression strength on a nine-point Likert scale. The stimuli for the human avatar, which are balanced based on the motion flow measure MF, exhibit the least variability in expression strength among participants and the highest overall expression strength. For the monkey avatar, the expression strength is rated similarly for stimuli balanced using the measures MF and DF, while it is significantly lower for stimuli balanced using the optic flow ($t(275) = 2.8; p = 0.0054$ and $t(269) = 3.95; p < 0.001$). A step-wise regression analysis, in which we predict the expression strength ratings from the remaining measures (MF and DF computed from the two-dimensional mesh motion), shows that the motion flow MF is sufficient, while the other predictor DF does not add significant additional information. Using a model comparison analysis exploiting the Bayesian Information Criterion (BIC), we find no significant difference in the explanatory values of the models including the predictor MF, and the predictors MF and DF ($\chi^2(1,284) = 3.49; p = 0.062$).

A.7 Anova results

A.8 Example-based model

For this model the recognition is accomplished by using a RNN that detects temporal sequences of learned keyframes from facial expressions (Figure 4.1A). This type of mechanism has been shown to reproduce data from cortical neurons during the recognition of body actions (e.g. [9, 7]. It has been show to reproduce activity data from action

Threshold human avatar

Source of variation	Sum of square	df	mean square	F	P
Stimulus type	0.00	2	0.00	0.00	0.993
Expression type	0.40	1	0.40	46.37	<0.0001
Stimulus * Expression	0.05	2	0.03	3.15	0.049
Error	0.57	60	0.01		
Total	1.02	65			

Table A.6: ANOVA for the threshold of the human avatar: two-way mixed model with expression type as within-subject factor and the stimulus type as between-subject factor for both the monkey and the human avatar. The mean square is defined as Mean Square = Sum of square/df; df = degree of freedom.

Steepness original motion

Source of variation	Sum of square	df	mean square	F	P
Avatar type	376.68	1	376.68	6.3	0.016
Expression type	0.36	1	0.36	0.01	0.939
Avatar * Expression	0.16	1	0.16	0	0.959
Error	2391.21	40	59.78		
Total	2768.41	43			

Table A.7: ANOVA for steepness of orignial motion: two-way ANOVA with avatar type and expression factor for original motion type. The mean square is defined as Mean Square = Sum of square/df; df = degree of freedom.

selective neurons in the STS and in premotor cortex. Expressions were represented by a total of 50 keyframes. The structure of the example-based encoding circuit for one expression is shown in Fig. 1 B. The output from the previous layer, signified as vector $\mathbf{z}$, is providing input to Radial Basis Function units (RBFs) that were trained by setting their centers to the vectors $\mathbf{z}_p^n$ for the individual expression frames of expression p. The actual outputs of these neurons are then given by: $f_k^p = exp((|\mathbf{z} - \mathbf{z}_p^k|^2/(2\sigma^2)$. (For sufficient selectivity to distinguish temporally distant keyframes we chose $\sigma = 0.1$.) The outputs of the RBFs were thresholded with a linear threshold function.

The output signals of the RBFs were used as input of a recurrent neural network, or discretely approximated neural field [18], that obeys the dynamics:

$$\tau \dot{u}_n^p(t) = -u_n^p(t) + \sum_m w(n-m)[u_m^p(t)]_+ + s_n^p(t) - h - w_c I_c^p(t) \tag{A8}$$

The activity $u_p^n(t)$ is the activity of the neuron in the RNN that encodes keyframe n of the facial expression type p. The resting level constant was $h = 1$, and the time constant $\tau = 5$ (using an Euler approximation). The lateral interaction was asymmetric, inducing

Steepness occluded motion

Source of variation	Sum of square	df	mean square	F	P
Avatar type	286.17	1	286.17	3.33	0.076
Expression type	0.02	1	0.02	0	0.988
Avatar * Expression	0.00	1	0.00	0	0.995
Error	3094.54	36	85.96		
Total	3380.73	39			

Table A.8: ANOVA for steepness of orignial motion: two-way ANOVA with avatar type and expression factor for occluded motion type. The mean square is defined as Mean Square = Sum of square/df; df = degree of freedom.

Steepness equiliobrated motion

Source of variation	Sum of square	df	mean square	F	P
Avatar type	1.57	1	1.57	0.4	0.533
Expression type	0.25	1	0.25	0.06	0.803
Avatar * Expression	0.02	2	0.02	0	0.945
Error	174.76	44	3.97		
Total	176.60	47			

Table A.9: ANOVA for steepness of equilibrated motion: two-way ANOVA with avatar type and expression factor for original motion type. The mean square is defined as Mean Square = Sum of square/df; df = degree of freedom.

a network dynamics that is sequence selective. It was given by the function

$$w(n) = Ae^{-\frac{(n-C)^2}{2\sigma_{ker}^2}} - B \tag{A9}$$

with the parameters $A = 1$, $B = 0.5$, and $C = 3.5$. A cross inhibition term leads to competition between the neural subnetworks that encode different expressions. It was defined by the equation: $I_c^p(t) = \sum_{p \leq n} [\mathbf{u}_m^{p'}(t)]_+$ with the cross-inhibition weight $w = 0.5$.

The input signals $s_n^p(t)$ was computer from the output signals of the RBF units that recognize the key frames of expression p. These can be described by the vector $b^p = [b_1^1 p, ..., b_{50}^p]^T$ for the actual time step. The components of this vector were smoothed along the neuron axis using a Gaussian filter with a width (standard deviation) of 2 neurons. The neurons in this RNN are called snapshot neurons in the following. They are expression-selective and fire phasically during the evolution of the expressions. In addition, they are sequence-selective, i.e. they fire less strongly if the same keyframe is presented as part of temporally inverted sequence. The thresholded output signals of the snapshot neurons belonging to the same expression are integrated by an expression neuron that computes the maximum over all snapshot neuron outputs (cf. Fig. 1 B).

Model	BFS	Human	Monkey	Cartoon
SVM-LBP Luo *et al.* (2013)	14.3	14.3	14.3	14.3
ResNet50v2 Ngo and Yoon (2020)	52.38	**100**	33.33	33.33
CORnet-S	63.1	91.67	53.33	50.00
DAN Wen *et al.* (2021)	37.37	75.47	28.21	17.91
EfficientFace Zhao *et al.* (2021)	33.8	75.47	15.38	22.39
MD-NRE	**100**	**100**	**100**	**100**

Table A.10: Detailed classification accuracies (%) for our domain generalisation task separated for each avatar type

These neurons have a time constant $\tau_v = 4$, so that their outputs can be describe by the dynamics:

$$\tau_v \dot{v}_p(t) = -v_p(t) + \sum_n [u_n^p(t)]_+. \tag{A10}$$

The expressions neuron thus fire continuously during the evolution of the corre- sponding expression p, but only weakly during the other ones.

A.8.1 CORNet-S training

To train our CORNet-S model, we follow the work by Ngo *et al.*Ngo and Yoon (2020) that investigates the training of CNNs in a FER task for an unbalanced dataset. We made one notable change in their original pipeline by using a loss function that depends on the raw outputs instead of the logit outputs from the last layer. Data augmentation, RGB image normalisation and class weights are all applied as in the original paper. For optimisation, we use a stochastic gradient descent algorithm with a learning decay of 0.04, a momentum of 0.9 and an L2 regularisation coefficient of 0.01. The initial learning rate is 0.0001 for CORnet-S models and train the model for 35 epochs. We also use a subset of the AffectNet dataset as Ngo *et al.*to train our models. This subset includes only the manually labelled images for 8 facial expressions, which reduces the training set to 280k+ images. We train our model on 2x Nvidia GTX 1080Ti with a batch size of 64 using the Tensorflow library.

A.8.2 Detailed Results for Domain Generalisation

Table A.10 shows detailed results of our *Domain Generalisation* task avatar type. We leave the SVM model for a detailed discussion (A.10). The table reveals that each CNN model performs well on our human avatar. They range from 75.47% to 100% with ResNet50v2. These results show the high quality of our human avatar and our set of facial expressions. Results show how the classifications' accuracy drops dramatically with the Monkey and the cartoon avatars. The worst performing model is the EfficientFace model on the monkey avatar (15.38%) and the DAN on the Cartoon avatar (17.91%). These results are close to the chance level (14.3%), demonstrating clearly how these CNNs based architecture does not transfer easily on new domains. Interestingly,

Model	25%	50%	75%	100%
CORNet-S	27.78	33.33	50	72.22
MD-NRE	**50**	**77.78**	**83.33**	**100**

Table A.11: Classification accuracy (%) for our MD-NRE method and the CORNet-S on the expression strength level test (BFS-L) for different expression levels (across all avatars).

CORNet-S model performs best, reaching more than 50& accuracy. CORNet-S authors claims that this architecture is more brain-like Kubilius *et al.* (2018). Nonetheless, this model is a feed-forward approach, but its extensive use of weight sharing seems to be beneficial for transfer of learned knowledge from domain to domain.

A.9 Detailed Results of Expression Strengths

The detail results for the BFS-L dataset (see 7.4), which requires to classify the expression type for pictures with different expression strength (models being trained only with 100% expression strength) show that our model outperforms the CORNet-S model for each expression strength. As expected, both models shows a clear monotonic increase of performance with the expression strength, with the CORNet-S model going from 27.82% to 72.22%, and our MD-NRE model going from 50% to a 100% classification accuracy. This makes sense since weaker expressions shows less deviation from the neutral expression, and thus are more difficult to classify. At the same time, this demonstrates a superior robustness of expression classification using our method in terms of variation of expression strength.

A.10 Discussion on SVM Results

Since deep learning architectures typically require huge datasets for training, and our technique is designed to work with relatively small amounts of training data, we also wanted to test against a well-established standard method for learning from moderate amounts of data. We chose a Support Vector Machine (SVM) classifier approach Luo *et al.* (2013), which was shown to achieve decent performance for facial expression recognition prior to the introduction of deep neural network architectures. This method combines a principle components analysis (PCA) for feature extraction with a SVM classifier. We used the same training and test sets as for our own method. For this architecture we found a poor performance of 14.3% (equal to chance level) for all face types, thus effectively no transfer learning. This is likely a consequence of PCA, as a holistic method, not being able to disentangle successfully the components that carry information about expression and basic head shape. Since PCA is mainly driven by the

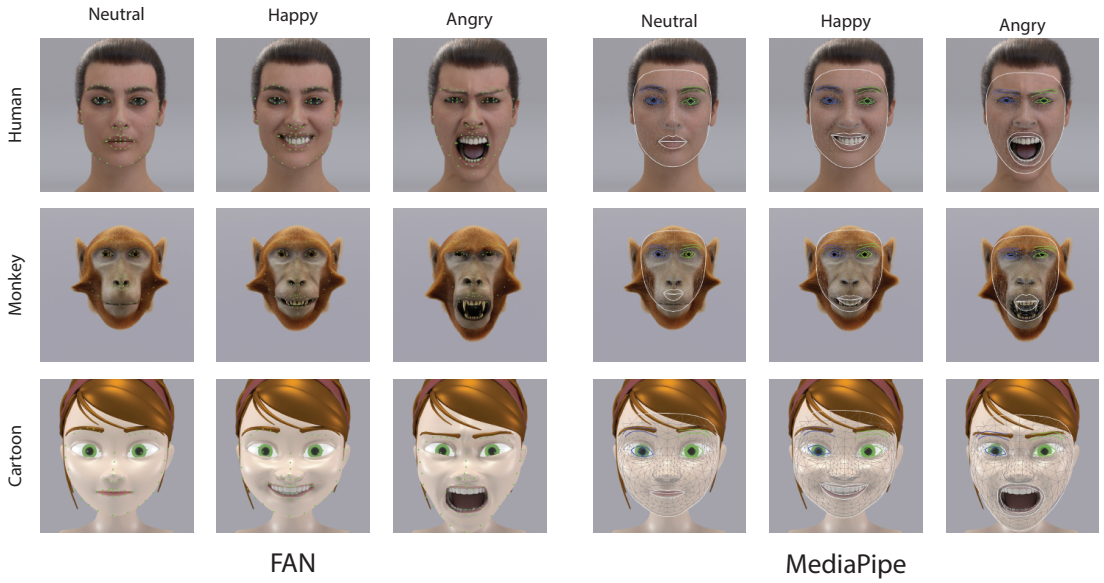

Figure A.1: Results obtained with other landmark detectors on examples from the BFS dataset.

variance in pixel space, this implies that the major components might be driven mainly by differences between the different head types that by the small variations induced by changes in expression.

A.11 Results with Other Landmark Detectors

Figure A.1 shows the result obtained by applying the FAN network Bulat and Tzimiropoulos (2017) and the MediaPipe (MP) method Kartynnik *et al.* (2019) to our BFS dataset. For both methods the landmark detection misses specifically the eyebrows, which might be due to their strong curvature on both the monkey and cartoon compared to the human head shape. This is most prominent for the angry condition. Both methods show particularly poor results for the mouth region of the monkey avatar. This lack or reliable tracking prevented us from using these established methods for the non-human face pictures.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application programming interface
AsI	asymmetry index
AU	Action Unit
BFS	Basic Face Shape
BNN	Biological Neural Network
CNN	Convolutional Neural Network
DF	Distance Flow
FER	Facial Expression Recognition
FR	Facial Recognition
MF	Motion Flow
NN	Neural Network
NRE	Norm-Referenced Mechanism
OF	Optical Flow
RNN	Recurrent Neural Network

License

Attribution 4.0 International

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

Adapted Material means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image. Adapter's License means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License. Copyright and Similar Rights means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section 2(b)(1)-(2) are not Copyright and Similar Rights. Effective Technological Measures means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements. Exceptions and Limitations means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material. Licensed Material means the artistic or literary work, database, or other material to which the Licensor applied this Public License. Licensed Rights means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of

the Licensed Material and that the Licensor has authority to license. Licensor means the individual(s) or entity(ies) granting rights under this Public License. Share means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them. Sui Generis Database Rights means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world. You means the individual or entity exercising the Licensed Rights under this Public License. Your has a corresponding meaning.

Section 2 – Scope.

License grant. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to: reproduce and Share the Licensed Material, in whole or in part; and produce, reproduce, and Share Adapted Material. Exceptions and Limitations . For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions. Term . The term of this Public License is specified in Section 6(a) . Media and formats; technical modifications allowed . The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material. Downstream recipients . Offer from the Licensor – Licensed Material . Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License. No downstream restrictions . You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material. No endorsement . Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i) . Other rights . Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or

other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise. Patent and trademark rights are not licensed under this Public License. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

Attribution. If You Share the Licensed Material (including in modified form), You must:

retain the following if it is supplied by the Licensor with the Licensed Material: identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated); a copyright notice; a notice that refers to this Public License; a notice that refers to the disclaimer of warranties; a URI or hyperlink to the Licensed Material to the extent reasonably practicable; indicate if You modified the Licensed Material and retain an indication of any previous modifications; and indicate the Licensed Material is licensed under this Public License, and include the text of, or the URI or hyperlink to, this Public License. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable. If You Share Adapted Material You produce, the Adapter's License You apply must not prevent recipients of the Adapted Material from complying with this Public License.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database; if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database

Rights (but not its individual contents) is Adapted Material; and You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database. For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You. To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect, incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:

automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or upon express reinstatement by the Licensor. For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.

For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License. Sections 1 , 5 , 6 , 7 , and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 – Interpretation.

For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce, limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Alreja, A., Ward, M. J., Colan, J. A., Ma, Q., Richardson, R. M., Morency, L.-P., and Ghuman, A. S. (2023). Reconstructing the neurodynamics of face perception during real world vision in humans using intracranial eeg recordings. *Journal of Vision*, **23**(9), 5487–5487.

Aneja, D., Colburn, A., Faigin, G., Shapiro, L., and Mones, B. (2016). Modeling stylized character expressions via deep learning. In *Asian conference on computer vision*, pages 136–153. Springer.

Ashir, A. M., Eleyan, A., and Akdemir, B. (2020). Facial expression recognition with dynamic cascaded classifier. *Neural Computing and Applications*, **32**, 6295–6309.

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision research*, **172**, 46–61.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Bernstein, M. and Yovel, G. (2015). Two neural pathways of face processing: A critical evaluation of current models. *Neuroscience & Biobehavioral Reviews*, **55**, 536–546.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030.

Caggiano, V., Fleischer, F., Pomper, J. K., Giese, M. A., and Thier, P. (2016). Mirror neurons in monkey premotor area f5 show tuning for critical features of visual causality perception. *Current Biology*, **26**(22), 3077–3082.

Calvo, M. G. and Lundqvist, D. (2008). Facial expressions of emotion (kdef): Identification under different display-duration conditions. *Behavior research methods*, **40**(1), 109–115.

Chang, L. and Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, **169**(6), 1013–1028.

Chang, L., Egger, B., Vetter, T., and Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*.

Chaudhuri, B., Vesdapunt, N., and Wang, B. (2019). Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728.

Chen, J., Konrad, J., and Ishwar, P. (2018). Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1570–1579.

Courville, P., Goodfellow, A., Mirza, I., and Bengio, Y. (2013). Fer-2013 face database. *Universit de Montreal: Montréal, QC, Canada*.

Curio, C. E., Bülthoff, H. H., and Giese, M. A. (2011). Dynamic faces: Insights from experiments and computation. In *COSYNE conference, Mar, 2008, Snowbird, UT, US; This book is an outgrowth of the aforementioned workshop.* MIT Press.

Dahl, C. D., Chen, C.-C., and Rasch, M. J. (2014). Own-race and own-species advantages in face perception: a computational view. *Scientific reports*, **4**(1), 1–9.

Darwin, C. and Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dobs, K., Isik, L., Pantazis, D., and Kanwisher, N. (2019). How face perception unfolds over time. *Nature communications*, **10**(1), 1–10.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Egger, B., Smith, W. A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., *et al.* (2020). 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, **39**(5), 1–38.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**(3-4), 169–200.

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

Felsen, G. and Dan, Y. (2005). A natural approach to studying vision. *Nature neuroscience*, **8**(12), 1643–1646.

Freiwald, W. A. and Hosoya, H. (2021). Neuroscience: A face's journey through space and time. *Current Biology*, **31**(1), R13–R15.

Gao, Y., Leung, M. K., Hui, S. C., and Tananda, M. W. (2003). Facial expression recognition from line-based caricatures. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **33**(3), 407–412.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018a). Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018b). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Georgescu, M.-I., Ionescu, R. T., and Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, **7**, 64827–64836.

Giannopoulos, P., Perikos, I., and Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on fer-2013. *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, pages 1–16.

Giese, M. A. and Leopold, D. A. (2005). Physiologically inspired neural model for the encoding of face spaces. *Neurocomputing*, **65**, 93–101.

Goldman, D. and Homa, D. (1977). Integrative and metric properties of abstracted information as a function of category discriminability, instance variability, and experience. *Journal of Experimental Psychology: Human Learning and Memory*, **3**(4), 375.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, **95**(2), 245–258.

Hasson, U. and Honey, C. J. (2012). Future trends in neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, **62**(2), 1272–1278.

Hasson, U., Nastase, S. A., and Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, **105**(3), 416–434.

Hawkins, J. (2021). *A thousand brains: A new theory of intelligence*. Hachette UK.

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, **4**(6), 223–233.

Ionescu, R. T., Popescu, M., and Grozea, C. (2013). Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on challenges in representation learning, ICML*. Citeseer.

Jack, R. E., Sun, W., Delis, I., Garrod, O. G., and Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, **145**(6), 708.

Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., and Liu, J. (2020). Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889.

Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, **58**(6), 1187–1211.

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., and Akamatsu, S. (2013). Dynamic properties influence the perception of facial expressions. *Perception*, **42**(11), 1266–1278.

Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*.

Kätsyri, J. and Sams, M. (2008). The effect of dynamics on identifying basic emotions from synthetic and natural faces. *International Journal of Human-Computer Studies*, **66**(4), 233–242.

Kaye, L. K., Malone, S. A., and Wall, H. J. (2017). Emojis: Insights, affordances, and possibilities for psychological science. *Trends in cognitive sciences*, **21**(2), 66–68.

Kouw, W. M. and Loog, M. (2019). A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, **43**(3), 766–785.

Koyano, K. W., Jones, A. P., McMahon, D. B., Waidmann, E. N., Russ, B. E., and Leopold, D. A. (2021). Dynamic suppression of average facial structure shapes neural tuning in three macaque face patches. *Current Biology*, **31**(1), 1–12.

Krumhuber, E. G., Kappas, A., and Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, **5**(1), 41–46.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D., and DiCarlo, J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. biorxiv, 408385.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, **40**.

Latinus, M., McAleer, P., Bestelmeyer, P. E., and Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, **23**(12), 1075–1080.

Lawrence, N. D. and Moore, A. J. (2007). Hierarchical gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pages 481–488.

Leopold, D. A., Bondar, I. V., and Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, **442**(7102), 572–575.

Li, S. and Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*.

Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. (2017). Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, **36**(6), 194–1.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Loffler, G., Yourganov, G., Wilkinson, F., and Wilson, H. R. (2005). fmri evidence for the neural representation of faces. *Nature neuroscience*, **8**(10), 1386–1391.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current biology*, **5**(5), 552–563.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE.

Luo, Y., Wu, C.-m., and Zhang, Y. (2013). Facial expression recognition based on fusion feature of pca and lbp with svm. *Optik-International Journal for Light and Electron Optics*, **124**(17), 2767–2770.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Matsumoto, D., Keltner, D., Shiota, M. N., O'Sullivan, M., and Frank, M. (2008). Facial expressions of emotion.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, **85**(3), 207.

Minaee, S., Minaei, M., and Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, **21**(9), 3046.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, **10**(1), 18–31.

Neumann, P. G. (1977). Visual prototype formation with discontinuous representation of dimensions of variability. *Memory & Cognition*, **5**(2), 187–197.

Ngo, Q. T. and Yoon, S. (2020). Facial expression recognition based on weighted-cluster loss and deep transfer learning using a highly imbalanced dataset. *Sensors*, **20**(9), 2639.

Niu, B., Gao, Z., and Guo, B. (2021). Facial expression recognition with lbp and orb features. *Computational Intelligence and Neuroscience*, **2021**, 1–10.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of experimental psychology: human perception and performance*, **17**(1), 3.

Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding v1? *Neural computation*, **17**(8), 1665–1699.

Panis, S., Wagemans, J., and Op de Beeck, H. P. (2011). Dynamic norm-based encoding for unfamiliar shapes in human visual cortex. *Journal of Cognitive Neuroscience*, **23**(7), 1829–1843.

Parr, L. A., Waller, B. M., Burrows, A. M., Gothard, K. M., and Vick, S.-J. (2010). Brief communication: Maqfacs: a muscle-based facial movement coding system for the rhesus macaque. *American journal of physical anthropology*, **143**(4), 625–630.

Pascalis, O., Scott, L. S., Kelly, D. J., Shannon, R. W., Nicholson, E., Coleman, M., and Nelson, C. A. (2005). Plasticity of face processing in infancy. *Proceedings of the national academy of sciences*, **102**(14), 5297–5300.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**(6255), 263–266.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.* (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ratté, S., Lankarany, M., Rho, Y.-A., Patterson, A., and Prescott, S. A. (2015). Subthreshold membrane currents confer distinct tuning properties that enable neurons to encode the integral or derivative of their input. *Frontiers in cellular neuroscience*, **8**, 452.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive psychology*, **3**(3), 382–407.

Rhodes, G., Brennan, S., and Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive psychology*, **19**(4), 473–497.

Ribera, R. B. I., Zell, E., Lewis, J. P., Noh, J., and Botsch, M. (2017). Facial retargeting with automatic range of motion alignment. *ACM Transactions on graphics (TOG)*, **36**(4), 1–12.

Schmidt, K. L. and Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, **116**(S33), 3–24.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., *et al.* (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., and DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*.

Scott, L. S. and Fava, E. (2013). The own-species face bias: A review of developmental and comparative data. *Visual Cognition*, **21**(9-10), 1364–1391.

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, **5**, 399–426.

Siebert, R., Taubert, N., Spadacenta, S., Dicke, P. W., Giese, M. A., and Thier, P. (2020). A naturalistic dynamic monkey head avatar elicits species-typical reactions and overcomes the uncanny valley. *Eneuro*, **7**(4).

Siebert, R., Stettler, M., Taubert, N., Dicke, P. W., Giese, M. A., and Thier, P. (2022). Encoding of dynamic facial expressions in the macaque superior temporal sulcus. *Program No. 053.11 2022 Neuroscience Meeting Planner. San Diego, CA:Society for Neuroscience*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stettler, M., Taubert, N., Azizpour, T., Siebert, R., Spadacenta, S., Dicke, P., Thier, P., and Giese, M. A. (2020). Physiologically-inspired neural circuits for the recognition of dynamic faces. In *International Conference on Artificial Neural Networks*, pages 168–179. Springer.

Stettler, M., Taubert, N., Siebert, R., Spadacenta, S., Dicke, P., Thier, P., and Giese, M. (2022). Norm-referenced neural mechanism for the recognition of facial expressions across fundamentally different face shapes. *Journal of Vision*, **22**(14), 3398–3398.

Stettler, M., Lappe, A., Taubert, N., and Giese, M. (2023a). Multi-domain norm-referenced encoding enables data efficient transfer learning of facial expression recognition. *arXiv preprint arXiv:2304.02309*.

Stettler, M., Lappe, A., Taubert, N., Siebert, R., Thier, P., and Giese, M. A. (2023b). Norm-reference encoding explains intra-cortical recording and behavioural results of dynamic facial expression perception. *BioXiv*.

Taubert, N., Christensen, A., Endres, D., and Giese, M. A. (2012). Online simulation of emotional interactive behaviors with hierarchical gaussian process dynamical models. In *Proceedings of the ACM Symposium on Applied Perception*, pages 25–32.

Taubert, N., Stettler, M., Siebert, R., Spadacenta, S., Sting, L., Dicke, P., Thier, P., and Giese, M. A. (2021). Shape-invariant encoding of dynamic primate facial expressions in human perception. *Elife*, **10**, e61197.

Thomaz, C. E. and Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and vision computing*, **28**(6), 902–913.

Valentine, T. (2005). 3 face-space models of face recognition. In *Computational, Geometric, and Process Perspectives on Facial Cognition*, pages 83–113. Psychology Press.

Valentine, T. *et al.* (2001). Face-space models of face recognition. *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*, pages 83–113.

Vetter, T. and Troje, N. F. (1995). A separate linear shape and texture space for modeling two-dimensional images of human faces.

Vetter, T., Jones, M. J., and Poggio, T. (1997). A bootstrapping algorithm for learning linear models of object classes. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 40–46. IEEE.

Vick, S.-J., Waller, B. M., Parr, L. A., Smith Pasqualini, M. C., and Bard, K. A. (2007). A cross-species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (facs). *Journal of nonverbal behavior*, **31**(1), 1–20.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. (2022a). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, J. M., Fleet, D. J., and Hertzmann, A. (2007). Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, **30**(2), 283–298.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022b). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Wang, X., Wang, X., and Ni, Y. (2018). Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience*, **2018**.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, **53**(3), 1–34.

Wardle, S. G., Taubert, J., Teichmann, L., and Baker, C. I. (2020). Rapid and dynamic processing of face pareidolia in the human brain. *Nature communications*, **11**(1), 4518.

Wen, Z., Lin, W., Wang, T., and Xu, G. (2021). Distract your attention: Multi-head cross attention network for facial expression recognition. *CoRR*, **abs/2109.07270**.

Wilson, E. O. *et al.* (2012). *The social conquest of earth*. WW Norton & Company.

Wilson, V. A., Kade, C., Moeller, S., Treue, S., Kagan, I., and Fischer, J. (2020). Macaque gaze responses to the primatar: a virtual macaque head for social cognition research. *Frontiers in Psychology*, **11**, 1645.

Zador, A., Richards, B., Ölveczky, B., Escola, S., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., *et al.* (2022). Toward next-generation artificial intelligence: Catalyzing the neuroai revolution. *arXiv preprint arXiv:2210.08340*.

Zhang, S., Liu, X., Yang, X., Shu, Y., Liu, N., Zhang, D., and Liu, Y.-J. (2021). The influence of key facial features on recognition of emotion in cartoon faces. *Frontiers in psychology*, **12**, 687974.

Zhao, H., Liu, Q., and Yang, Y. (2018). Transfer learning with ensemble of multiple feature representations. In *2018 IEEE 16th international conference on software engineering research, management and applications (SERA)*, pages 54–61. IEEE.

Zhao, J., Meng, Q., An, L., and Wang, Y. (2019). An event-related potential comparison of facial expression processing between cartoon and real faces. *PLoS One*, **14**(1), e0198868.

Zhao, Z., Liu, Q., and Zhou, F. (2021). Robust lightweight facial expression recognition network with label distribution training. In *AAAI Conference on Artificial Intelligence*.