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

Tim Büchner*[1] [0000-0002-6879-552X], Niklas Penzel*[1] [0000-0001-8002-4130], Orlando Guntinas-Lichius[2] [0000-0001-9671-0784], and Joachim Denzler[1] [0000-0002-3193-3300]

[1] Computer Vision Group, Friedrich Schiller University Jena, 07743 Jena, Germany
[2] Dept. of Otorhinolaryngology, Jena University Hospital, 07747 Jena, Germany
tim.buechner@uni-jena.de

1 Structural Causal Models

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

Definition 1 (Structural Causal Model in [15, Sec. 7.1.1] and [1, 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, $V = \{V_1, ..., V_n\}$ is the set of endogenous variables we measure in our model, $F = \{f_2, ..., f_n\}$ 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 $F_x = \{f_j : V_j \neq V_i\} \cup \{V_i \leftarrow x\}$ [1].

2 Measuring Systematic Change - Significance Test

In Algorithm 1, we provide detailed pseudo code for our proposed shuffle hypothesis test regarding the significance of $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)} | I_{\varphi^{(e)}})$. Further in Fig. 1, we

[*] These authors contributed equally to this work.

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. 1, is not typical for the estimated null-distribution leading to a low p-value. Table 1 contains the number of individuals per classifier and expression for which Algorithm 1 together with the Holm-Bonferroni correction [8] 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.

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

```
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
  $p \leftarrow 0.0$
  $\sigma_{orig.} \leftarrow \hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)}|I_{\varphi^{(e)}})$                       ▷ Estimate the original statistic
  for $i \in \{1, ..., K\}$ do
    $\mathbb{F}_\theta^{(e)}(I_{\varphi^{(e)}}^{(perm.)}(s,t)) \leftarrow \texttt{permute}(\mathbb{F}_\theta^{(e)}(I_{\varphi^{(e)}}(s,t)), \texttt{axis} = 0)$
                                                                      ▷ Shuffle along Symmetry Axis
    $\sigma_{perm.} \leftarrow \hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)}|I_{\varphi^{(e)}}^{(perm.)})$
    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 $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)}|I_{\varphi^{(e)}})$ is significant.
  else
    return $\hat{\mathcal{S}}(\mathbb{F}_\theta^{(e)}|I_{\varphi^{(e)}})$ is not significant.
  end if
```

3 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.

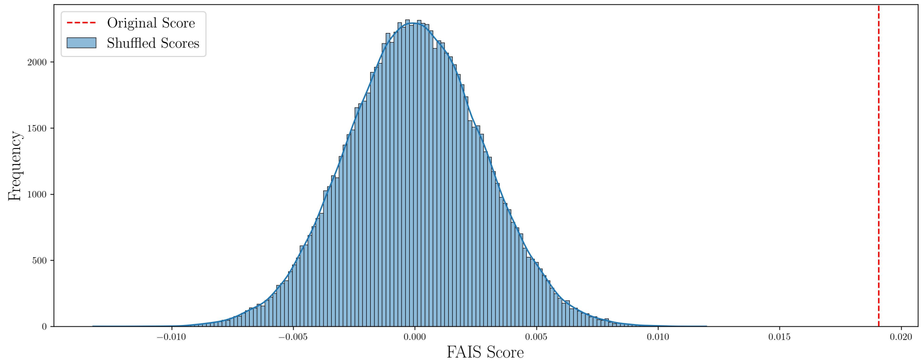


Fig. 1: Using the shuffle test, outlined in Algorithm 1, we plot the resulting scores for 100000 permutations in a histogram. Our tests use a significance threshold of $p < (0.05)$. The original model score for a single individual, shown as a red dashed line, lies clearly outside the computed null distribution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

3.1 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 2, 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 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.

Table 1: We report how many of the 200 individuals, using the Holm-Bonferroni [8] 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 [26]	200	200	200	200	200	200
	DDAMFN++ [29]	200	200	200	200	200	200
	HSEmotion [23]	200	200	200	200	200	200
	PosterV2 [12]	200	199	200	200	200	200
AffectNet8	DAN [26]	200	200	200	200	200	200
	DDAMFN++ [29]	200	200	198	200	200	200
	HSEmotion [23]	200	200	200	200	200	200
	PosterV2 [12]	200	200	200	200	200	200
FER2013	EmoNeXt-Tiny† [5]	200	200	200	200	200	200
	EmoNeXt-Small† [5]	199	194	200	200	200	200
	EmoNeXt-Base† [5]	163	137	200	200	200	200
	EmoNeXt-Large† [5]	133	187	196	195	198	187
	ResidualMaskingNet [18]	200	194	199	200	200	200
	Segmentation-VGG19† [25]	199	148	200	148	200	129
RAFDB	DAN [26]	200	199	200	200	200	200
	DDAMFN++ [29]	200	200	200	200	200	200
	PosterV2 [12]	200	194	200	200	195	200

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.

3.2 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 [21] toward this goal. This method frames supervised learning as an SCM [15] 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 [19].

Clearly, the choice of conditional independence test is an important hyperparameter choice to ensure that the results are reliable. Further, Shah and Peters [24] prove that there is no optimal test that can control type-I errors, i.e.,

Table 2: 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 darks 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 [26]	57.50%	99.00%	100.00%	94.44%	62.82%
	DDAMFN++ [29]	0.00%	0.00%	19.50%	0.35%	0.20%
	HSEmotion [23]	96.00%	100.00%	100.00%	0.00%	0.00%
	PosterV2 [12]	87.00%	100.00%	100.00%	97.92%	61.23%
	Average	80.17%	99.67%	79.88%	64.24%	41.42%
AffectNet8	DAN [26]	0.00%	30.50%	88.50%	91.32%	45.92%
	DDAMFN++ [29]	0.00%	8.00%	45.00%	0.35%	6.56%
	HSEmotion [23]	0.00%	0.00%	97.50%	0.00%	0.00%
	PosterV2 [12]	0.00%	20.50%	99.00%	89.93%	36.58%
	Average	0.00%	19.67%	82.50%	60.53%	29.69%
FER2013	EmoNeXt-Base† [5]	10.50%	71.00%	97.50%	99.65%	70.38%
	EmoNeXt-large† [5]	28.50%	77.50%	100.00%	98.26%	63.42%
	EmoNeXt-small† [5]	10.50%	68.50%	100.00%	98.96%	66.40%
	EmoNeXt-tiny† [5]	0.00%	12.50%	87.00%	97.92%	58.45%
	ResidualMaskingNet [18]	33.50%	89.00%	100.00%	92.36%	59.05%
	Segmentation-VGG19† [25]	2.00%	39.50%	99.00%	97.22%	31.81%
	Average	17.00%	59.67%	97.25%	97.40%	58.25%
RAFDB	DAN [26]	30.00%	65.50%	94.50%	32.29%	54.27%
	DDAMFN++ [29]	69.00%	99.00%	100.00%	90.97%	76.34%
	PosterV2 [12]	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%

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 [2, 16, 17, 20] and select multiple non-linear tests. Specifically, we select conditional HSIC [6], CMIknn [22], and FCIT [3]. We consider the result from all three tests and report the majority decision [20].

The selected conditional independence tests themselves have different hyperparameter choices. First, for conditional HSIC [6], 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. [7] to approximate suitable kernel widths for all of our three variables. Second, similarly for CMIknn [22], 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 [3], 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.

3.3 Additional Visualizations Regarding Logit Activations

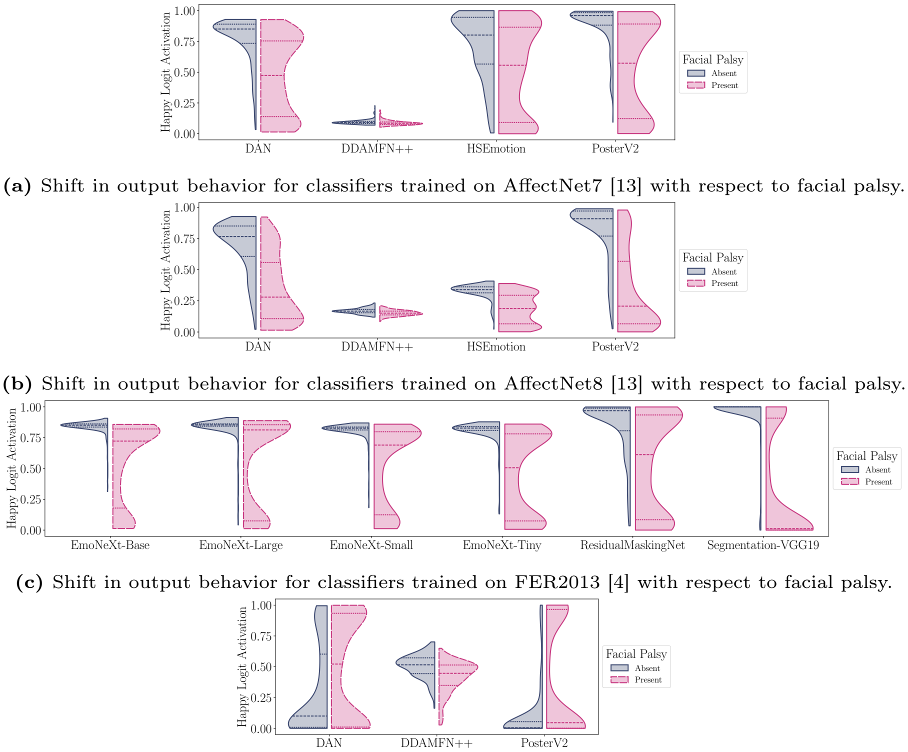
Following previous work [2], we visualize the difference in the *happy* logit behavior between the healthy probands and facial palsy patients. Fig. 2 contains these results split between the training datasets of the 17 models we investigate in this work. However, these are associational investigations, i.e., of the first level of the PCH [1]. 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. 2 shows a decrease in *happy* activations for most models. This is congruent with the aggregated performance results in Table 2. 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 [26], and PosterV2 [12] show higher activations and improved performance. Nonetheless, both models still struggle with facial palsy patients and are outperformed by DDAMMFN++ [29] trained on the same dataset.

Additionally, we also visualize the results for the continuous LPIPS [28] symmetry. For regressing the mean and standard deviation, we use a window regression approach as described in [16]. We display these visualizations in Fig. 3.

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. 2. Furthermore, we again observe a very small effect size for DDAMFS++ [29] for both features. These findings are in agreement with the noted performance in Table 2.

Nevertheless, we want to highlight two additional observations: First, in Fig. 3d, 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. 2d. Second, while for most models in Fig. 3, we observe a decrease in logit activations for lower facial symmetry, we note a smaller increase again for the most asymmetric faces.



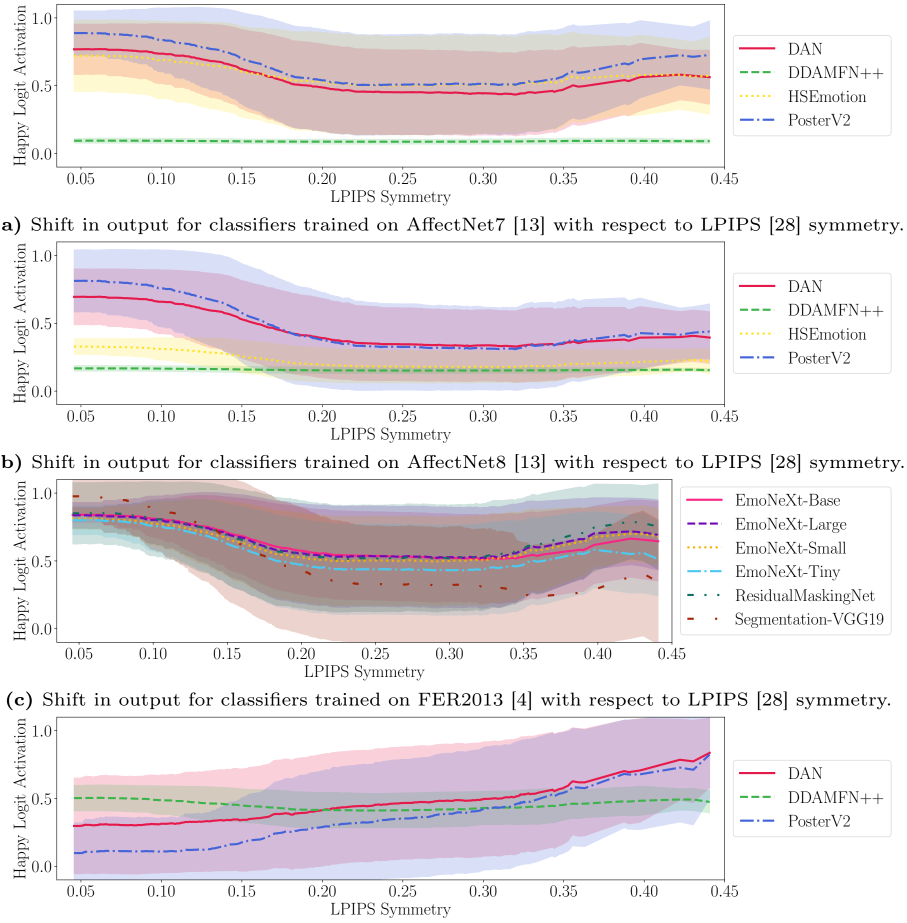
(a) Shift in output behavior for classifiers trained on AffectNet7 [13] with respect to facial palsy.

(b) Shift in output behavior for classifiers trained on AffectNet8 [13] with respect to facial palsy.

(c) Shift in output behavior for classifiers trained on FER2013 [4] with respect to facial palsy.

(d) Shift in output behavior for classifiers trained on RAFDB [9, 10] with respect to facial palsy. Note that we find the behavior shift for the DAN [26] model is not significant.

Fig. 2: We follow [2] and visualize the differences in the classifiers *happy* logit distribution for healthy probands and facial palsy patients. Here 2a - 2d contain models trained on the indicated dataset respectively.



(a) Shift in output for classifiers trained on AffectNet7 [13] with respect to LPIPS [28] symmetry.

(b) Shift in output for classifiers trained on AffectNet8 [13] with respect to LPIPS [28] symmetry.

(c) Shift in output for classifiers trained on FER2013 [4] with respect to LPIPS [28] symmetry.

(d) Shift in output for classifiers trained on RAFDB [9, 10] with respect to LPIPS [28] symmetry.

Fig. 3: We follow [2, 16] and regress the shift in the classifiers *happy* logit distribution for measured LPIPS [28] symmetry scores of healthy probands and facial palsy patients. Here 3a - 3d contain models trained on the indicated dataset respectively. Note that higher LPIPS corresponds to lower symmetry [28].

Table 3: 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 [26]	0.862	0.853	0.446	0.842	0.702	0.917
	DDAMFN++ [29]	0.356	0.331	0.136	0.220	0.233	0.292
	HSEmotion [23]	0.915	0.913	0.403	0.979	0.824	0.954
	PosterV2 [12]	0.835	0.931	0.505	0.950	0.747	0.931
AffectNet8	DAN [26]	0.776	0.805	0.416	0.464	0.732	0.881
	DDAMFN++ [29]	0.237	0.211	0.122	0.228	0.212	0.316
	HSEmotion [23]	0.595	0.759	0.349	0.340	0.590	0.826
	PosterV2 [12]	0.814	0.911	0.499	0.666	0.725	0.928
FER2013	EmoNeXt-Tiny† [5]	0.400	0.083	0.269	0.508	0.278	0.821
	EmoNeXt-Small† [5]	0.727	0.088	0.342	0.752	0.345	0.885
	EmoNeXt-Base† [5]	0.548	0.076	0.432	0.720	0.465	0.884
	EmoNeXt-Large† [5]	0.902	0.352	0.644	0.886	0.829	0.862
	ResidualMaskingNet [18]	0.959	0.995	0.500	0.884	0.581	0.997
	Segmentation-VGG19† [25]	0.818	0.067	0.725	0.976	0.846	0.919
RAFDB	DAN [26]	0.991	0.877	0.088	0.874	0.947	0.999
	DDAMFN++ [29]	0.077	0.070	0.013	0.585	0.808	0.771
	PosterV2 [12]	0.993	0.996	0.312	0.982	0.987	1.000

4 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.

4.1 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 3. 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 [25]. The *surprise* facial expression has a high activation among all classifiers, whereas DDAMFN++ [29] is the sole outlier; overall, reached activation is low.

4.2 Individuals

We provide an overview of all created individuals in Fig. 4. 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 [11] and the texture from the BaselFace-Model [14].

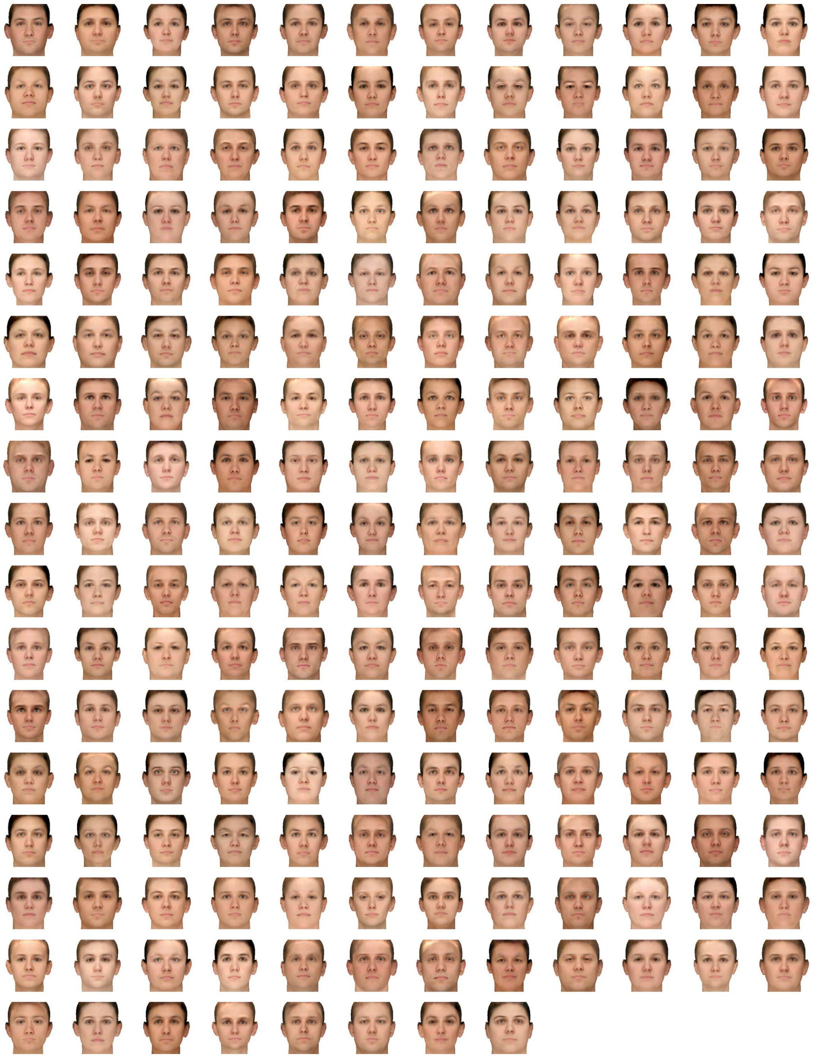


Fig. 4: 200 individual population $\mathfrak{I}$

4.3 Average Facial Expression

Together with the reached logit activations, see Table 3, 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. 5. This visualization gives an intuitive feeling about the underlying facial expression per FER benchmark [4,9,10,13]. 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.

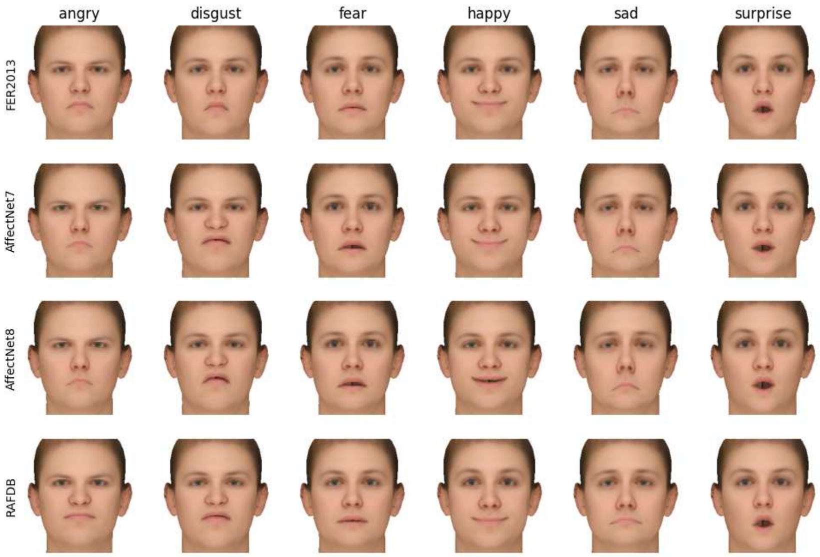


Fig. 5: Average Facial Expression used for classification based on the underlying training dataset.

Average Facial Expression - AffectNet7 Fig. 6 contains the average facial expressions for models trained on AffectNet7 [13].

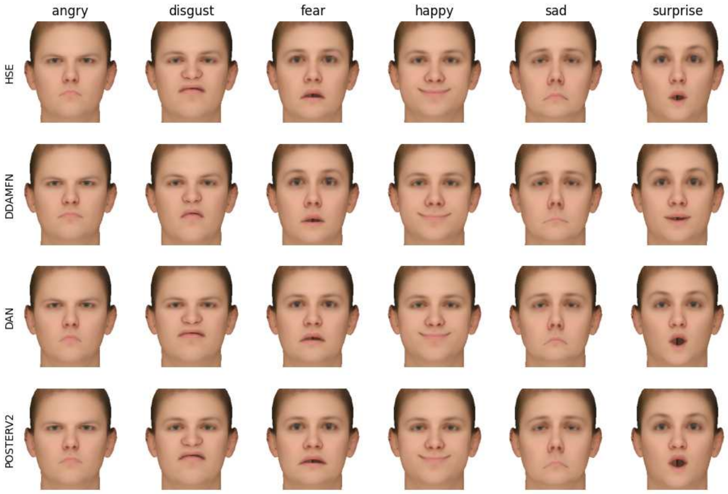


Fig. 6: The average facial expressions for models trained on AffectNet7 [13]

Average Facial Expression - AffectNet8 Fig. 7 contains the average facial expressions for models trained on AffectNet8 [13].

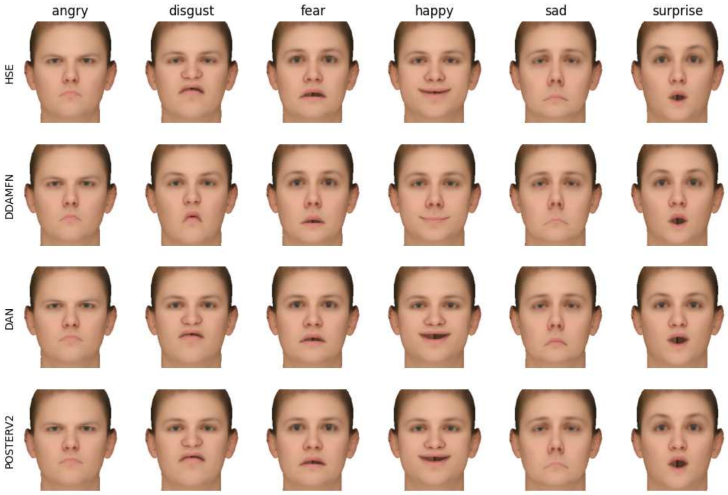


Fig. 7: The average facial expressions for models trained on AffectNet8 [13]

Average Facial Expression - FER2013 Fig. 8 contains the average facial expressions for models trained on FER2013 [4].

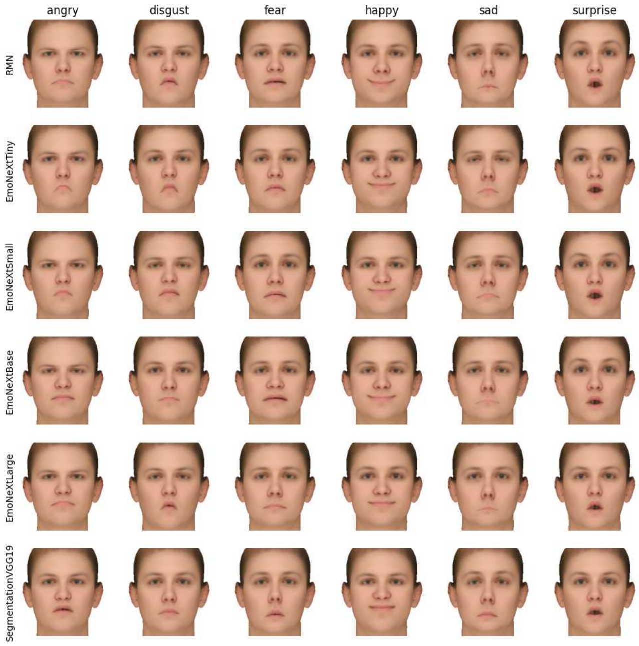


Fig. 8: The average facial expressions for models trained on FER2013 [4]

Average Facial Expression - RAFDB Fig. 9 contains the average facial expressions for models trained on RAFDB [9, 10].

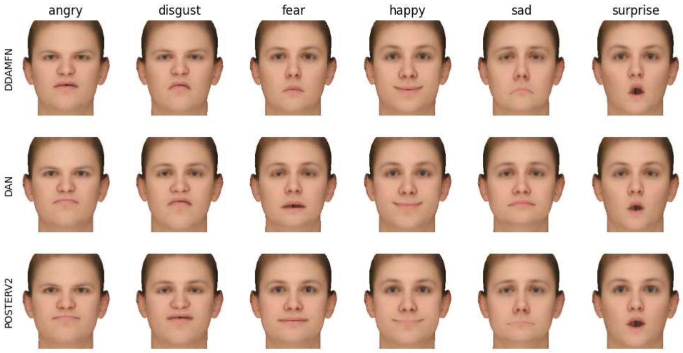
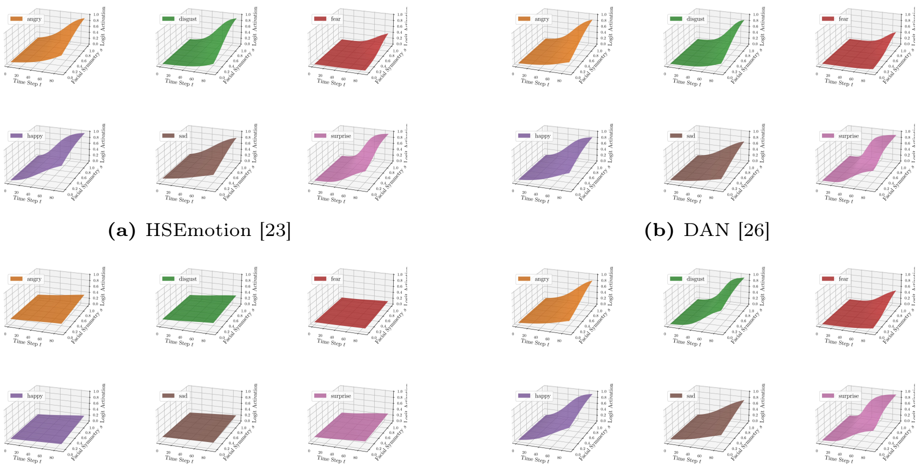


Fig. 9: The average facial expressions for models trained on RAFDB [9, 10]

4.4 Model Activation Surfaces

The main paper shows that we use a finite grid over $\mathfrak{T}$ and $\mathfrak{S}$ 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.



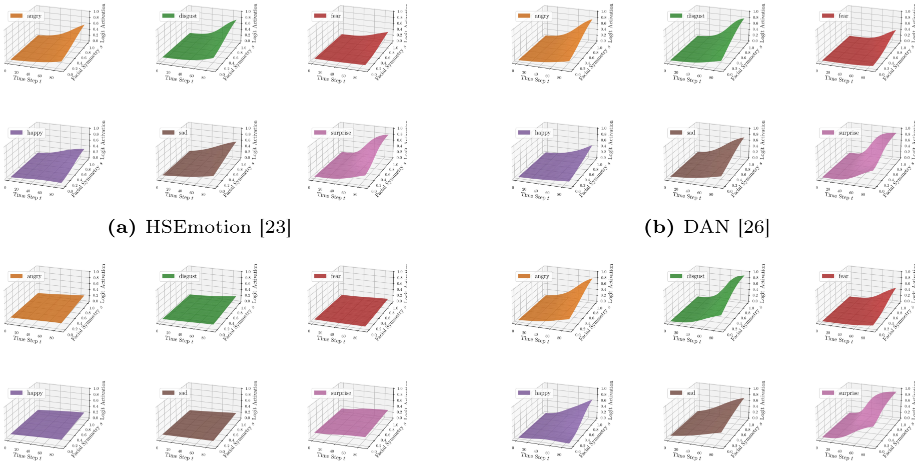
(a) HSEmotion [23]

(b) DAN [26]

(c) DDAMFN++ [29]

(d) PosterV2 [12]

Fig. 10: AffectNet7 [13]



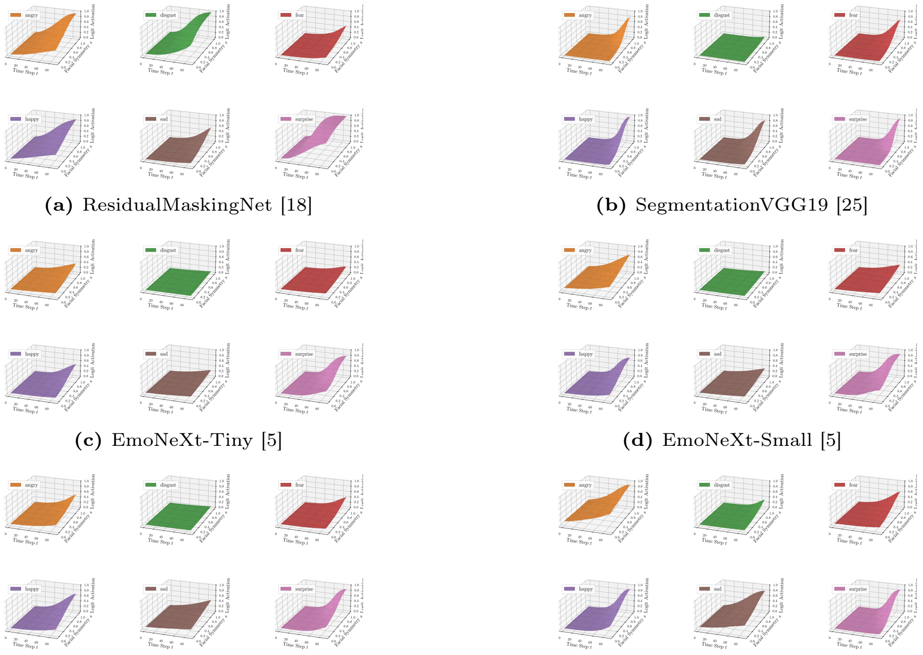
(a) HSEmotion [23]

(b) DAN [26]

(c) DDAMFN++ [29]

(d) PosterV2 [12]

Fig. 11: AffectNet8 [13]



(a) ResidualMaskingNet [18]

(b) SegmentationVGG19 [25]

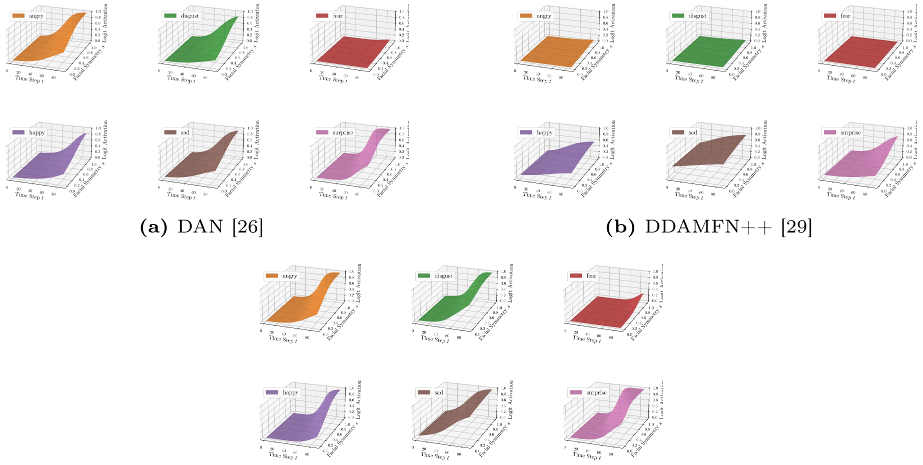
(c) EmoNeXt-Tiny [5]

(d) EmoNeXt-Small [5]

(e) EmoNeXt-Base [5]

(f) EmoNeXt-Large [5]

Fig. 12: FER2013 [4]



(a) DAN [26]

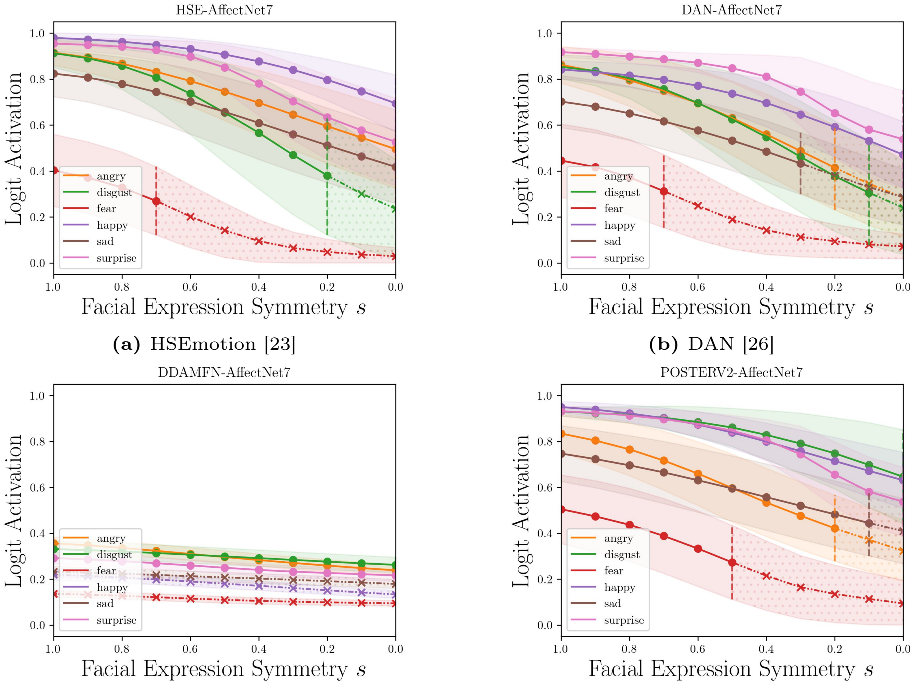
(b) DDAMFN++ [29]

(c) PosterV2 [12]

Fig. 13: RAFDB [9, 10]

4.5 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.



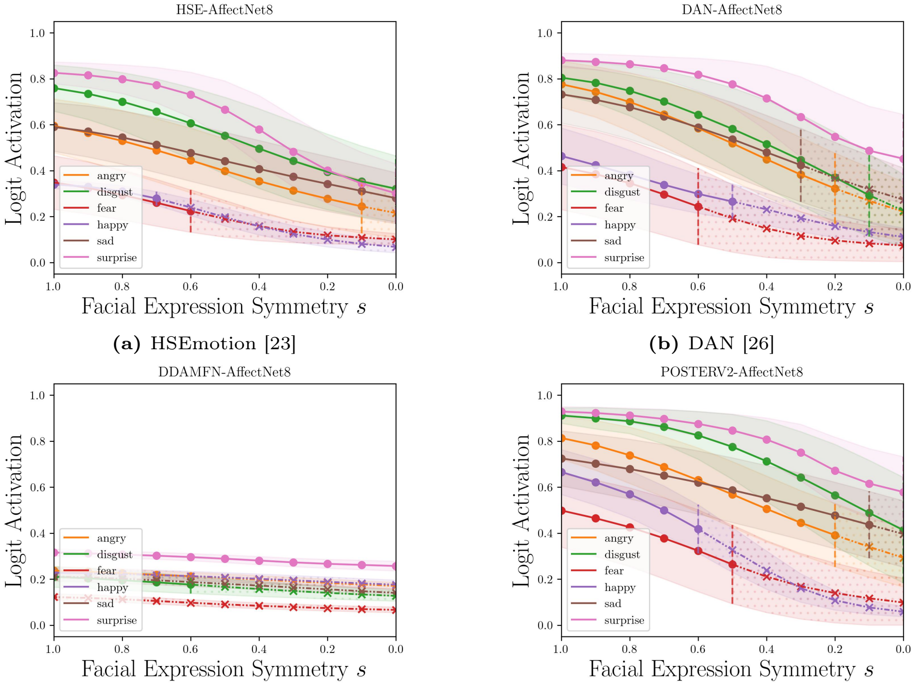
(a) HSEmotion [23]

(b) DAN [26]

(c) DDAMFN++ [29]

(d) PosterV2 [12]

Fig. 14: AffectNet7 [13]



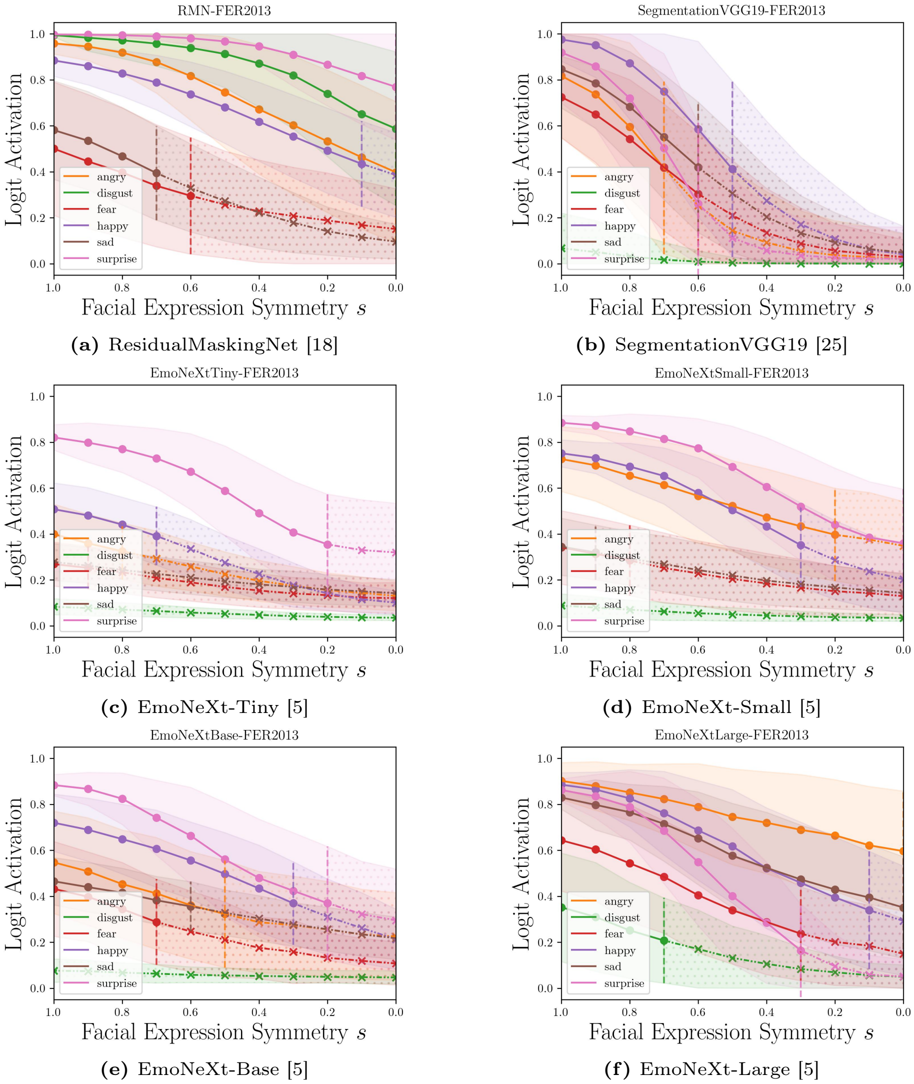
(a) HSEmotion [23]

(b) DAN [26]

(c) DDAMFN++ [29]

(d) PosterV2 [12]

Fig. 15: AffectNet8 [13]



(a) ResidualMaskingNet [18]

(b) SegmentationVGG19 [25]

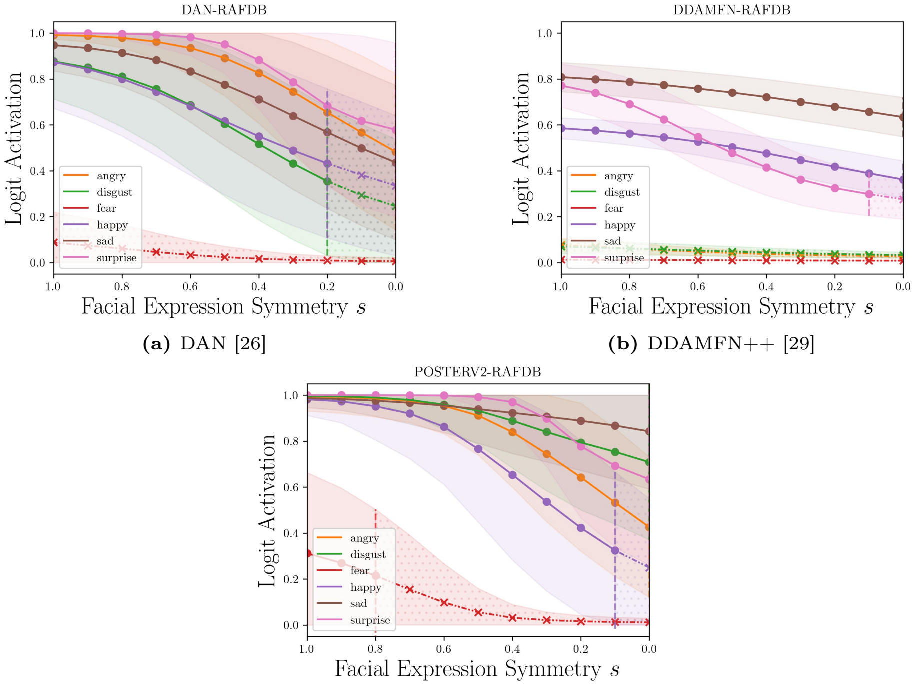
(c) EmoNeXt-Tiny [5]

(d) EmoNeXt-Small [5]

(e) EmoNeXt-Base [5]

(f) EmoNeXt-Large [5]

Fig. 16: FER2013 [4]



(a) DAN [26]

(b) DDAMFN++ [29]

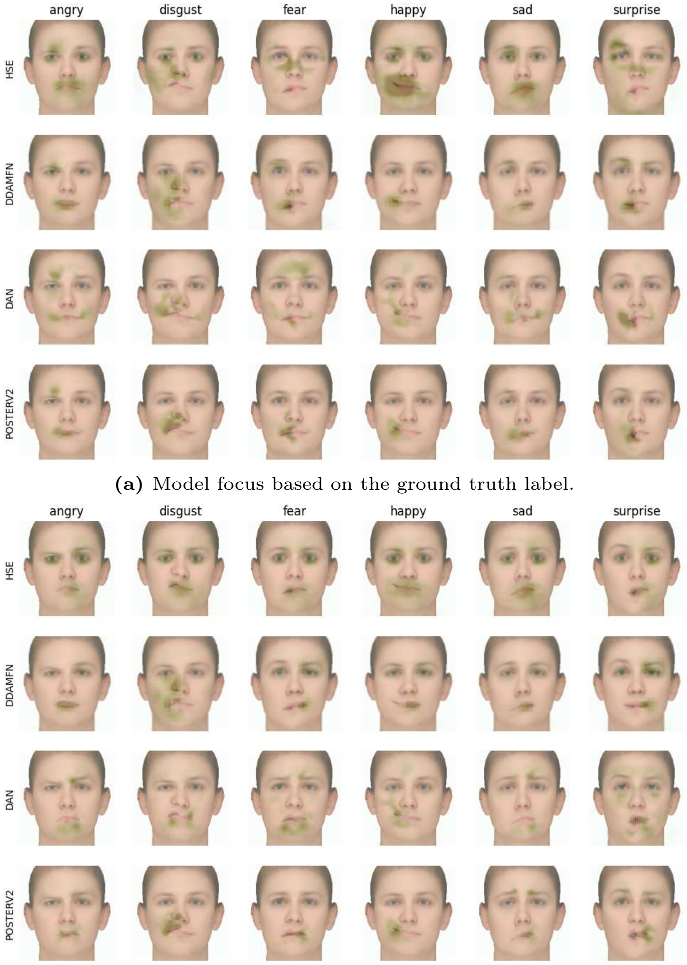
(c) PosterV2 [12]

Fig. 17: RAFDB [9, 10]

4.6 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 [27] 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.

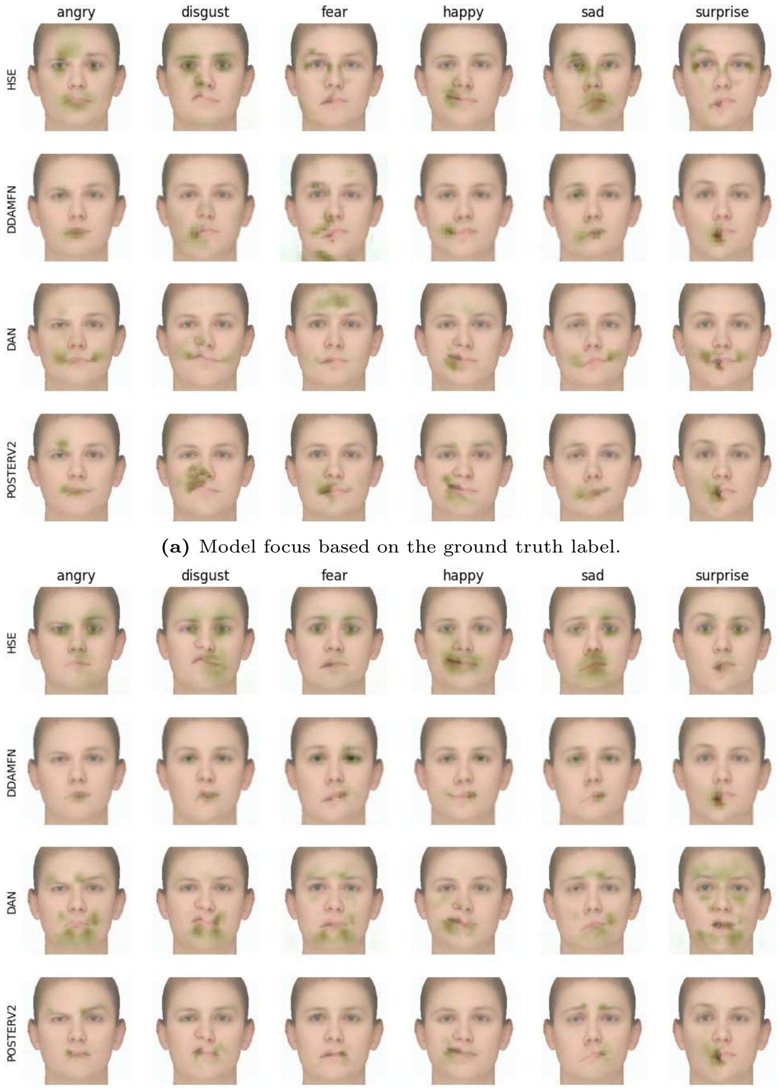


(a) Model focus based on the ground truth label.

(b) Model focus based on the predicted truth label.

Fig. 18: The occlusion-based saliency maps for models trained on AffectNet7 [13]

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.

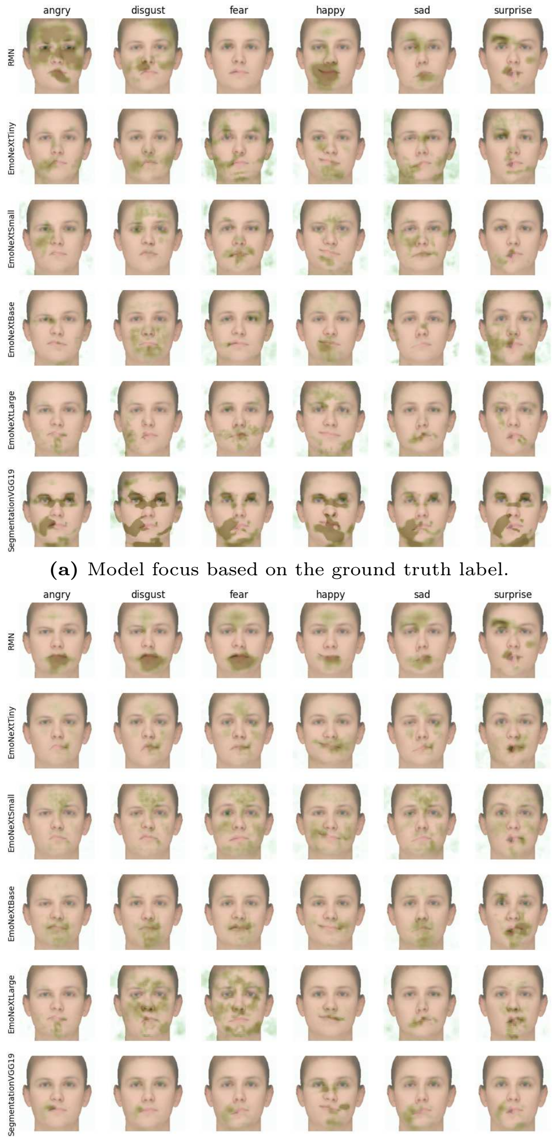


(a) Model focus based on the ground truth label.

(b) Model focus based on the predicted truth label.

Fig. 19: TThe occlusion-based saliency maps for models trained on AffectNet8 [13]

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.

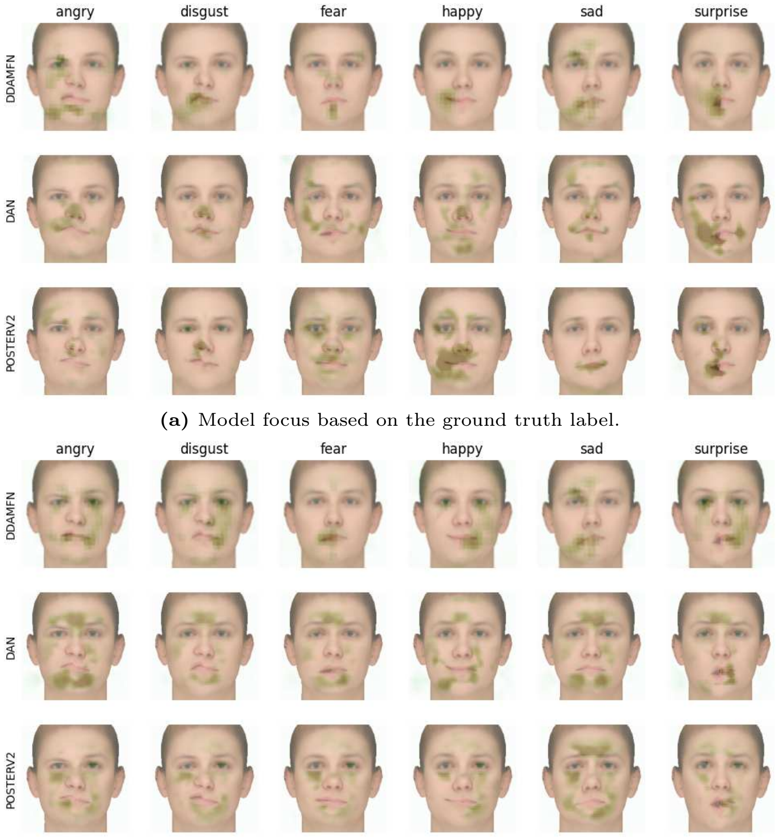


(a) Model focus based on the ground truth label.

(b) Model focus based on the predicted truth label.

Fig. 20: The occlusion-based saliency maps for models trained on FER2013 [4]

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.



(a) Model focus based on the ground truth label.

(b) Model focus based on the predicted truth label.

Fig. 21: The occlusion-based saliency maps for models trained on RAFDB [9, 10]

References

1. 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)
2. Büchner, 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)
3. Chalupka, K., Perona, P., Eberhardt, F.: Fast conditional independence test for vector variables with large sample sizes. arXiv preprint arXiv:1804.02747 (2018)

4. 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`
5. 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`
6. Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. Advances in neural information processing systems **20** (2007)
7. 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)
8. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979), `http://www.jstor.org/stable/4615733`
9. 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)
10. 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)
11. 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`
12. Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A.: POSTER++: A simpler and stronger facial expression recognition network (Feb 2023)
13. 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`
14. 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`
15. Pearl, J.: Causality. Cambridge university press (2009)
16. 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)
17. 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)
18. 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`
19. Reichenbach, H.: The direction of time, vol. 65. Univ of California Press (1956)
20. 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)

21. Reimers, C., Runge, J., Denzler, J.: Determining the relevance of features for deep neural networks. In: European Conference on Computer Vision. Springer (2020)
22. 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)
23. 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`
24. 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)
25. 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`
26. 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`
27. 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)
28. 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`
29. 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`