

Facing Asymmetry - Uncovering the Causal Link between Facial Symmetry and Expression Classifiers using Synthetic Interventions

Tim Büchner^{*1}[0000-0002-6879-552X], Niklas Penzel^{*1}[0000-0001-8002-4130],
Orlando Guntinas-Lichius²[0000-0001-9671-0784], and Joachim
Denzler¹[0000-0002-3193-3300]

¹ Computer Vision Group, Friedrich Schiller University Jena, 07743 Jena, Germany
`{firstname.surname}@uni-jena.de`

² Dept. of Otorhinolaryngology, Jena University Hospital, 07747 Jena, Germany
`orlando.guntinas@uni-jena.de`

Abstract. Understanding expressions is vital for deciphering human behavior, and nowadays, end-to-end trained black box models achieve high performance. Due to the black-box nature of these models, it is unclear how they behave when applied out-of-distribution. Specifically, these models show decreased performance for unilateral facial palsy patients. We hypothesize that one crucial factor guiding the internal decision rules is facial symmetry. In this work, we use insights from causal reasoning to investigate the hypothesis. After deriving a structural causal model, we develop a synthetic interventional framework. This approach allows us to analyze how facial symmetry impacts a network’s output behavior while keeping other factors fixed. All 17 investigated expression classifiers significantly lower their output activations for reduced symmetry. This result is congruent with observed behavior on real-world data from healthy subjects and facial palsy patients. As such, our investigation serves as a case study for identifying causal factors that influence the behavior of black-box models.

Keywords: Facial Expressions · Facial Asymmetry · Unilateral Facial Palsy · Causal Inference · Intervention

1 Introduction

Emotional expressiveness is a crucial topic in our daily life for communicating our internal state and for understanding other people [76, 95]. The state-of-the-art for automatically classifying the six base emotions [23] is achieved by end-to-end trained black box neural networks [17, 24, 65, 82, 102, 110]. However, it remains unclear how the internal decision-making processes of these models respond to out-of-distribution inputs that were not accounted for during training. Specifically, we observe a performance degradation when classifying facial expressions

^{*} These authors contributed equally to this work.

in individuals with unilateral facial palsy, a condition impairing the ability to produce symmetrical facial expressions due to underlying nerve damage. Although intuition suggests that facial asymmetry could influence model behavior, we lack a quantifiable way to test and validate this hypothesis.

We leverage causal reasoning principles to address this limitation and move beyond empirical analytics to uncover a contributing factor in the underlying mechanisms driving model decision-making. Specifically, our work answers the interventional question [4]: “If we only change the facial symmetry for an input, then how does the output of an expression classifier behave?” First, we provide evidence that a symmetry bias exists for real-world data inside all models using associational methods [9, 61, 71, 73]. Second, moving up on the causal hierarchy [4], we build an interventional framework derived from a structural causal model that allows us to generate synthetic faces and connect symmetry with classifier outputs. To accurately quantify this relationship, we develop an interpretable score and an accompanying hypothesis test. As a case study, we analyze 17 expression classifiers and find significant changes in their predictions for all of them. Specifically, we find that decreases in facial symmetry result in lower logit activations. Our study highlights the importance of symmetry influencing expression classifiers, emphasizing the general need for investigations beyond predictive performance.

2 Related Work

Synthetic data has become a widely accepted tool for evaluating and training computer vision models in diverse applications, such as object detection [56, 94, 97, 103], pose estimation [15, 38, 94, 96], segmentation [13, 77, 80, 81, 92], 3D reconstruction [17, 25, 37, 69, 75], and also for facial tasks [6, 7, 15, 43, 75]. We develop a generative interventional framework that fixes possible other confounding factors to isolate the impact of facial asymmetry on expression classification.

Facial Expression Classification. Since the standardization of facial expression into six base *emotions* by Ekman [23], state-of-the-art performance for automated classification is achieved by end-to-end trained black-box models [1, 2, 12, 17, 50, 65, 82–84, 100, 111]. While such models reach high performance, their inner workings remain opaque. Hence, the relationship between facial symmetry and predictions remains unclear, especially in medical contexts like facial palsy [3, 8, 10, 18, 40, 42, 57, 58]. To address this uncertainty, we study the effects of facial asymmetry on expression classifiers in a controlled setting: the expression space of 3D Morphable Models [5, 21, 29, 112], more specifically FLAME [48]. While EMOCA [19], an extension of DECA, also relies on the FLAME expression space for classification, our approach takes a different route. We maximize the logit activation output for each model and emotion combination by leveraging the expression space, ensuring optimal performance. We then rely on methods from explainability to perform an in-depth investigation into the model behavior.

Explaining Model Decisions Behavior. Local explainability methods, e.g., [74, 85, 87, 88, 90], are used to investigate the behavior of machine learn-

ing models for singular inputs, e.g., highlight important image regions. Further, in [44], such local explanations are summarized to form a more conclusive general understanding of a classifier (global explanation). The focus is put on shortcut biases leading to so-called “Clever Hans predictors” [44]. However, such local attribution methods necessitate a semantic interpretation, for example, by a domain expert. Especially for medical applications, more abstract but relevant features increase the complexity [54,55,78]. We are interested in analyzing facial symmetry, a complex feature not directly part of the input. Global approaches, for example, [41,73], can determine the usage of such abstract features. Especially, [73] is based on causal principles [60,64,70] and tests for conditional dependence between the feature and the network predictions given the labels. In [9,61,62,72], this approach is applied to various application domains such as skin lesion analysis, digital agriculture, or emotion classification. However, here we go one step beyond and extend their approach by an interventional framework to generate more in-depth answers according to Pearl’s causal hierarchy [4].

Synthetic Face Generation. Generative models have long been the go-to approach for modeling human faces. Ranging from parametric 3D Morphable Models (3DMMs) [5,21,29,59], Active Appearance Models [16,28,34,51], or learned in an entirely data-driven manner [29,48,69,99,104,105,112]. The disentanglement of identity, expression, pose, and appearance is a powerful tool for bias identification [15,43] or image manipulation [19,52,66,93]. In contrast, Generative Adversarial Networks prioritize photorealism over control, embedding multiple facial properties into a single latent representation, making it challenging to have specific control over the generation [6,7,14,39,63,67,68]. Many approaches utilize neural networks to compute the 3DMM parameters from 2D images, either by reconstructing faces [25,33,49,69,99,104,105] or training in an adversarial manner [19,52,66,93]. We aim to quantify the impact of facial asymmetry on the predictive behavior of expression classifiers. To achieve this, we prioritize control over photorealism in our generative pipeline using DECA [25]. Hence, we can fix other confounding factors, like appearance, lighting, and pose, at the same time. Building upon FLAME [48], we alter the geometric face model to induce subtle variations, thereby creating realistic facial asymmetry.

3 Evaluating Models by Intervening on Facial Symmetry

Studying mimicry is crucial when analyzing facial palsy, which impacts the mobility of the facial muscles. An objective evaluation of the nerve damage is commonly done via data-driven methods [10,32,40,42]. In this work, we focus on one facial feature likely impacting the downstream performance of expression classifiers: facial symmetry. We start by detailing our investigation’s causal model and framing the question we are trying to answer in Pearl’s causal hierarchy [4]. Afterward, we describe the adapted 3D Morphable Model to perform interventions by changing one face side’s geometry. Lastly, we derive an interpretable score and corresponding significance test to quantify the impact of facial symmetry on a model’s prediction.

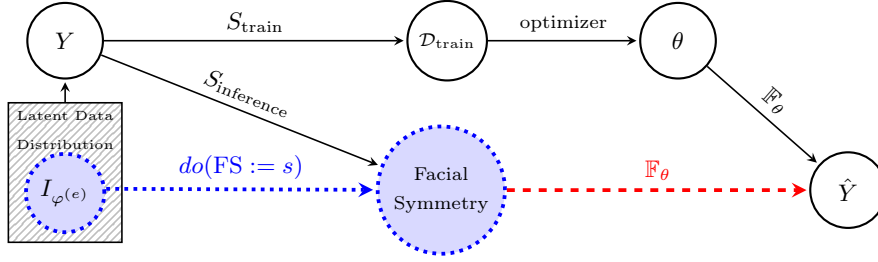


Fig. 1: Expression classifier structural causal model: Y is the expression influenced by the latent distribution of all facial images (hatched box), S_* samples from this latent distribution [73]. $\mathcal{D}_{\text{train}}$ is the training data distribution. The model architecture is an exogenous variable, and weights θ are learned using an optimizer, i.e., a training algorithm. The model’s predictions $\hat{Y}$ result from the trained model $\mathbb{F}_\theta$. We investigate whether $\mathbb{F}_\theta$ is independent of the model predictions $\hat{Y}$ (dashed red arrow). Additionally, we analyze the changes in behavior for varying facial symmetries. Toward this goal, we perform synthetic interventions ($do(\text{Facial Symmetry} := s)$) on facial symmetry variable using 3d morphable models $I_{\varphi(e)}$. Note that these $I_{\varphi(e)}$ are a part (subpopulation) of the latent distribution of all facial images. Adapted from Figure 2 in [73].

3.1 Preliminaries & Causal Point of View

Causal inference tries to answer causal questions from data [64]. This includes interventions, i.e., additional experiments, and purely observational data. Importantly, causal questions can be categorized into a hierarchy. This so-called Pearl’s causal hierarchy (PCH) [4] consists of the following three levels ordered by increasing difficulty: associational, interventional, and counterfactual questions. The latter two are analyzed using the *do*-operator [60], which changes a variable to a constant value, e.g., we would write $do(\text{Facial Symmetry} := s)$ for the variable facial symmetry.

Furthermore, framing data-generating processes and complex interactions of our physical reality as directed graphs enables us to precisely define and investigate the underlying causal mechanisms [60, 64]. The resulting models are called structural causal models (SCMs), and we include a formal definition in the supplementary material. Nevertheless, to understand this framework, it is important to interpret the dependencies, i.e., connections in the graph, as assignments and not as algebraic mappings [64]. Specifically, the connections between variables in such an SCM function like physical mechanisms and not like instantaneous equations. This work extends a specific SCM to model supervised learning [73].

We visualize our SCM for expression classification in Fig. 1, enabling us to study different questions about the decision process. For example, Reimers et al. [73] answer associational questions of whether a feature, such as facial symmetry, is used during the prediction, i.e., does the red dashed arrow exist in Fig. 1. Intuitively, they measure if there is a statistically significant shift in classifier outputs for inputs of the same class but with different feature manifestations.

Other works visualize such significant changes for feature variations [9, 61]. However, they lack actionable descriptions of how the model would behave if a particular feature, e.g., facial symmetry, changes for a specific individual. In this work, we go one step up on the PCH. We employ a synthetic rendering pipeline to alter the facial symmetry while controlling other factors. Hence, we answer the interventional question: “If we change the facial symmetry for an input, then how does the output of the expression classifier behave?” Please note that the levels of the PCH are disjunct and increasing in difficulty. In [4], the authors prove the Causal Hierarchy Theorem (CHT), which states that one needs data of at least the corresponding level to answer causal questions of that level.

In the following, we describe how we generate synthetic data ($I_{\varphi(e)}$ in Fig. 1), where we have fine-grained control over facial symmetry and realized emotional expressions. Using this framework, we generate new interventional data. Hence, we do not violate the CHT [4]. Further, while our approach of synthetic generation necessarily introduces a domain shift (see Fig. 1), we argue that it enables us to go beyond simple interventions. Specifically, our framework allows us to vary the facial symmetry for a specific individual and measure changes in the classifier outputs while fixing other confounding factors. Finally, we discuss how we quantify systematic output changes and determine significance.

3.2 Facial Symmetry Intervention Framework

We require a controllable face generation method to answer interventional questions of the form: “If we change the facial symmetry for an input, then how does the output of the emotion classifier behave?” Additionally, the generation process has to ensure that only facial expressions contribute to the changes measured by the expression classifier. Therefore, we select a 3D Morphable Model (3DMM) [5, 21, 29, 48, 59], to be precise FLAME [48], used in the DECA architecture to create synthetic facial images [25]. Although the generated faces introduce a domain shift, the underlying representation of identity, expression, and appearance gives us complete control over individual changes. Therefore, this disentanglement ensures we can causally link facial changes to the model’s predictive behavior. In the following, we detail our face generation framework to (a) find the expression parameters for optimal classifier activation and (b) introduce a controllable symmetry value s for interventional reasoning.

For all synthetic face images I in this work, we utilize the DECA pipeline [25]: $I = \mathcal{R}(\mathcal{M}, \mathcal{B}, c)$, composed of the face model $\mathcal{M}$, camera position $c \in \mathbb{R}^3$ (fixed to $[0, 0, 0]^T$ in this work) and illumination process $\mathcal{B}$ used in the differential renderer $\mathcal{R}(\cdot)$ [25]. To study facial asymmetry, we alter the face model geometry $\mathcal{M}$ formally defined as $\mathcal{M}(\beta, \vartheta, \varphi, \alpha) = \{\mathcal{G}(\beta, \vartheta, \varphi), \mathcal{A}(\alpha)\}$. DECA employs the geometric components of FLAME $\mathcal{G}$ using the identity $\beta \in \mathbb{R}^{100}$, expression $\varphi \in \mathbb{R}^{50}$, and pose $\vartheta \in \mathbb{R}^6$ blendshape parameters [48]. The texture is computed from the appearance model $\mathcal{A}$ from the Basel Face Model using the parameter α [25, 29, 59]. In FLAME, the face geometry is modeled as

$$\mathcal{G}(\beta, \vartheta, \varphi) = W(T + B_I(\beta, \mathcal{I}) + B_P(\vartheta, \mathcal{P}) + B_E(\varphi, \mathcal{E}), J(\vartheta), \vartheta, \mathcal{W}), \quad (1)$$

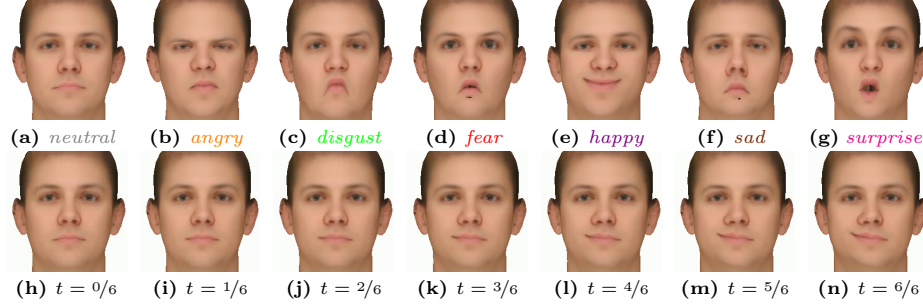


Fig. 2: We display the optimized synthetic face images $I_{\varphi(e)}$ for the *neutral* expression (a) and for the six base emotions (b) - (g) based on the ResidualMaskingNet classifier [65]. Furthermore, we simulate with our geometric face model $\mathcal{G}_{s,t}(\cdot)$ different interpolations t for a symmetry of $s = 0.0$. At $t = 0.0$ (h) we have a *neutral* expression morphing into an asymmetric *happy* expression at $t = 1.0$ (n).

with W being a standard skinning function to rotate the modified N face vertices of the template model $T \in \mathbb{R}^{N \times 3}$ around predefined FLAME joints $J \in \mathbb{R}^{3K}$. B denotes a linear blend skinning (LBS) [45] function of the according blend shapes with identity $\mathcal{I} \in \mathbb{R}^{100 \times N \times 3}$, expression $\mathcal{E} \in \mathbb{R}^{50 \times N \times 3}$, and pose $\mathcal{P} \in \mathbb{R}^{6 \times N \times 3}$. We use the blending weights $\mathcal{W} \in \mathbb{R}^{K \times N}$ of the original FLAME model [48].

Our changes must ensure that (a) under full facial symmetry, the original geometry holds, and (b) a symmetry scalar s specifies facial symmetry and enables interventional queries. Furthermore, we formalize a time parameter t to control the interpolation between *neutral* and a *target* facial expressions [21, 49].

We extend a recent approach by freezing geometry parts to simulate facial asymmetry [105]. Using a scaling parameter s , we can simulate different *freeze* states ranging from 0.0 defining complete asymmetry to 1.0 defining complete symmetry. We artificially induce facial asymmetry by changing only the left side of the face (person’s point of view). Therefore, we recompose the FLAME expression space such that $B_E(\varphi, \mathcal{E}) = B_E(\varphi, \mathcal{E}^L) + B_E(\varphi, \mathcal{E}^R)$. Thus, we define

$$\mathcal{E}_i^L = \begin{cases} \mathcal{E}_i, & \text{if the vertex } i \text{ is on the left side of the face} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

such that the linear blend skinning function $B_E(\varphi, \mathcal{E}^L)$ changes only vertices on the left side of the face [45, 105]. The same applies to $\mathcal{E}^R$. Scaling the blend-shape vectors in $\mathcal{E}^L$ with s induces a symmetry difference between the faces’ sides. Lastly, we multiply the expression parameters φ with t to create dynamic expressions. Our geometric face model $\mathcal{G}_{s,t}$ with symmetry parameter is

$$\mathcal{G}_{s,t}(\beta, \vartheta, \varphi, s, t) = W(T + B_S(\beta, \mathcal{S}) + B_P(\vartheta, \mathcal{P}) + B_E(t \cdot \varphi, \mathcal{E}^R) + B_E(t \cdot \varphi, s \cdot \mathcal{E}^L), J(\vartheta), \vartheta, \mathcal{W}). \quad (3)$$

Thus, the synthetic face image $I_{\varphi^{(e)}}$ updates for a single individual, i.e., β and α are fixed (omitted for clarity), and target expression (e) vector $\varphi^{(e)}$ with symmetry s and temporal dynamic t to $I_{\varphi^{(e)}}(s, t) = \mathcal{R}(\mathcal{G}_{s,t}(\beta, \vartheta, \varphi^{(e)}, s, t), \mathcal{A}(\alpha), \mathcal{B}, c)$.

Within this framework, for an individual, we can (a) modify expressions, (b) simulate facial asymmetry, and (c) simulate movements, ensuring that only changes in facial expression result in changes in the classifier’s behavior. While our synthetic faces are a domain shift for most classifiers, the comparisons are all relative and contained within this new domain. Hence, representing out-of-domain scenarios in which they are applied [6, 7, 65, 82–84]. Furthermore, we optimize the facial expression parameters such that a classifier output $\mathbb{F}_\theta^{(e)}$ correctly identifies the given image $I_{\varphi^{(e)}}$ as the target emotion via

$$\varphi^{(e)} = \arg \min_{\hat{\varphi}^{(e)} \in [-3, 3]^{|\varphi^{(e)}|}} 1 - \mathbb{F}_\theta^{(e)}(I_{\hat{\varphi}^{(e)}}). \quad (4)$$

For this estimation problem, all parameters apart from φ are fixed during the optimization, minimizing other confounding factors [9, 15, 22, 101], enforcing that only changes in facial expression influence the classifier output. Given that we cannot use a gradient-based optimizer as changes in φ result in no changes in the parameters of $\mathbb{F}_\theta$, we use the *differential evolution algorithm* [89] for optimization using a search range of $[-3, 3]$ [21, 48]. In Fig. 2, we visualize renderings for the six base emotions given the ResidualMaskingNet as classifier [65]. The supplementary material provides more examples and expressions parameters $\varphi^{(e)}$.

3.3 Measuring Systematic Change

Given our rendering pipeline $I_{\varphi^{(e)}}$, specified in the previous section, we need a score function to measure systematic changes in expression classifier behavior concerning facial symmetry. Hence, we define a facial symmetry impact score for a specific trained model $\mathbb{F}_\theta$. To be precise, we measure one score for each possible expression e predicted by the selected classifier, henceforth, $\mathbb{F}_\theta^{(e)}$.

Using $I_{\varphi^{(e)}}$ with a sampled identity, i.e., fixed α and β , we generate synthetic images for timesteps t and facial symmetries s . Now, $\mathbb{F}_\theta^{(e)}(I_{\varphi^{(e)}}(s, t))$ defines a surface, where for each s and t , we have an output activation of $\mathbb{F}_\theta$ for emotion e . Fig. 3a visualizes two of these surfaces for the *neutral* and *happy* emotion. Ideally, we would want to see no changes along the s axis in these surfaces, i.e., the model is unbiased concerning symmetry. We can measure these changes by investigating the partial derivatives $\nabla_s \mathbb{F}_\theta^{(e)}(I_{\varphi^{(e)}}(s, t))$. A positive ∇_s indicates higher model outputs for increased symmetry, which is reversed for negative ∇_s . An unbiased model activation surface (blue) is visualized in Fig. 3b. This optimal surface is characterized by $\nabla_s \mathbb{F}_\theta^{(e)}$ being zero for any valid s and t .

Of course, in reality, we do not expect the outputs of any model to stay constant for changing symmetry values. Many factors can impact the model outputs, even for small visual changes. Nevertheless, we expect an unbiased model to show no systematic behavioral changes, e.g., categorically lower outputs

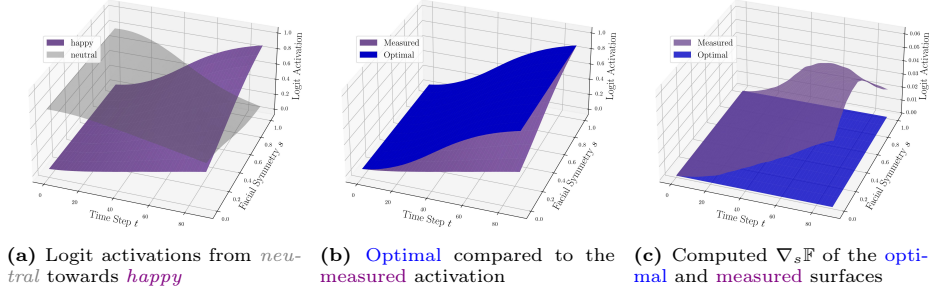


Fig. 3: Visualization of our impact score for a classifier’s *happy* logit activation: In a synthetic setting, a model was shown a face transition from *neutral* to a *happy* expression (a). A model would be invariant toward changes along the symmetry axis if $\nabla_s \mathbb{F} = 0$. However, the actual activation logits (*happy*) show a lower activation (b). This is more evident in the visualization of the estimated ∇_s in (c).

for smaller symmetry values s . Hence, a more realistic ideal surface would be a noisy version of the visualization in Fig. 3b. In other words, for an unbiased model that does not change behavior for different facial symmetry, we expect that $\mathbb{E}_{s,t}[\nabla_s \mathbb{F}_\theta^{(e)}(I_{\varphi(e)}(s, t))]$ is approximately zero for some joint distribution of symmetry values s and timesteps t . Without loss of generality, let $[0, 1]$ be a valid domain for s and t respectively, then we define our facial symmetry impact score $\mathcal{S}$ for a specific model $\mathbb{F}_\theta^{(e)}$ and for a fixed individual $I_{\varphi(e)}$ as

$$\begin{aligned} \mathcal{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)}) &= \mathbb{E}_{s,t}[\nabla_s \mathbb{F}_\theta^{(e)}(I_{\varphi(e)}(s, t))] \\ &= \int_0^1 \int_0^1 \nabla_s \mathbb{F}_\theta^{(e)}(I_{\varphi(e)}(s, t)) \cdot p(s, t) dt ds, \end{aligned} \quad (5)$$

where p is the density function describing the joint distribution of s and t .

Calculating $\mathcal{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ directly is intractable. Hence, we assume that s and t are independent and uniformly distributed. Although this is a strong assumption, we can utilize our rendering pipeline (see Sec. 3.2) to ensure these conditions in our synthetic data. By doing so, we can approximate $\mathcal{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ by evaluating $\mathbb{F}_\theta^{(e)}$ at a grid of finitely many equidistant samples of $I_{\varphi(e)}(s, t)$.

Let $\mathfrak{T}$ and $\mathfrak{S}$ be a set of equidistant time and symmetry steps in $[0, 1]$, then

$$\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)}) = \frac{1}{|\mathfrak{S}| \cdot |\mathfrak{T}|} \sum_{s \in \mathfrak{S}} \sum_{t \in \mathfrak{T}} \nabla_s \mathbb{F}_\theta^{(e)}(I_{\varphi(e)}(s, t)), \quad (6)$$

approximates $\mathcal{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$. To estimate the gradient on our finite grid of $\mathfrak{T}$ and $\mathfrak{S}$, we use the implementation of [26] by the library NumPy [35]. This algorithm minimizes the error between the actual gradient and the estimate at a grid position by solving a system of linear equations of the neighboring grid points.

While $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ enables us to investigate changes for a single individual $I_{\varphi(e)}$ with fixed α and β , we are additionally interested in explaining models $\mathbb{F}_\theta$

more globally concerning specific emotions. Hence, we define a global score for an emotion e as $\mathcal{S}(\mathbb{F}_\theta^{(e)}) = \mathbb{E}_{\alpha, \beta}[\mathcal{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})]$. For a set of N individuals $\mathcal{I}$, by using the same assumptions as in Eq. (6), we approximate $\mathcal{S}(\mathbb{F}_\theta^{(e)})$ with

$$\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)}) = \frac{1}{N} \sum_{I_{\varphi(e)} \in \mathcal{I}} \hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)}). \quad (7)$$

Testing for Statistical Significance: Values for $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$ and $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ close to zero indicate less change for varying facial symmetry values s . Further, the sign of our scores can be interpreted as over- (positive) or under-predicting (negative) an emotion for increasing symmetry. However, we also need to specify at which point values of $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$ or $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ are statistically significant.

We utilize permutation hypothesis tests [30] for this goal in which the values of $\mathbb{F}_\theta^{(e)}$ in our grid of synthetic inputs are shuffled. To control for the influence of t , i.e., the onset expression’s strength, we only shuffle values of $\mathbb{F}_\theta^{(e)}$ while fixing t . In other words, we permute along the symmetry axis in Fig. 3a. Afterward, the corresponding $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)}^{(perm.)})$ is recalculated (Eq. (6)). The process repeats K -times to generate our distribution under the null hypothesis H_0 , which is that the observed value $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ is zero. By counting how often we observe permutations where $|\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)}^{(perm.)})| > |\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})|$, we can determine a p -value. Note that the absolutes are needed because negative scores are valid. If the p -value is smaller than a significance level δ , we discard H_0 , i.e., the observed score $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi(e)})$ is statistically significant. We provide the hypothesis test pseudo-code in the supplementary material (Alg. 1).

While our approach tests for significance concerning a certain $I_{\varphi(e)}$ with fixed α and β , regarding $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$, we additionally apply the Holm-Bonferroni correction on our results [36]. This is a sequential correction method to control the family-wise error rate for repeated hypothesis tests. In other words, ensuring we do not overestimate significance, i.e., increase type-I errors, for a pre-specified δ . Finally, we report the ratio of significant results of the corrected tests, which intuitively captures how often we observe significant changes in behavior. Note that, in contrast, $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$ measures how strong (and in which direction) these changes are on average. This means statistical significance is possible even if the effect size, i.e., the actual systematic change, is low. With the scores derived above combined with the corresponding statistical hypothesis tests, we can investigate the interventional question posed in Sec. 3.1: “If the facial symmetry for an input changes, then how does the output of the expression classifier behave?”

4 Experiments and Results

Our investigation focuses on the impact of facial symmetry on data-driven expression classifiers. Before we detail our experiments, we want to state our hypothesis clearly: We expect that most classifiers show systematic differences in

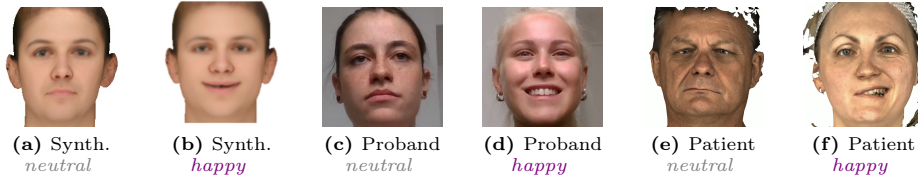


Fig. 4: We display two example faces per analyzed group (synthetic, probands, and patients). Both healthy probands and patients with unilateral facial palsy were instructed to mimic the shown emotions, similar to the FER2013 benchmark [20].

their behavior when intervening on facial symmetry. Given that one face half exhibits less movement for unilateral facial palsy and our synthetic symmetry data, we assume a reduction in logit-activations for reduced symmetry based on observations in related studies [9, 106]. Because most facial expression datasets, e.g., [20, 46, 47, 53], contain mostly healthy symmetric faces. Our study is limited to the six base emotions [23], omitting *neutral* and *contempt* for comparability.

Compared to other research that requires training [15, 43], we can assess 17 existing expression classifiers. Please note that our selection of models is not intended to be comprehensive or representative of all possible classifiers but as a first case study. We focus on a subset of models that provide code and model weights, which are likely to be applied more frequently in out-of-domain scenarios like medicine and psychology. To test our hypothesis, we perform two groups of experiments. First, we investigate if a symmetry bias exists for real-world data inside classifiers using associational methods [7, 61, 71, 73]. Second, moving up on the causal hierarchy [4], we connect facial symmetry and model output using our synthetic intervention framework and score.

4.1 Experiment 1: Observations on Real-World Data

Although intuition suggests that facial asymmetry influences model behavior, we must first check if this bias is present in expression classifiers. We utilize associational methods [7, 61, 71, 73] to show that classifiers significantly change their behavior, i.e., the red arrow exists in Fig. 1. Specifically, we attempt to validate our hypothesis on the first level of Pearl’s hierarchy [4], using real-world data recorded on 36 healthy probands (18-67 years, 17 ♂, 19 ♀) and 36 patients (25-72 years, 8 ♂, 28 ♀) with unilateral chronic synkinetic facial palsy. Probands were recorded using a RealSense camera (Intel Corporation, Santa Clara, USA) and patients with the 3dMD face system (3dMD LLC, Georgia, USA), see Fig. 4. We model two symmetry properties: the presence of facial palsy (binary) and LPIPS [108] similarity (continuous) among face sides.

The participants’ expressions are captured while mimicking the six basic emotions four times in a random order [6, 7, 9, 32] following FER2013 [20]. We focus on the *happy* expression as it most impacts emotional expressiveness [8, 76, 95]. Examples of each face group are shown in Fig. 4. We measure an average classification accuracy of 97.40% for probands and 58.25% for patients among the

	DAN [102] DDAMFN++ [110] HSEmotion [82] PosterV2 [50]	DAN [102] DDAMFN++ [110] HSEmotion [82] PosterV2 [50]	EmoNeXt-Base† [24] EmoNeXt-Large† [24] EmoNeXt-Small† [24] EmoNeXt-Tiny† [24] ResidualMaskingNet [65] Segmentation-VGG19† [98]	DAN [102] DDAMFN++ [110] PosterV2 [50]
	AffectNet7	AffectNet8	FER2013	RAFDB
Facial Palsy	✓✓✓✓	✓✓✓✓	✓✓✓✓✓✓	✗✓✓
LPIPS symmetry [108]	✓✓✓✓	✓✓✓✓	✓✓✓✓✓✓	✓✓✓

Table 1: Significance results ($p < 0.01 \rightarrow \checkmark$) of [73] on our data for three symmetry features. We investigate *happy* logits with respect to the binary facial palsy state and LPIPS [108] similarity between the face sides for real-world images of facial palsy patients and healthy probands.

FER2013 models (due to the similar recording setup), indicating the decreased performance under facial palsy. A more fine-grained analysis can be found in the supplementary material.

Following related work [9, 61, 73], we denote all significant behavior changes in Table 1 using the majority decision of three conditional independence tests [11, 27, 79] (for detailed hyperparameters see supplementary material). We find all 17 models show a statistically significant shift in their *happy* activations for varying facial symmetries in the real-world data. Please note that DAN [102], trained on RAFDB [46, 47], is the only classifier where we find no significance regarding binary facial palsy, which is a highly discretized form of symmetry. However, the continuous symmetry measure LPIPS [108] indicates the same behavior changes. These results provide first evidence for our hypothesis that expression classifiers are biased toward facial symmetry, especially concerning downstream applications.

Furthermore, we provide more qualitative investigations of the output changes in the supplementary material. For most classifiers, we observe, on average, a decrease in activations. Given the performance decrease, this observation is expected and indicates uncertainty in predicting the *happy* class. However, these are associational investigations [4], i.e., we do not isolate changes in facial symmetry from confounding factors. Hence, while we observe changes in classifier behavior on real-world data, our following interventional investigation in Sec. 4.2 provides more reliable, actionable insights.

4.2 Experiment 2: Synthetic Facial Symmetry Interventions

To confirm our hypothesis on how facial symmetry affects expression classification beyond the associational level, we perform synthetic interventions using the framework described in Sec. 3. These enable us to measure the impact of facial

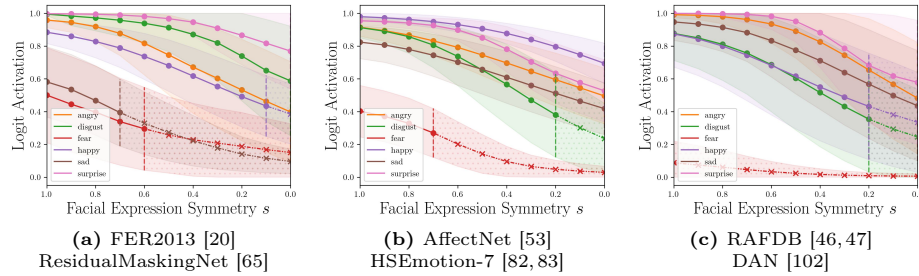


Fig. 5: We display a model’s activation (mean and std.) curve at $t = 1.0$ for each expression recognition dataset. Note that the x-axis is inverted, so we start with high symmetry. Lower symmetry generally results in lower logit activations across all expressions, with hatched lines indicating misclassification. We show the surface variants, such as Fig. 3a, in the supplementary material.

asymmetry and model output. Therefore, we create a population $\mathcal{I}$ of 200 identities sampled from a standard normal distribution (different α and β). Following Eq. (4), we optimize the facial expression $\varphi^{(e)}$ for each model and identity at $t = 1.0$ and $s = 1.0$. To apply interventions, our finite grid spans ten equidistant symmetry steps ($s \in [0, 1]$) and 90-time steps ($t \in [0, 1]$), simulating three-second expression onset. We then derive the mean logit surfaces (compare Fig. 3) of all classifiers over $\mathcal{I}$ and display selected classifiers at $t = 1.0$ in Fig. 5.

First, we observe that facial symmetry impacts each expression’s logit activation for the models irrespective of the dataset. The ResidualMaskingNet [65] trained on FER2013 [20] does not reach high activations for *fear* and *sad*. Further, they decrease even more for lower symmetry values. Expressions such as *angry*, *disgust*, or *surprise* have higher activations and seem to be affected only by more pronounced asymmetry. Especially *fear* also seems problematic for the other classifiers highlighted in Fig. 5. Notably, the same behavior holds for HSEmotion [82] and DAN [102]: Lower symmetry leads to lower output activations for all expressions, providing further evidence to confirm our hypothesis.

Next, we go one step further and quantitatively measure the impact of facial symmetry using our proposed score (Sec. 3.3). Table 2 summarizes these results averaged over all individuals $\mathcal{I}$. Remember that a positive score corresponds to increased activations for increased symmetry. Further, we report the ratio of significant systematic changes over the set of individuals $\mathcal{I}$ in Table 1 of the supplementary material. In most cases, we observe a significant impact of facial symmetry for all classifiers and expressions. Hence, this enables us to interpret the patterns we observe in Table 2 concerning the expressions and training datasets. We note the highest impact of 0.0373 for *surprise* of PosterV2 [50]. This score indicates that, on average, over the complete *surprise* onset, increasing the symmetry by one step in our simulation increased the softmax output of PosterV2 by 3.7 percentage points. However, while all 17 classifiers are significantly impacted by changes in facial symmetry, the effect size can still be small (see, for example, *fear* in Fig. 5c).

Table 2: We compute $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$, defined in Eq. (7), among our population $\mathcal{I}$. The models are grouped by the training dataset, and the $\dagger$ annotates models trained by us (see the supplementary material for details); otherwise, the provided models’ weights were used respectively. All $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$ are significant for the majority of individuals.

Dataset	$\mathbb{F}_\theta$ Model	Angry	Disgust	Fear	Happy	Sad	Surprise
AffectNet7	DAN [102]	0.0249	0.0234	0.0098	0.0242	0.0192	0.0279
	DDAMFN++ [110]	0.0050	0.0035	0.0019	0.0040	0.0027	0.0037
	HSEmotion [82]	0.0229	0.0308	0.0107	0.0239	0.0211	0.0268
	PosterV2 [50]	0.0214	0.0192	0.0133	0.0234	0.0168	0.0290
AffectNet8	DAN [102]	0.0209	0.0222	0.0091	0.0112	0.0207	0.0281
	DDAMFN++ [110]	0.0030	0.0032	0.0026	0.0028	0.0034	0.0036
	HSEmotion [82]	0.0136	0.0156	0.0076	0.0111	0.0150	0.0250
	PosterV2 [50]	0.0205	0.0228	0.0129	0.0214	0.0171	0.0269
FER2013	EmoNeXt-Small † [24]	0.0157	0.0017	0.0078	0.0195	0.0074	0.0238
	EmoNeXt-Tiny † [24]	0.0092	0.0017	0.0050	0.0124	0.0051	0.0219
	EmoNeXt-Base † [24]	0.0089	0.0009	0.0096	0.0184	0.0096	0.0227
	EmoNeXt-Large † [24]	0.0149	0.0065	0.0163	0.0207	0.0236	0.0228
	ResidualMaskingNet [65]	0.0298	0.0307	0.0099	0.0251	0.0137	0.0280
	Segm-VGG19 † [98]	0.0186	0.0010	0.0174	0.0238	0.0221	0.0206
RAFDB	DAN [102]	0.0319	0.0262	0.0017	0.0205	0.0246	0.0314
	DDAMFN++ [110]	0.0013	0.0013	0.0002	0.0134	0.0117	0.0181
	PosterV2 [50]	0.0326	0.0228	0.0054	0.0313	0.0166	0.0373

We start with broad insights about the results in Table 2, before focusing on specific models: First, all $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$ contained in Table 2 are positive. Hence, expression classifiers show, on average, lower logit activations for decreased symmetry. This result provides *interventional evidence* for our previously stated hypothesis. Second, similar to Fig. 5, we see low scores for *fear* expressions irrespective of the classifier and dataset. However, this is likely due to the often lower activations for *fear* (Fig. 5 and supplementary material). The FLAME expression space may limit accurately modeling *fear*. The shift in model outputs is, nevertheless, significant. In contrast to *fear*, the overall high scores for *happy*, *surprise*, and *angry* suggest stronger changes in model behavior for these expressions.

Seen in Table 2, models trained on the same dataset often show similar $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)})$, likely due to the latent training data distribution [91]. For FER2013 [20] classifiers, facial symmetry has a lower impact on the *disgust* expression, excluding ResidualMaskingNet [65]. Similarly, models trained on this dataset display lower scores for *sad* and *angry*. We analyze different EmoNeXt [24] model sizes. *Large* being impacted the most, indicating higher capacity could consolidate biases in the training data. In contrast, DDAMFN++ [12] shows a small effect size irrespective of the dataset. Visualizing the graphs at $t = 1.0$ for different s and the classification accuracies on the real-world data (both in the supplementary

material), we assume that the model is likely overfitting the training data. Thus, low output activations result in smaller ∇_s .

Nevertheless, we conclude that all analyzed classifiers show significant increases in output activations for higher facial symmetry, confirming our hypothesis. Given that we use interventions beyond the associational level, we verify a causal link between facial symmetry and expression classifiers’ behavior.

5 Limitations and Social Impact

Our work relies on the statistical shape model, which induces a domain shift. While many classifiers advertise the out-of-domain usage, e.g., [65, 82–84], and we additionally optimize the synthetic faces regarding model and expression (see Sec. 3.2). Nevertheless, other factors like camera angle could jointly influence the behavior. Secondly, while we focus on facial symmetry, other facial features, e.g., age or skin color, can impact classifier performance and should be considered in the future. Our framework cannot model the entire range of facial nerve damage effects on facial mimicry, as we do not account for all possible forms of facial asymmetry, e.g., synkinetic effects.

Regarding societal impact, our work investigates only existing expression classifiers. We move from the associational level to causal interventions to better understand how these black-box models operate. Such an approach could benefit other disciplines, primarily psychological and medical applications. We provide our experiments’ generative framework and evaluation code so that other researchers can simply evaluate their models.

6 Conclusion

Emotional expressiveness is crucial for communicating our internal state and for understanding other people [76, 95]. In this work, we investigate the impact of facial symmetry on 17 different expression classifiers trained on four different datasets [20, 46, 47, 53]. Extending empirical analysis, we try to answer an interventional question [4] by following insights from causal inference and explainability [73] and using an SCM (Fig. 1) together with a generative framework. We control expression and facial symmetry using a modified statistical shape model [48] to measure systematic changes with a proposed interpretable score.

We tested our hypothesis on real-world data of facial palsy patients and healthy probands using associational methods [9, 61, 73]. Here, we saw that facial palsy and the similarity between the face halves led to 33 out of 34 tests being significant. To verify these results in a more controlled manner, moving up the causal hierarchy, we employed our interventional framework to test the impact of symmetry on the models’ behavior. We observed that many classifiers, on average, decrease logit activations for lower facial symmetry.

While, in retrospect, our results align with the pre-specified intuition, we stress that our framework provides a structured way to test such hypotheses. Further, it could be extended to other features, e.g., age or skin color, given the

controllable nature of statistical shape models. These insights could also be used to grade facial palsy or to correct the classifier output for patients posthoc. Hence, we hope that our work can help researchers understand the prediction behavior of their trained expression classifiers beyond simple performance metrics.

Ethics Approval Written informed consent was obtained from all participants, including displayed individuals. The ethics committee of the Jena University Hospital approved the study (No. 2019-1539).

Acknowledgement Partially supported by Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) project 427899908 BRIDGING THE GAP: MIMICS AND MUSCLES (DE 735/15-1 and GU 463/12-1).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Baltrušaitis, T., Robinson, P., Morency, L.: OpenFace: An open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–10 (Mar 2016). <https://doi.org/10.1109/WACV.2016.7477553>
2. Baltrušaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: OpenFace 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 59–66 (2018). <https://doi.org/10.1109/FG.2018.00019>
3. Banks, C.A., Bhama, P.K., Park, J., Hadlock, C.R., Hadlock, T.A.: Clinician-Graded Electronic Facial Paralysis Assessment: The eFACE. *Plastic and Reconstructive Surgery* **136**(2), 223e (Aug 2015). <https://doi.org/10.1097/PRS.0000000000001447>
4. Bareinboim, E., Correa, J.D., Ibeling, D., Icard, T.F.: On pearl’s hierarchy and the foundations of causal inference. *Probabilistic and Causal Inference* (2022)
5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*. pp. 187–194. ACM Press, Not Known (1999). <https://doi.org/10.1145/311535.311556>
6. Büchner, T., Guntinas-Lichius, O., Denzler, J.: Improved obstructed facial feature reconstruction for emotion recognition with minimal change cyclegrams. In: *Advanced Concepts for Intelligent Vision Systems (Acivs)*. pp. 262–274. Springer-Nature (august 2023). https://doi.org/10.1007/978-3-031-45382-3_22
7. Büchner, T., Sickert, S., Volk, G.F., Anders, C., Guntinas-Lichius, O., Denzler, J.: Let’s get the faces straight - reconstructing obstructed facial features. In: *International Conference on Computer Vision Theory and Applications (VISAPP)*. SciTePress (march 2023). <https://doi.org/10.5220/0011619900003417>
8. Büchner, T., Sickert, S., Volk, G.F., Guntinas-Lichius, O., Denzler, J.: From Faces to Volumes - Measuring Volumetric Asymmetry in 3D Facial Palsy Scans. In: *Advances in Visual Computing. Lecture Notes in Computer Science*, Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-47969-4_10

9. B chner, T., Penzel, N., Guntinas-Lichius, O., Denzler, J.: The power of properties: Uncovering the influential factors in emotion classification. In: International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI) (2024), <https://arxiv.org/abs/2404.07867>, (accepted)
10. B chner, T., Sickert, S., Gra me, R., Anders, C., Guntinas-Lichius, O., Denzler, J.: Using 2d and 3d face representations to generate comprehensive facial electromyography intensity maps. In: International Symposium on Visual Computing (ISVC). pp. 136–147 (2023). https://doi.org/10.1007/978-3-031-47966-3_11, https://link.springer.com/chapter/10.1007/978-3-031-47966-3_11
11. Chalupka, K., Perona, P., Eberhardt, F.: Fast conditional independence test for vector variables with large sample sizes. arXiv preprint arXiv:1804.02747 (2018)
12. Chen, Y., Li, J., Shan, S., Wang, M., Hong, R.: From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos (Dec 2023)
13. Chen, Y., Li, W., Chen, X., Gool, L.V.: Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1841–1850 (2019)
14. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018)
15. Choithwani, M., Almeida, S., Egger, B.: PoseBias: On Dataset Bias and Task Difficulty - Is there an Optimal Camera Position for Facial Image Analysis? In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3088–3096. IEEE, Paris, France (Oct 2023). <https://doi.org/10.1109/ICCVW60793.2023.00334>
16. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on pattern analysis and machine intelligence **23**(6), 681–685 (2001)
17. Dane c ek, R., Black, M.J., Bolkart, T.: EMOCA: Emotion Driven Monocular Face Capture and Animation. CVPR p. 12 (2022)
18. Demeco, A., Marotta, N., Moggio, L., Pino, I., Marinaro, C., Barletta, M., Petraroli, A., Palumbo, A., Ammendolia, A.: Quantitative analysis of movements in facial nerve palsy with surface electromyography and kinematic analysis. Journal of Electromyography and Kinesiology **56**, 102485 (Feb 2021). <https://doi.org/10.1016/j.jelekin.2020.102485>
19. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5154–5163 (2020)
20. Dumitru, Goodfellow, I., Cukierski, W., Bengio, Y.: Challenges in representation learning: Facial expression recognition challenge (2013), <https://kaggle.com/competitions/challenges-in-representation-learning-facial-expression-recognition-challenge>
21. Egger, B., Smith, W.A.P., Tewari, A., Wuhler, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., Theobalt, C., Blanz, V., Vetter, T.: 3D Morphable Face Models-Past, Present, and Future. ACM Transactions on Graphics **39**(5), 157:1–157:38 (Jun 2020). <https://doi.org/10.1145/3395208>

22. Egger, B., Sutherland, S., Medin, S.C., Tenenbaum, J.: Identity-Expression Ambiguity in 3D Morphable Face Models. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–7. IEEE Press, Jodhpur, India (Dec 2021). <https://doi.org/10.1109/FG52635.2021.9667002>
23. Ekman, P.: An argument for basic emotions. *Cognition and Emotion* **6**(3-4), 169–200 (1992). <https://doi.org/10.1080/02699939208411068>
24. El Boudouri, Y., Bohi, A.: Emonext: an adapted convnext for facial emotion recognition. In: 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6 (2023). <https://doi.org/10.1109/MMSP59012.2023.10337732>
25. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics* **40**(4), 1–13 (Aug 2021). <https://doi.org/10.1145/3450626.3459936>
26. Fornberg, B.: Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation* **51**, 699–706 (1988), <https://api.semanticscholar.org/CorpusID:119513587>
27. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. *Advances in neural information processing systems* **20** (2007)
28. Gao, X., Su, Y., Li, X., Tao, D.: A review of active appearance models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **40**(2), 145–158 (2010)
29. Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schoenborn, S., Vetter, T.: Morphable Face Models - An Open Framework. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 75–82. IEEE, Xi'an (May 2018). <https://doi.org/10.1109/FG.2018.00021>
30. Good, P.: *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Series in Statistics, Springer New York (2013), <https://books.google.de/books?id=pk3hBwAAQBAJ>
31. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. *Adv. Neural. Inf. Process. Syst.* **19** (2006)
32. Guntinas-Lichius, O., Trentzsch, V., Mueller, N., Heinrich, M., Kutteneich, A.M., Dobel, C., et al.: High-resolution surface electromyographic activities of facial muscles during the six basic emotional expressions in healthy adults: a prospective observational study. *Scientific Reports* **13**(1), 19214 (2023)
33. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3D dense face alignment. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2020)
34. Haase, D., Rodner, E., Denzler, J.: Instance-weighted transfer learning of active appearance models. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1426–1433 (2014)
35. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. *Nature* **585**(7825), 357–362 (Sep 2020). <https://doi.org/10.1038/s41586-020-2649-2>, <https://doi.org/10.1038/s41586-020-2649-2>
36. Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2), 65–70 (1979), <http://www.jstor.org/stable/4615733>

37. Hu, Y.T., Wang, J., Yeh, R.A., Schwing, A.G.: Sail-vos 3d: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1418–1428 (2021)
38. Josifovski, J., Kerzel, M., Pregizer, C., Posniak, L., Wermter, S.: Object detection and pose estimation based on convolutional neural networks trained with synthetic data. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 6269–6276. IEEE (2018)
39. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks (Mar 2019). <https://doi.org/10.48550/arXiv.1812.04948>
40. Katsumi, S., Esaki, S., Hattori, K., Yamano, K., Umezaki, T., Murakami, S.: Quantitative analysis of facial palsy using a three-dimensional facial motion measurement system. *Auris Nasus Larynx* **42**(4), 275–283 (Aug 2015). <https://doi.org/10.1016/j.anl.2015.01.002>
41. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
42. Knoedler, L., Baecher, H., Kauke-Navarro, M., Prantl, L., Machens, H.G., Scheuermann, P., Palm, C., Baumann, R., Kehrer, A., Panayi, A.C., Knoedler, S.: Towards a Reliable and Rapid Automated Grading System in Facial Palsy Patients: Facial Palsy Surgery Meets Computer Science. *Journal of Clinical Medicine* **11**(17), 4998 (Aug 2022). <https://doi.org/10.3390/jcm11174998>
43. Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., Vetter, T.: Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2174–217409. IEEE, Salt Lake City, UT, USA (Jun 2018). <https://doi.org/10.1109/CVPRW.2018.00283>
44. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. *Nature communications* **10**(1), 1096 (2019)
45. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. pp. 165–172. Siggraph '00, ACM Press/Addison-Wesley Publishing Co., USA (2000). <https://doi.org/10.1145/344779.344862>
46. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* **28**(1), 356–370 (2019)
47. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
48. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* **36**(6), 1–17 (Nov 2017). <https://doi.org/10.1145/3130800.3130813>
49. Lin, C.Z., Nagano, K., Kautz, J., Chan, E.R., Iqbal, U., Guibas, L., Wetzstein, G., Khamis, S.: Single-Shot Implicit Morphable Faces with Consistent Texture Parameterization. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. pp. 1–12 (Jul 2023). <https://doi.org/10.1145/3588432.3591494>

50. Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A.: POSTER++: A simpler and stronger facial expression recognition network (Feb 2023)
51. Matthews, I., Baker, S.: Active appearance models revisited. *International journal of computer vision* **60**, 135–164 (2004)
52. Medin, S.C., Egger, B., Cherian, A., Wang, Y., Tenenbaum, J.B., Liu, X., Marks, T.K.: MOST-GAN: 3D Morphable StyleGAN for Disentangled Face Image Manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence* **36**(2), 1962–1971 (Jun 2022). <https://doi.org/10.1609/aaai.v36i2.20091>
53. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* **10**(1), 18–31 (2019). <https://doi.org/10.1109/TAFFC.2017.2740923>
54. Nachbar, F., Stolz, W., Merkle, T., Cognetta, A.B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O., Plewig, G.: The abcd rule of dermatoscopy. high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology* **30** 4, 551–9 (1994), <https://api.semanticscholar.org/CorpusID:4860343>
55. Neumann, T., Lorenz, A., Volk, G., Hamzei, F., Schulz, S., Guntinas-Lichius, O.: Validierung einer Deutschen Version des Sunnybrook Facial Grading Systems. *Laryngo-Rhino-Otologie* **96**(03), 168–174 (Nov 2016). <https://doi.org/10.1055/s-0042-111512>
56. Nowruzzi, F.E., Kapoor, P., Kolhatkar, D., Hassanat, F.A., Laganieri, R., Rebut, J.: How much real data do we actually need: Analyzing object detection performance using synthetic and real data. *arXiv preprint arXiv:1907.07061* (2019)
57. Özsoy, U., Uysal, H., Hizay, A., Sekerci, R., Yildirim, Y.: Three-dimensional objective evaluation of facial palsy and follow-up of recovery with a handheld scanner. *Journal of Plastic, Reconstructive & Aesthetic Surgery* p. S1748681521002552 (Jun 2021). <https://doi.org/10.1016/j.bjps.2021.05.003>
58. Patel, A., Islam, S.M.S., Murray, K., Goonewardene, M.S.: Facial asymmetry assessment in adults using three-dimensional surface imaging. *Progress in Orthodontics* **16**(1), 36 (Oct 2015). <https://doi.org/10.1186/s40510-015-0106-9>
59. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D Face Model for Pose and Illumination Invariant Face Recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 296–301. IEEE, Genova, Italy (Sep 2009). <https://doi.org/10.1109/AVSS.2009.58>
60. Pearl, J.: *Causality*. Cambridge university press (2009)
61. Penzel, N., Kierdorf, J., Roscher, R., Denzler, J.: Analyzing the behavior of cauliflower harvest-readiness models by investigating feature relevances. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 572–581. IEEE (2023)
62. Penzel, N., Reimers, C., Bodesheim, P., Denzler, J.: Investigating neural network training on a feature level using conditional independence. In: *European Conference on Computer Vision*. pp. 383–399. Springer (2022)
63. Perarnau, G., Van De Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355* (2016)
64. Peters, J., Janzing, D., Schölkopf, B.: *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press (2017)
65. Pham, L., Vu, T.H., Tran, T.A.: Facial expression recognition using residual masking network. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4513–4519 (2021). <https://doi.org/10.1109/ICPR48806.2021.9411919>

66. Piao, J., Sun, K., Wang, Q., Lin, K.Y., Li, H.: Inverting generative adversarial renderer for face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15628 (2021)
67. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: GAN-imation: Anatomically-aware facial animation from a single image. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 835–851. Springer International Publishing, Cham (2018)
68. Pumarola, A., Agudo, A., Sanfeliu, A., Moreno-Noguer, F.: Unsupervised person image synthesis in arbitrary poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8620–8628 (2018)
69. Qiu, Z., Li, Y., He, D., Zhang, Q., Zhang, L., Zhang, Y., Wang, J., Xu, L., Wang, X., Zhang, Y., Yu, J.: SCULPTOR: Skeleton-Consistent Face Creation Using a Learned Parametric Generator. *ACM Transactions on Graphics* **41**(6), 213:1–213:17 (Nov 2022). <https://doi.org/10.1145/3550454.3555462>
70. Reichenbach, H.: The direction of time, vol. 65. Univ of California Press (1956)
71. Reimers, C., Bodesheim, P., Runge, J., Denzler, J.: Conditional adversarial debiasing: Towards learning unbiased classifiers from biased data. In: DAGM German Conference on Pattern Recognition. pp. 48–62. Springer (2021)
72. Reimers, C., Penzel, N., Bodesheim, P., Runge, J., Denzler, J.: Conditional dependence tests reveal the usage of abcd rule features and bias variables in automatic skin lesion classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1810–1819 (2021)
73. Reimers, C., Runge, J., Denzler, J.: Determining the relevance of features for deep neural networks. In: European Conference on Computer Vision. Springer (2020)
74. Ribeiro, M.T., Singh, S., Guestrin, C.: “why should i trust you?”: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), <https://api.semanticscholar.org/CorpusID:13029170>
75. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 2016 fourth international conference on 3D vision (3DV). pp. 460–469. IEEE (2016)
76. Roberts, W., Strayer, J.: Empathy, emotional expressiveness, and prosocial behavior. *Child development* **67**(2), 449–470 (1996)
77. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016)
78. Ross, B.G., Fradet, G., Nedzelski, J.M.: Development of a Sensitive Clinical Facial Grading System. *Otolaryngology–Head and Neck Surgery* **114**(3), 380–386 (Mar 1996). <https://doi.org/10.1016/S0194-59989670206-1>
79. Runge, J.: Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: International Conference on Artificial Intelligence and Statistics. PMLR (2018)
80. Saleh, F.S., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M.: Effective use of synthetic data for urban scene semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 84–100 (2018)
81. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3752–3761 (2018)

82. Savchenko, A.: Facial expression recognition with adaptive frame rate based on multiple testing correction. In: International Conference on Machine Learning. vol. 202. PMLR (2023), <https://proceedings.mlr.press/v202/savchenko23a.html>
83. Savchenko, A.V.: Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 2359–2366 (Jun 2022)
84. Savchenko, A.V., Savchenko, L.V., Makarov, I.: Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing* (2022)
85. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* **128**, 336 – 359 (2016), <https://api.semanticscholar.org/CorpusID:15019293>
86. Shah, R.D., Peters, J.: The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* **48**(3), 1514–1538 (2020)
87. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. *ArXiv abs/1706.03825* (2017), <https://api.semanticscholar.org/CorpusID:11695878>
88. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. *CoRR abs/1412.6806* (2014), <https://api.semanticscholar.org/CorpusID:12998557>
89. Storn, R., Price, K.: Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization* **11**(4), 341–359 (Dec 1997). <https://doi.org/10.1023/A:1008202821328>
90. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning (2017), <https://api.semanticscholar.org/CorpusID:16747630>
91. Sutton, R.: The bitter lesson (2019)
92. Takmaz, A., Schult, J., Kaftan, I., Akçay, M., Leibe, B., Sumner, R., Engelmann, F., Tang, S.: 3d segmentation of humans in point clouds with synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1292–1304 (2023)
93. Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H.P., Pérez, P., Zöllhofer, M., Theobalt, C.: StyleRig: Rigging StyleGAN for 3D control over portrait images, CVPR 2020. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2020)
94. Thalhammer, S., Patten, T., Vincze, M.: Sydpote: Object detection and pose estimation in cluttered real-world depth images trained using only synthetic data. In: 2019 International Conference on 3D Vision (3DV). pp. 106–115. IEEE (2019)
95. Thompson, R.A.: Empathy and emotional understanding: The early development of empathy. *Empathy and its development* **119**, 145 (1987)
96. Tremblay, J., To, T., Birchfield, S.: Falling things: A synthetic dataset for 3d object detection and pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2038–2041 (2018)
97. Vanherle, B., Moonen, S., Van Reeth, F., Michiels, N.: Analysis of training object detection models with synthetic data. *arXiv preprint arXiv:2211.16066* (2022)
98. Vignesh, S., Savithadevi, M., Sridevi, M., Sridhar, R.: A novel facial emotion recognition model using segmentation VGG-19 architecture. *International Journal*

- of Information Technology **15**(4), 1777–1787 (Apr 2023). <https://doi.org/10.1007/s41870-023-01184-z>
99. Wagner, N., Botsch, M., Schwanecke, U.: SoftDECA: Computationally Efficient Physics-Based Facial Animations. In: Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. pp. 1–11. MIG '23, Association for Computing Machinery, New York, NY, USA (Nov 2023). <https://doi.org/10.1145/3623264.3624439>
 100. Wasi, A.T., Šerbetar, K., Islam, R., Rafi, T.H., Chae, D.K.: ARBEx: Attentive Feature Extraction with Reliability Balancing for Robust Facial Expression Learning (Jul 2023)
 101. Weiherer, M., Klein, F., Egger, B.: Approximating Intersections and Differences Between Linear Statistical Shape Models Using Markov Chain Monte Carlo. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6352–6361. IEEE, Waikoloa, HI, USA (Jan 2024). <https://doi.org/10.1109/WACV57701.2024.00624>
 102. Wen, Z., Lin, W., Wang, T., Xu, G.: Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition. *Biomimetics* **8**(2), 199 (May 2023). <https://doi.org/10.3390/biomimetics8020199>
 103. Wu, Z., Wang, L., Wang, W., Shi, T., Chen, C., Hao, A., Li, S.: Synthetic data supervised salient object detection. In: Proceedings of the 30th ACM international conference on multimedia. pp. 5557–5565 (2022)
 104. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
 105. Yang, L., Zoss, G., Chandran, P., Gross, M., Solenthaler, B., Sifakis, E., Bradley, D.: Learning a Generalized Physical Face Model From Data (Feb 2024)
 106. Yang, Y., Zhang, H., Katabi, D., Ghassemi, M.: Change is hard: A closer look at subpopulation shift. arXiv preprint arXiv:2302.12254 (2023)
 107. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)
 108. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Apr 2018). <https://doi.org/10.48550/arXiv.1801.03924>
 109. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Apr 2018). <https://doi.org/10.48550/arXiv.1801.03924>
 110. Zhang, S., Zhang, Y., Zhang, Y., Wang, Y., Song, Z.: A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition. *Electronics* **12**(17), 3595 (Jan 2023). <https://doi.org/10.3390/electronics12173595>
 111. Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., Qiao, Y.: Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In: 2019 International Conference on Multimodal Interaction. pp. 562–566 (Oct 2019). <https://doi.org/10.1145/3340555.3355713>
 112. Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Wu, Menghua and Shen, Q., Yang, R., Cao, X.: FaceScape: 3D facial dataset and benchmark for single-view 3D face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023)

A Structural Causal Models

Here we include a technical definition of structural causal models used in our work.

Definition 1 (Structural Causal Model in [60, Sec. 7.1.1] and [4, Def. 1]).

A structural causal model (SCM) is defined as a 4-tuple $M = (U, V, F, P)$, where U is a set of exogenous variables describing outside factors, $V = \{V_1, \dots, V_n\}$ is the set of endogenous variables we measure in our model, $F = \{f_1, \dots, f_n\}$ is a set containing functions f_i that describe the functional relationships, and P is a joint probability distribution over U . Further, each V_i has a set of parents PA_i that functionally determine V_i together with some exogenous variables $U_i \subseteq U$. These parents PA_i are a subset of $V \setminus \{V_i\}$. For settings pa_i of parents PA_i and u_i of the exogenous variables U_i , f_i determines the value $v_i = f_i(pa_i, u_i)$ of V_i .

Each causal model M can be visualized as directed graphs. Here, each variable V_i in V defines a node, and we draw directed links from all parents PA_i into V_i . Using such a model M , we can investigate questions of the following nature: given observed evidence, e.g., $V_j = v_j$, what is the probability of a statement A happening? Further, performing a *do*-action on $V_i \in V$ is equivalent to removing the dependency f_i and instead forcing V_i to a constant value x . In other words, we set F to F_x with $F_x = \{f_j : V_j \neq V_i\} \cup \{V_i \leftarrow x\}$ [4].

B Measuring Systematic Change - Significance Test

In Algorithm 1, we provide detailed pseudo code for our proposed shuffle hypothesis test regarding the significance of $\hat{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(e)}})$. Further in Fig. 6, we visualize the estimated null-distribution as well as the originally measured score. We see that randomly shuffling observations along the symmetry axis results in a symmetrical distribution centered around zero, i.e., no systematic dependence on the facial symmetry. The original score, for the example in Fig. 6, is not typical for the estimated null-distribution leading to a low p-value. Table 3 contains the number of individuals per classifier and expression for which Algorithm 1 together with the Holm-Bonferroni correction [36] is significant. For this analysis, we perform the shuffle test with 10K iterations. All 17 classifiers show significant behavior changes concerning facial symmetry for all expressions and a majority of individuals.

C Additional Details Experiment 1

This section gives an overview of the prediction accuracy of all 17 expression classifiers achieved on our real-world data, consisting of healthy probands and patients with unilateral facial palsy. Further, we detail the hyperparameter choices in our experiments regarding the associational methods to infer whether a causal link exists between facial symmetry and model prediction behavior. Lastly, we include some additional visualizations regarding the symmetry features.

Algorithm 1 Testing for statistical significance of $\hat{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(e)}})$.

Require: grid of predictions $\mathbb{F}_\theta^{(e)}(I_{\varphi^{(e)}}(s, t))$ ▷ Gridsize is $S \times T$
Require: integer $K > 0$ ▷ Number of Permutations
Require: $\delta \in (0, 1)$ ▷ Significance Level

$p \leftarrow 0.0$
 $\sigma_{orig.} \leftarrow \hat{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(e)}})$ ▷ Estimate the original statistic
for $i \in \{1, \dots, K\}$ **do**
 $\mathbb{F}_\theta^{(e)}(I_{\varphi^{(perm.)}}^{(perm.)}(s, t)) \leftarrow \text{permute}(\mathbb{F}_\theta^{(e)}(I_{\varphi^{(e)}}(s, t)), \text{axis} = 0)$ ▷ Shuffle along Symmetry Axis
 $\sigma_{perm.} \leftarrow \hat{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(perm.)}}^{(perm.)})$
 if $|\sigma_{perm.}| > |\sigma_{orig.}|$ **then** ▷ Absolutes because our statistic is two sided
 $p \leftarrow p + 1/K$ ▷ Increment the p -value
 end if
end for
if $p < \delta$ **then**
 return $\hat{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(e)}})$ is significant.
else
 return $\hat{S}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(e)}})$ is not significant.
end if

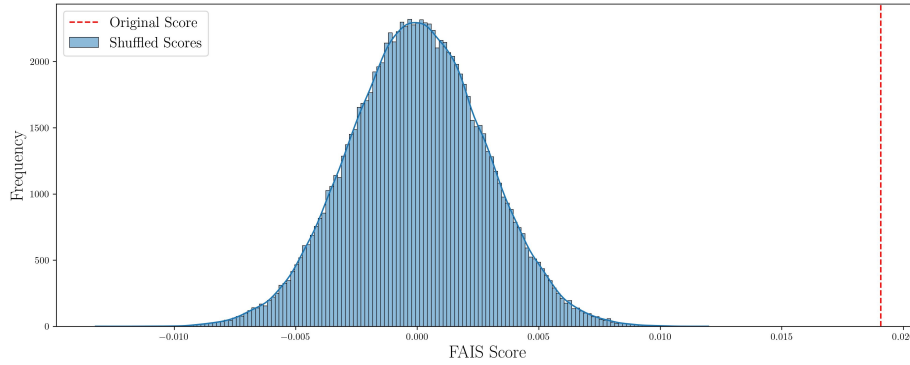


Fig. 6: Using the shuffle test, outlined in Algorithm 1, we plot the resulting scores for 100000 permutations in a histogram. Our tests use a significance threshold of $p < (0.05)$. The original model score for a single individual, shown as a red dashed line, lies clearly outside the computed null distribution and is thus significant.

Table 3: We report how many of the 200 individuals, using the Holm-Bonferroni [36] corrected p-values, have been significant ($p < 0.05$). We can see that the majority of all results are significant, confirming our hypothesis that facial symmetry impacts the internal decision rules.

Dataset	Model	Angry	Disgust	Fear	Happy	Sad	Surprise
AffectNet7	DAN [102]	200	200	200	200	200	200
	DDAMFN++ [110]	200	200	200	200	200	200
	HSEmotion [82]	200	200	200	200	200	200
	PosterV2 [50]	200	199	200	200	200	200
AffectNet8	DAN [102]	200	200	200	200	200	200
	DDAMFN++ [110]	200	200	198	200	200	200
	HSEmotion [82]	200	200	200	200	200	200
	PosterV2 [50]	200	200	200	200	200	200
FER2013	EmoNeXt-Tiny [†] [24]	200	200	200	200	200	200
	EmoNeXt-Small [†] [24]	199	194	200	200	200	200
	EmoNeXt-Base [†] [24]	163	137	200	200	200	200
	EmoNeXt-Large [†] [24]	133	187	196	195	198	187
	ResidualMaskingNet [65]	200	194	199	200	200	200
	Segmentation-VGG19 [†] [98]	199	148	200	148	200	129
RAFDB	DAN [102]	200	199	200	200	200	200
	DDAMFN++ [110]	200	200	200	200	200	200
	PosterV2 [50]	200	194	200	200	195	200

C.1 Real-World Prediction Accuracy

We are interested in the overall prediction accuracy of the model on our real-world data set consisting of 36 healthy probands and 36 patients with unilateral facial palsy. Both were instructed to mimic a *happy* expression. The probands repeated the information four times in two sessions, yielding 288 images. The patients followed the same instruction video during a ten-day biofeedback training at the hospital. They also repeated the exercise four times during a session on the first, third, and last day of therapy. An additional fourth session was offered after six months but was not followed up by some patients. Thus, we obtained 503 images for the patients.

In Table 4, we display the prediction accuracy of the *happy* emotion. We see strong differences per model and group. Therefore, we also denote the average accuracy per model and dataset to understand how we can see a particular trend per dataset. As expected and shown in the main paper, the performance of the models degrades for images that contain some form of facial asymmetry (either simulated at $s = 0.0$, $s = 0.5$, or actual facial palsy). Thus, we assume that facial symmetry is the underlying cause impacting the internal decision rules of the black box classifiers. We also see that the DDAMFN++ model trained on the AffectNet similarly performs worse on our real-world data than on the synthetic

Table 4: We evaluated each classifier on the faces of the healthy probands and patients with unilateral facial palsy mimicking the *happy* facial expression. Low accuracy is displayed in a **dark shade**, and high accuracy is displayed in a **light shade**. Models trained on FER2013 especially seem to work well on our data set. Models trained on RAFDB seem to be less suitable. Further, we provide the mean accuracy of all models per data set (with at least one correct classification).

Dataset	Model	$s = 0.0$	$s = 0.5$	$s = 1.0$	Probands	Patients
AffectNet7	DAN [102]	57.50%	99.00%	100.00%	94.44%	62.82%
	DDAMFN++ [110]	0.00%	0.00%	19.50%	0.35%	0.20%
	HSEmotion [82]	96.00%	100.00%	100.00%	0.00%	0.00%
	PosterV2 [50]	87.00%	100.00%	100.00%	97.92%	61.23%
	Average	80.17%	99.67%	79.88%	64.24%	41.42%
AffectNet8	DAN [102]	0.00%	30.50%	88.50%	91.32%	45.92%
	DDAMFN++ [110]	0.00%	8.00%	45.00%	0.35%	6.56%
	HSEmotion [82]	0.00%	0.00%	97.50%	0.00%	0.00%
	PosterV2 [50]	0.00%	20.50%	99.00%	89.93%	36.58%
	Average	0.00%	19.67%	82.50%	60.53%	29.69%
FER2013	EmoNeXt-Base [†] [24]	10.50%	71.00%	97.50%	99.65%	70.38%
	EmoNeXt-large [†] [24]	28.50%	77.50%	100.00%	98.26%	63.42%
	EmoNeXt-small [†] [24]	10.50%	68.50%	100.00%	98.96%	66.40%
	EmoNeXt-tiny [†] [24]	0.00%	12.50%	87.00%	97.92%	58.45%
	ResidualMaskingNet [65]	33.50%	89.00%	100.00%	92.36%	59.05%
	Segmentation-VGG19 [†] [98]	2.00%	39.50%	99.00%	97.22%	31.81%
	Average	17.00%	59.67%	97.25%	97.40%	58.25%
RAFDB	DAN [102]	30.00%	65.50%	94.50%	32.29%	54.27%
	DDAMFN++ [110]	69.00%	99.00%	100.00%	90.97%	76.34%
	PosterV2 [50]	25.50%	80.00%	100.00%	8.68%	38.17%
	Average	41.50%	81.50%	98.17%	43.98%	56.26%
Total	Average	40.91%	64.03%	89.85%	72.71%	48.77%

data we use for our intervention framework. Interestingly, the RAFDB-provided checkpoints seem more robust, at least in the case of *happy*.

Given that we also follow a similar experimental setup as in FER2013, models trained on it have the best performance on our data, observable in the table. Several reasons could be involved; either mimicry and *natural* facial expression have some inherent differences, the model source on public data (and human annotated) cannot differentiate, or the impact of confounding factors like camera pose. Lighting ensures that the models focus more on facial expressions.

Models such as PosterV2 perform well in our synthetic framework (likely due to the optimized expression parameters). Still, they seemed to overfit on the training data RAFDB as they performed worse on the probands but somehow better on the patients.

C.2 Feature Attribution Hyperparameter Choices

Our main experiment 2 tests the statistical dependence between expression classifier outputs and facial symmetry. We focus on the *happy* logit and find that most models change their behavior significantly for variations in facial symmetry. We employ the feature attribution method described in [73] toward this goal. This method frames supervised learning as an SCM [60] and tests whether network predictions and a pre-defined feature (facial symmetry) are conditionally independent given the reference annotation. If we have to discard this null hypothesis, we know that the classifier output values vary significantly for changes in the investigated feature. This procedure is motivated by Reichenbach’s common cause principle [70].

Clearly, the choice of conditional independence test is an important hyperparameter choice to ensure that the results are reliable. Further, Shah and Peters [86] prove that there is no optimal test that can control type-I errors, i.e., false positives, irrespective of the joint latent distribution in the non-parametric case. Because we have no knowledge about the joint distribution of all variables important in our analysis, we are exactly in the non-parametric case. Here, we follow previous work [9, 61, 62, 72] and select multiple non-linear tests. Specifically, we select conditional HSIC [27], CMiknn [79], and FCIT [11]. We consider the result from all three tests and report the majority decision [72].

The selected conditional independence tests themselves have different hyperparameter choices. First, for conditional HSIC [27], we have to select a suitable kernel function. We follow the suggestion of the authors and select the common radial basis functions kernel. Additionally, we use the heuristic by Gretton et al. [31] to approximate suitable kernel widths for all of our three variables. Second, similarly for CMiknn [79], we follow the suggested hyperparameter settings. Specifically, we set $k_{perm.}$, i.e., the neighborhood size, to five and use ten percent of the data to estimate the conditional mutual information ($k_{CMI} = 0.1 \cot n$ for n data points). Lastly, for FCIT [11], we again follow the suggestions by the authors. In other words, we set the number of data permutations to eight and use ten percent of the data to calculate the test statistic.

C.3 Additional Visualizations Regarding Logit Activations

Following previous work [9], we visualize the difference in the *happy* logit behavior between the healthy probands and facial palsy patients. Fig. 7 contains these results split between the training datasets of the 17 models we investigate in this work. However, these are associational investigations, i.e., of the first level of the PCH [4]. In other words, we do not isolate changes in facial symmetry from confounding factors and it is highly likely that other features correlate with the presence of facial palsy. Hence, while we observe changes in classifier behavior on real data, our interventional investigation is more reliable and provides actionable insights.

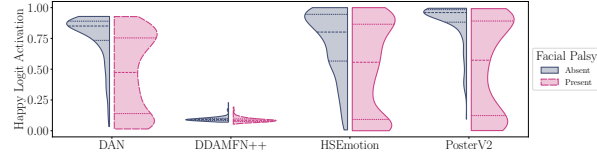
Nevertheless, Fig. 7 shows a decrease in *happy* activations for most models. This is congruent with the aggregated performance results in Table 4. Further,

these results are in line with our insights gained using our interventional framework: asymmetry results in lower activations for the *happy* class. Interestingly, we observe a slight deviation for models trained on the RAFDB. Here DAN [102], and PosterV2 [50] show higher activations and improved performance. Nonetheless, both models still struggle with facial palsy patients and are outperformed by DDAMMFN++ [110] trained on the same dataset.

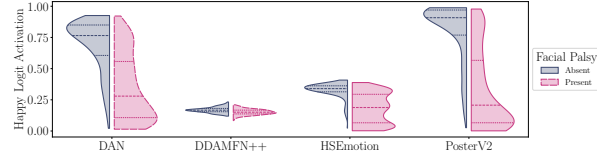
Additionally, we also visualize the results for the continuous LPIPS [109] symmetry. For regressing the mean and standard deviation, we use a window regression approach as described in [61]. We display these visualizations in Fig. 8.

Overall, we observe for most models a decrease in logit activations for decreasing facial symmetry. Hence, these results are congruent with the findings made in the main paper and Fig. 7. Furthermore, we again observe a very small effect size for DDAMFS++ [110] for both features. These findings are in agreement with the noted performance in Table 4.

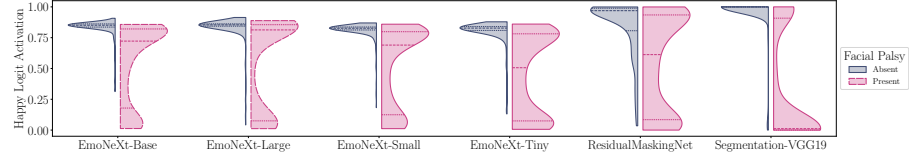
Nevertheless, we want to highlight two additional observations: First, in Fig. 8d, we observe an unexpected increase in activations. While these are associational insights, i.e., there are many possible reasons, these increases are also visible in Fig. 7d. Second, while for most models in Fig. 8, we observe a decrease in logit activations for lower facial symmetry, we note a smaller increase again for the most asymmetric faces.



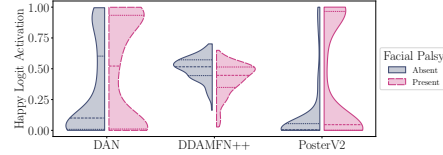
(a) Shift in output behavior for classifiers trained on AffectNet7 [53] with respect to facial palsy.



(b) Shift in output behavior for classifiers trained on AffectNet8 [53] with respect to facial palsy.



(c) Shift in output behavior for classifiers trained on FER2013 [20] with respect to facial palsy.



(d) Shift in output behavior for classifiers trained on RAFDB [46,47] with respect to facial palsy. Note that we find the behavior shift for the DAN [102] model is not significant.

Fig. 7: We follow [9] and visualize the differences in the classifiers *happy* logit distribution for healthy probands and facial palsy patients. Here 7a - 7d contain models trained on the indicated dataset respectively.

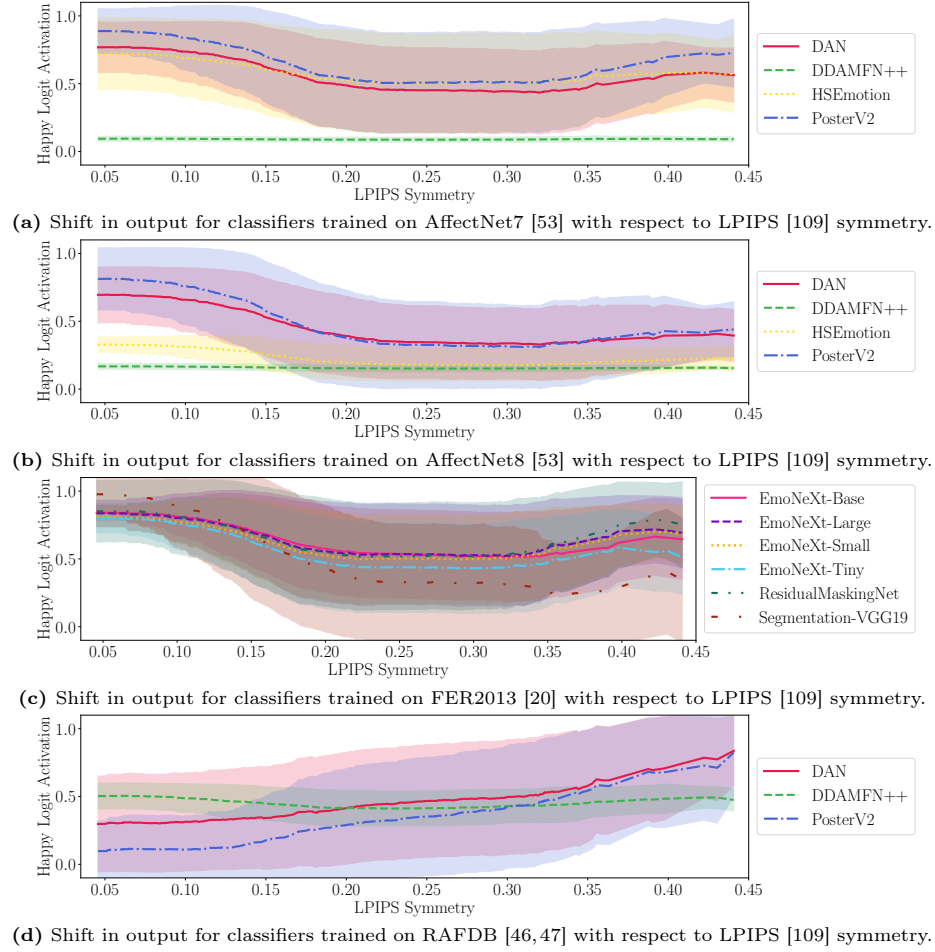


Fig. 8: We follow [9,61] and regress the shift in the classifiers *happy* logit distribution for measured LPIPS [109] symmetry scores of healthy probands and facial palsy patients. Here 8a - 8d contain models trained on the indicated dataset respectively. Note that higher LPIPS corresponds to lower symmetry [109].

Table 5: Using our intervention framework, we optimized each expression classifier $\mathbb{F}_\theta$ using logit activation for each of the six base emotions. We display the average logit activation per model and emotion. Low activation is displayed in a **darks shade**, and high activation is displayed in a **light shade**. We observe that *fear* has a generally low activation, indicating that the models have issues classifying fear or that the FLAME expression cannot model fear.

Dataset	Model	Angry	Disgust	Fear	Happy	Sad	Surprise
AffectNet7	DAN [102]	0.862	0.853	0.446	0.842	0.702	0.917
	DDAMFN++ [110]	0.356	0.331	0.136	0.220	0.233	0.292
	HSEmotion [82]	0.915	0.913	0.403	0.979	0.824	0.954
	PosterV2 [50]	0.835	0.931	0.505	0.950	0.747	0.931
AffectNet8	DAN [102]	0.776	0.805	0.416	0.464	0.732	0.881
	DDAMFN++ [110]	0.237	0.211	0.122	0.228	0.212	0.316
	HSEmotion [82]	0.595	0.759	0.349	0.340	0.590	0.826
	PosterV2 [50]	0.814	0.911	0.499	0.666	0.725	0.928
FER2013	EmoNeXt-Tiny [†] [24]	0.400	0.083	0.269	0.508	0.278	0.821
	EmoNeXt-Small [†] [24]	0.727	0.088	0.342	0.752	0.345	0.885
	EmoNeXt-Base [†] [24]	0.548	0.076	0.432	0.720	0.465	0.884
	EmoNeXt-Large [†] [24]	0.902	0.352	0.644	0.886	0.829	0.862
	ResidualMaskingNet [65]	0.959	0.995	0.500	0.884	0.581	0.997
	Segmentation-VGG19 [†] [98]	0.818	0.067	0.725	0.976	0.846	0.919
RAFDB	DAN [102]	0.991	0.877	0.088	0.874	0.947	0.999
	DDAMFN++ [110]	0.077	0.070	0.013	0.585	0.808	0.771
	PosterV2 [50]	0.993	0.996	0.312	0.982	0.987	1.000

D Additional Details Experiment 2

This section details the behavior analysis of the 17 expression classifiers on our synthetic intervention data. We start with setup and information about the facial expression optimization before displaying the sampled individuals. Afterward, we display the facial expressions achieved during the optimization per classifier for an individual. Finally, we visualize the resulting activation surfaces.

D.1 Classifier Facial Expression Optimization

Our experiments optimized each classifier $\mathbb{F}_\theta$ regarding the six base emotions. Therefore, we report the average logit activation per model and emotion reached in Table 5. We can observe several interesting properties in the logit activation. First, not all models can reach high logit activation based on facial expression changes. This indicates that models also leverage other facial information while classifying facial expressions. Furthermore, we observe that fear has a low activation among all classifiers except SegmentationVgg19 [98]. The *surprise* facial expression has a high activation among all classifiers, whereas DDAMFN++ [110] is the sole outlier; overall, reached activation is low.

D.2 Individuals

We provide an overview of all created individuals in Fig. 9. The data can be downloaded here: <https://doi.org/10.6084/m9.figshare.27074587.v1>. All resemblance to existing people is not intended and could only result from the underlying FLAME geometry model [48] and the texture from the BaselFace-Model [59].

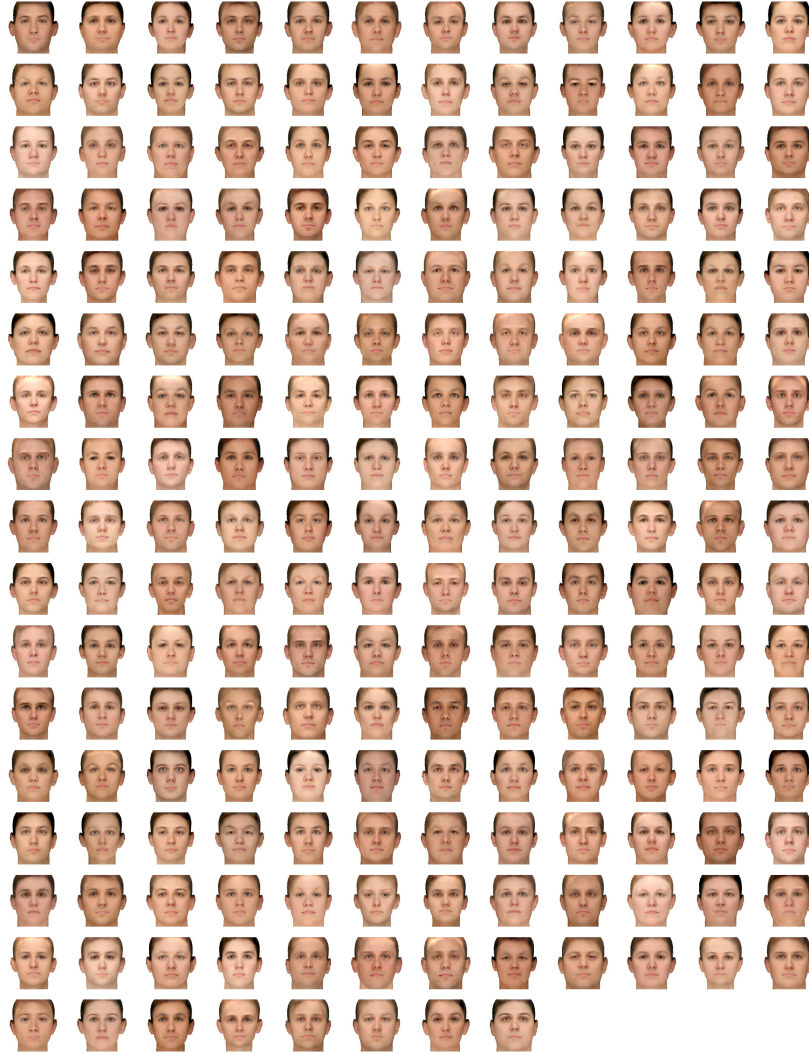


Fig. 9: 200 individual population $\mathcal{I}$

D.3 Average Facial Expression

Together with the reached logit activations, see Table 5, we are interested in the resulting facial expression. These should depict the internal representation of the respective emotion and give insight into what each classifier assumes. Furthermore, we assume the underlying base dataset influences the expression.

Average Facial Expression - Per Dataset Using our generative facial expression network, we can now create a representation of how different classifiers represent the underlying training dataset. This means the expression vectors of all 200 individuals per dataset and model are averaged and shown in Fig. 10. This visualization gives an intuitive feeling about the underlying facial expression per FER benchmark [20, 46, 47, 53]. Looking at the expression columns, we see that all interpretations of a face are slightly different. For example, for *angry*, the mouth frowning angles are different. For *disgust* the mouth is slightly opened compared to *angry*. For *fear* we can clearly see that raising the eye brows is common. The *happy* expression varies either with a wider grin or the opening state of the eye. The *sad* expression varies in the intensity of the frowning, but the eyebrows are not activated by the corrugator muscle. Also, the eyes are generally closed. For the *surprise* expression, we can see wide-open eyes and raised eyebrows in the shared interpretation.

Even though they are similar in their visual state, the intensity and expressiveness are different per model and could be the underlying reason for differences in the model architectures or the data used in the benchmark.

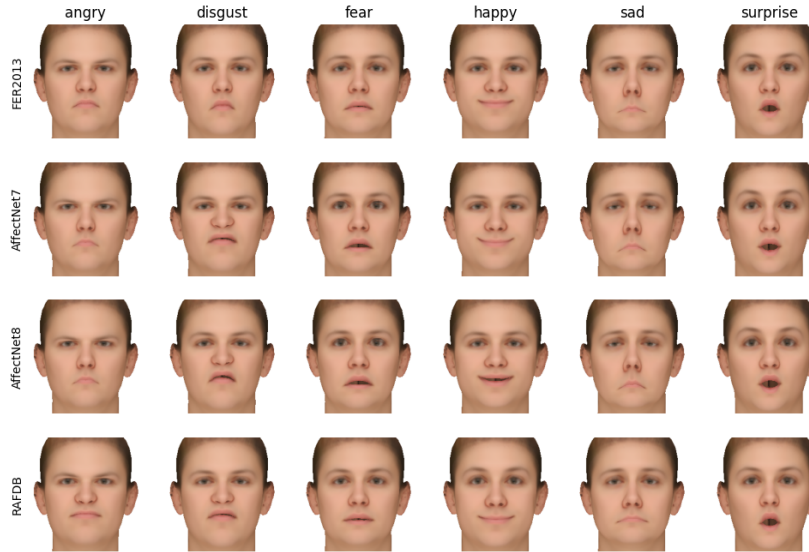


Fig. 10: Average Facial Expression used for classification based on the underlying training dataset.

Average Facial Expression - AffectNet7 Fig. 11 contains the average facial expressions for models trained on AffectNet7 [53].

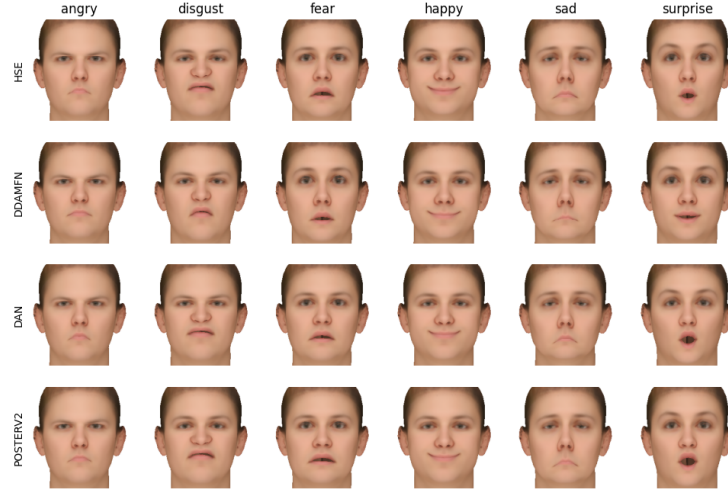


Fig. 11: The average facial expressions for models trained on AffectNet7 [53]

Average Facial Expression - AffectNet8 Fig. 12 contains the average facial expressions for models trained on AffectNet8 [53].

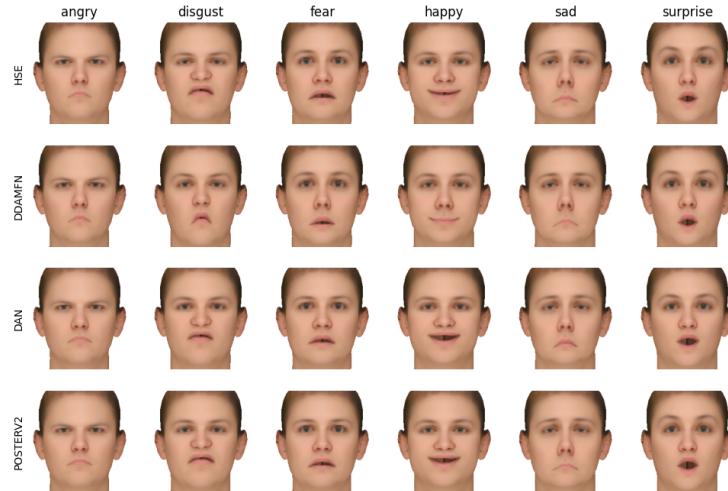


Fig. 12: The average facial expressions for models trained on AffectNet8 [53]

Average Facial Expression - FER2013 Fig. 13 contains the average facial expressions for models trained on FER2013 [20].

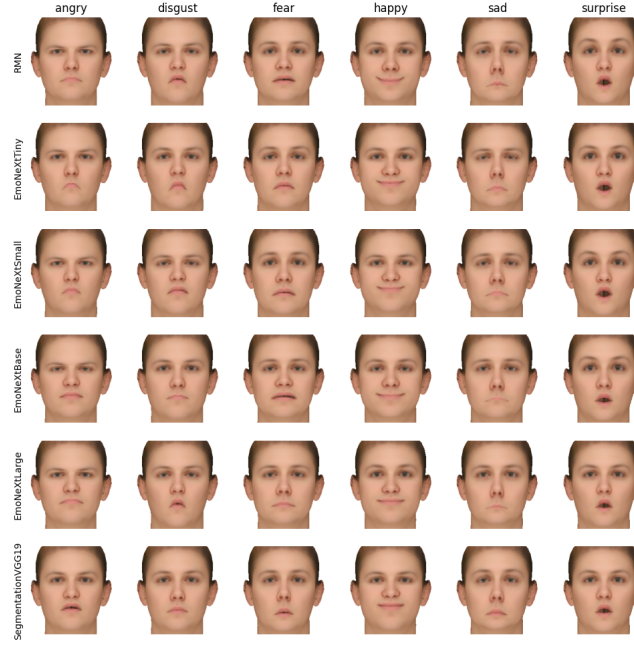


Fig. 13: The average facial expressions for models trained on FER2013 [20]

Average Facial Expression - RAFDB Fig. 14 contains the average facial expressions for models trained on RAFDB [46, 47].

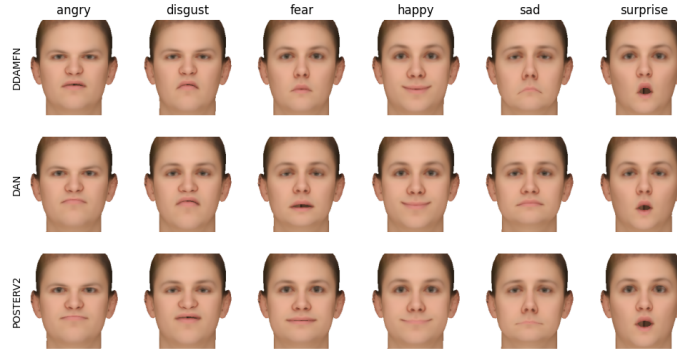


Fig. 14: The average facial expressions for models trained on RAFDB [46, 47]

D.4 Model Activation Surfaces

The main paper shows that we use a finite grid over $\mathcal{T}$ and $\mathcal{S}$ to compute the FAIS score. Given that we only highlighted the final time step $t = 1.0$, we show here the full logit activation surfaces used to compute our score.

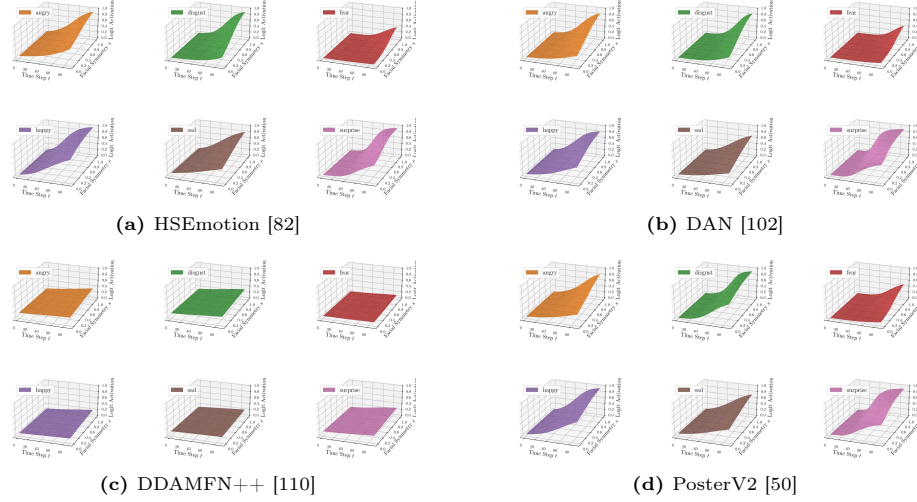


Fig. 15: AffectNet7 [53]

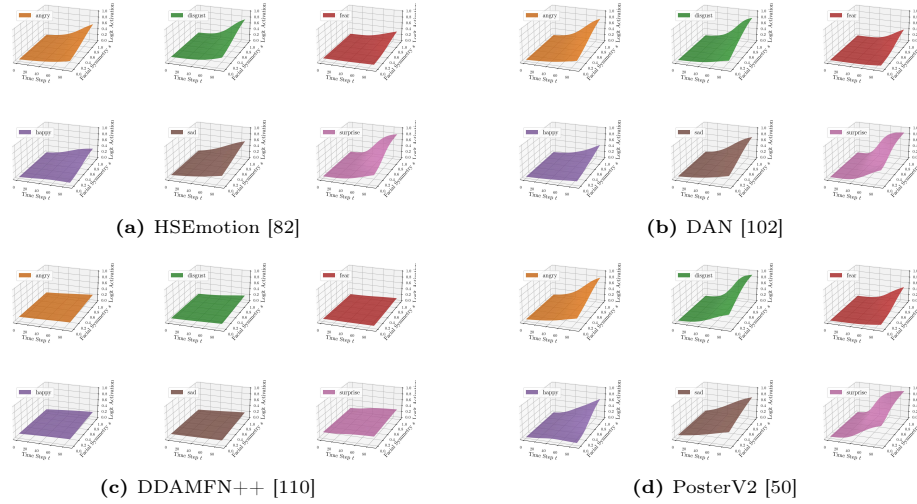


Fig. 16: AffectNet8 [53]

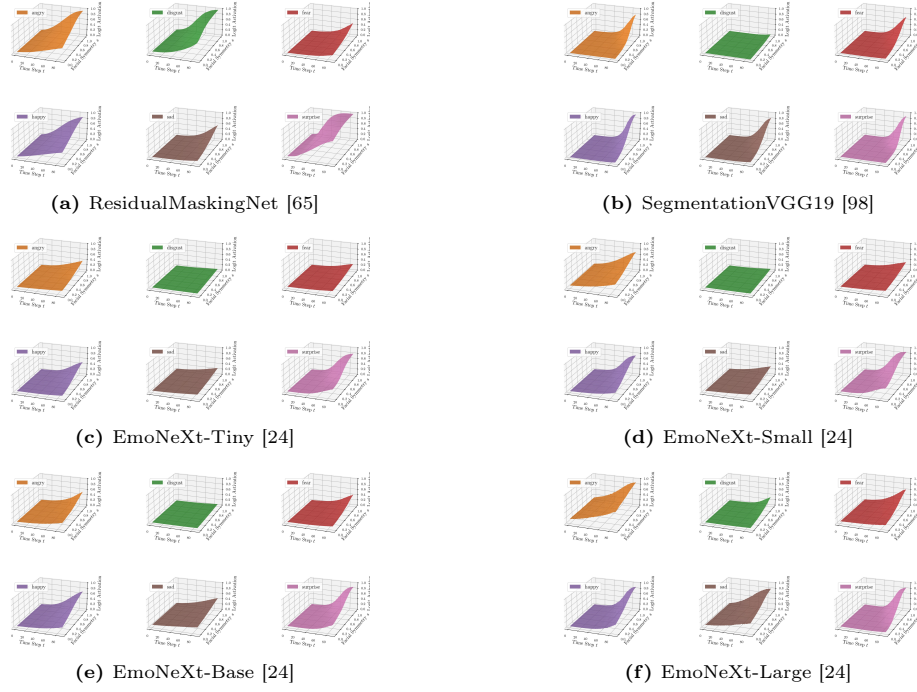


Fig. 17: FER2013 [20]

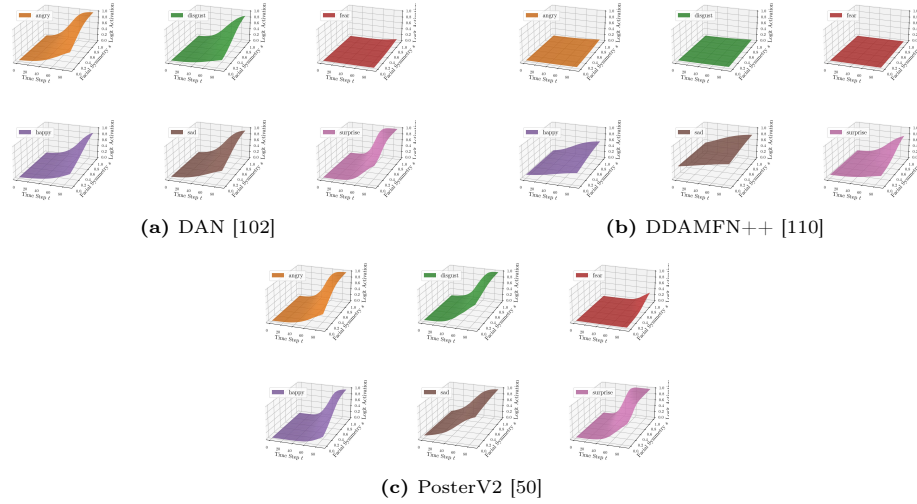


Fig. 18: RAFDB [46, 47]

D.5 Model Symmetry Impact

The main paper shows that we compute the interpretable asymmetry score using a finite grid over $\mathfrak{T}$ and $\mathfrak{S}$. Given that we only highlighted the final time step $t = 1.0$, we show here the full logit activation surfaces used to compute our score.

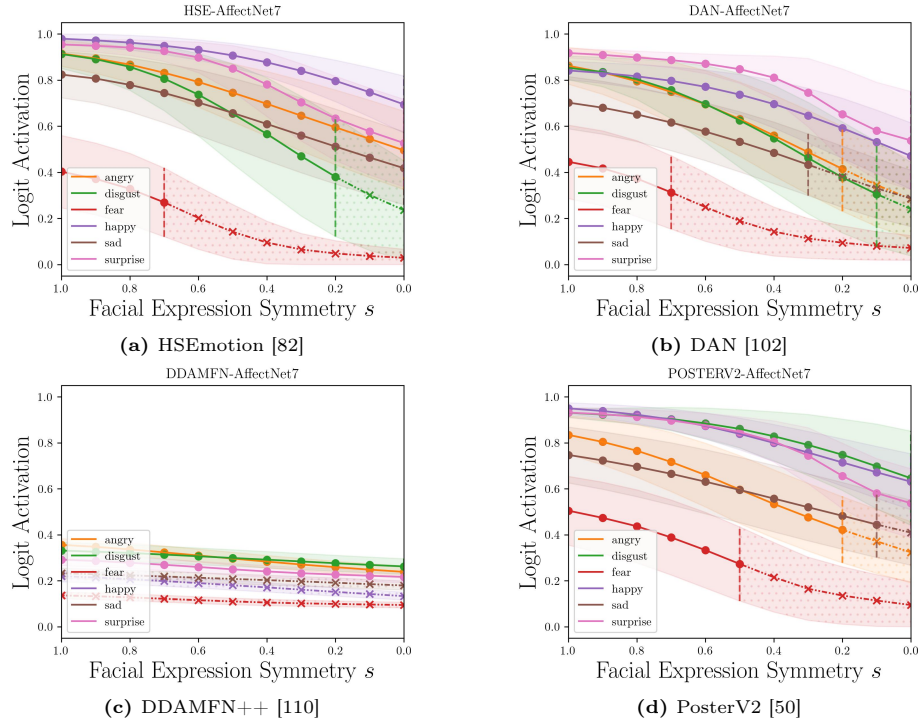
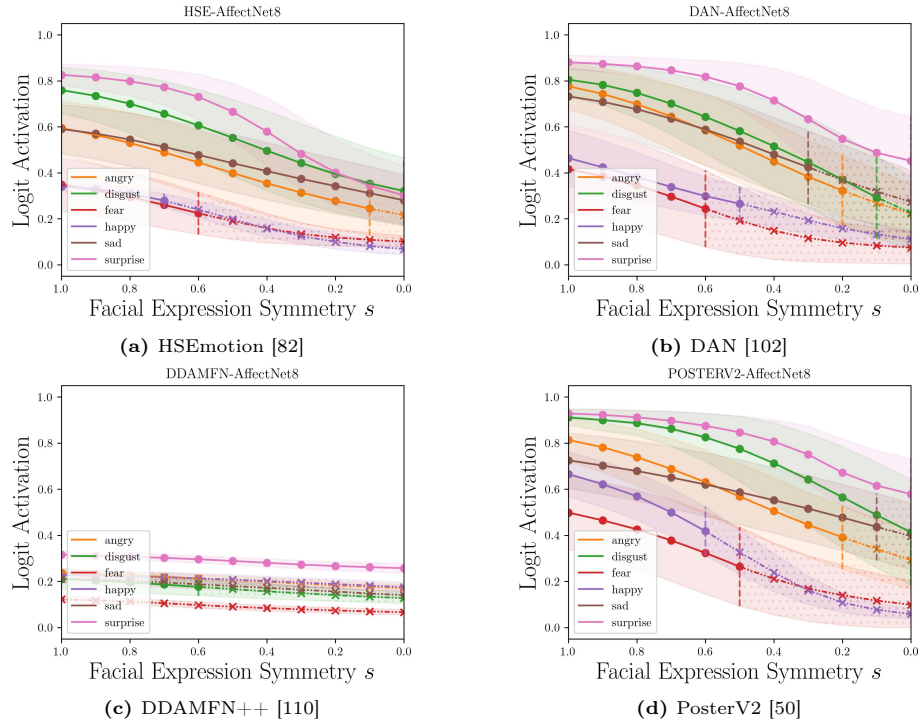


Fig. 19: AffectNet7 [53]

**Fig. 20:** AffectNet8 [53]

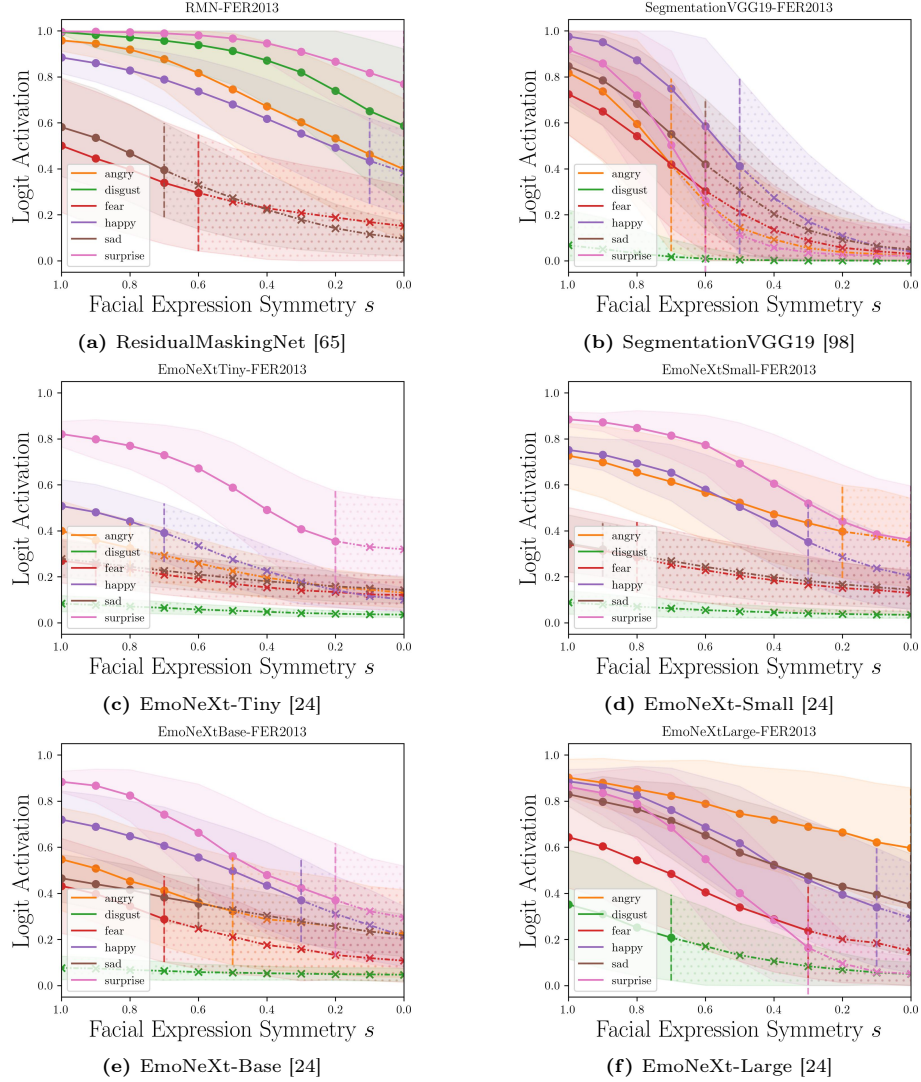
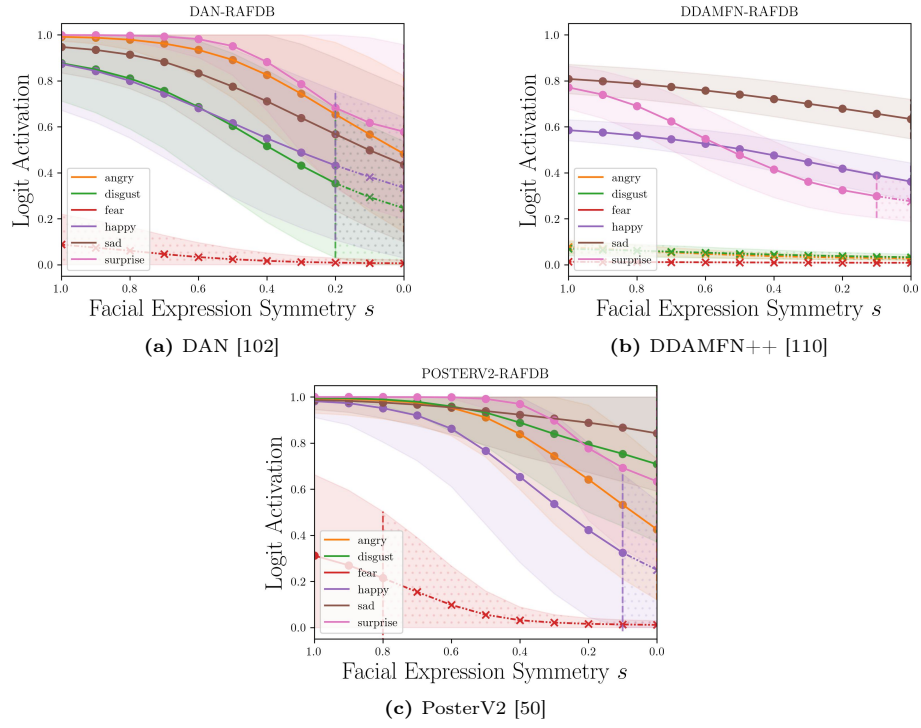


Fig. 21: FER2013 [20]

**Fig. 22:** RAFDB [46, 47]

D.6 Local Explanations - Saliency Maps

We aim to understand the impact of facial asymmetry globally; local explanations via saliency maps still offer insights but require human interpretation. We use an occlusion-based interpretation approach [107] for the ground truth and predicted label using the average emotion simulated with the default identity.

Local Explanations - AffectNet7 The saliency maps indicate that independent of the predicted or the ground truth label, the majority impact is only on one side of the face. This supports our global observation that facial symmetry has a strong impact on the model behavior.

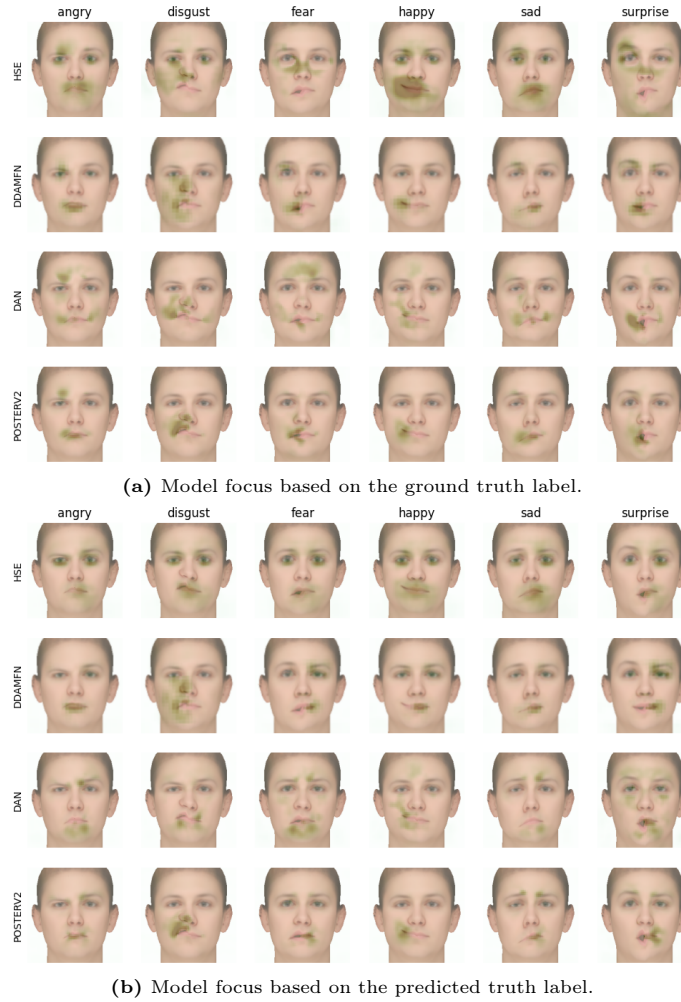


Fig. 23: The occlusion-based saliency maps for models trained on AffectNet7 [53]

Local Explanations - AffectNet8 The saliency maps indicate that independent of the predicted or the ground truth label, the majority impact is only on one side of the face. This supports our global observation that facial symmetry has a strong impact on the model behavior.

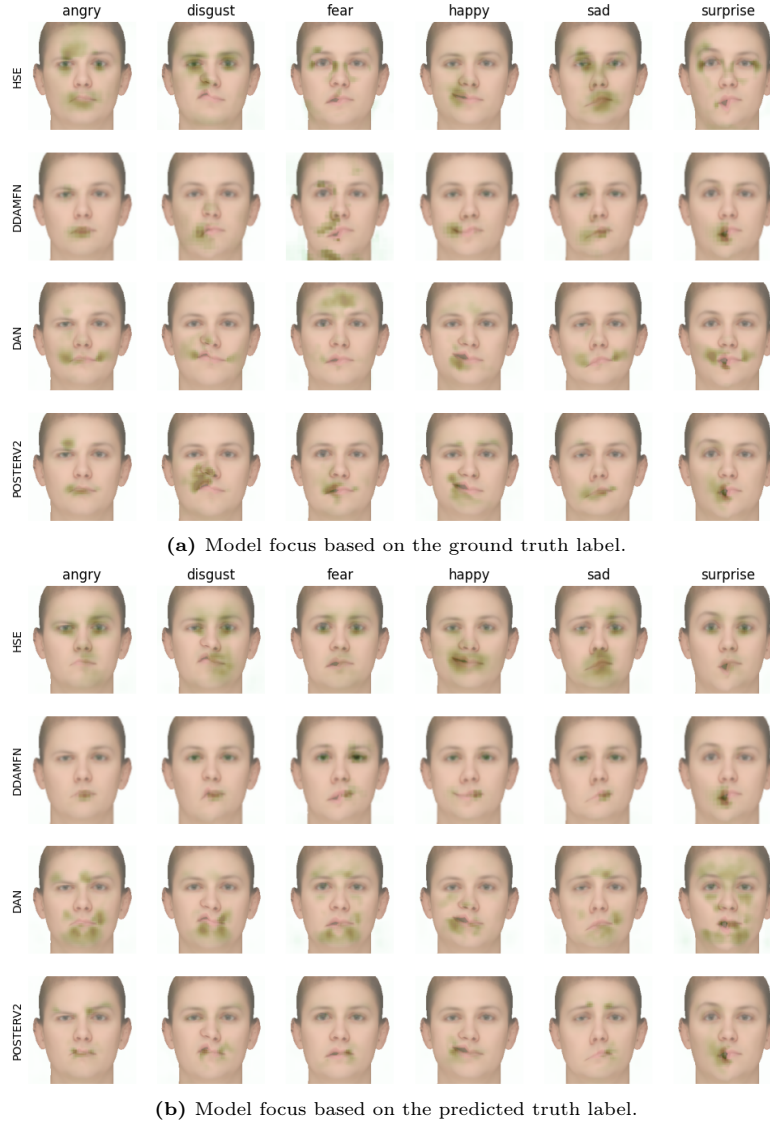


Fig. 24: The occlusion-based saliency maps for models trained on AffectNet8 [53]

Local Explanations - FER2013 The saliency maps indicate that independent of the predicted or the ground truth label, the majority impact is only on one side of the face. This supports our global observation that facial symmetry has a strong impact on the model behavior.

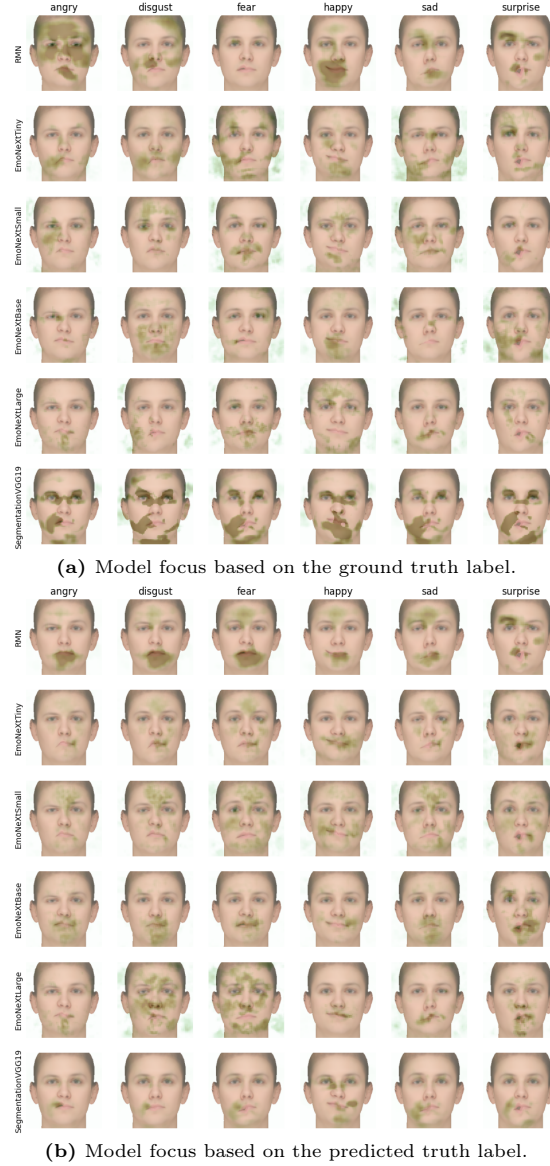


Fig. 25: The occlusion-based saliency maps for models trained on FER2013 [20]

Local Explanations - RAFDB The saliency maps indicate that independent of the predicted or the ground truth label, the majority impact is only on one side of the face. This supports our global observation that facial symmetry has a strong impact on the model behavior.

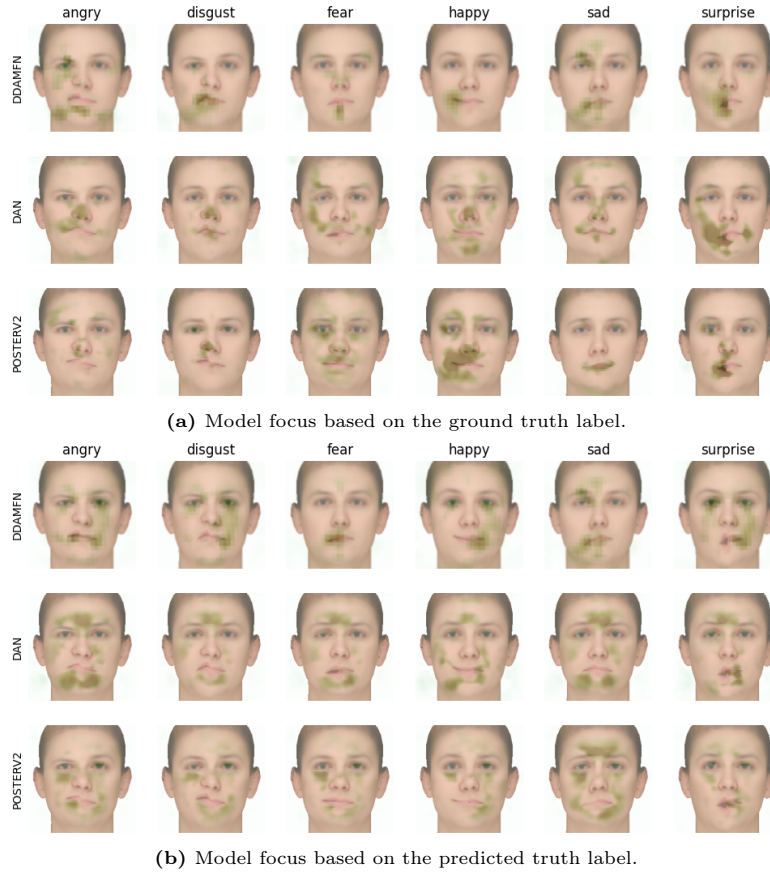


Fig. 26: The occlusion-based saliency maps for models trained on RAFDB [46, 47]