Canonical Saliency Maps: Decoding Deep Face Models

Thrupthi Ann John, Vineeth N. Balasubramanian, *Senior Member, IEEE*, and C. V. Jawahar, *Member, IEEE*

***Abstract*—As Deep Neural Network models for face processing tasks approach human-like performance, their deployment in critical applications such as law enforcement and access control has seen an upswing, where any failure may have far-reaching consequences. We need methods to build trust in deployed systems by making their working as transparent as possible. Existing visualization algorithms are designed for object recognition and do not give insightful results when applied‘ to the face domain. In this work, we present ‘Canonical Saliency Maps’, a new method which highlights relevant facial areas by projecting saliency maps onto a canonical face model. We present two kinds of Canonical Saliency Maps: image-level maps and model-level maps. Image-level maps highlight facial features responsible for the decision made by a deep face model on a given image, thus helping to understand how a DNN made a prediction on the image. Model-level maps provide an understanding of what the entire DNN model focuses on in each task, and thus can be used to detect biases in the model. Our qualitative and quantitative results show the usefulness of the proposed canonical saliency maps, which can be used on any deep face model regardless of the architecture.**

***Index Terms*—Deep neural networks, face understanding, explainability/accountability/transparency, canonical model.**

I. Introduction

DEEP learning achieves state-of-the-art performance in most computer vision tasks, surpassing earlier methods by a large margin. The performance of deep neural networks is improving in leaps and bounds for face processing tasks such as face recognition and detection. In 2014, DeepFace [1] approached human-like performance for the first time on the LFW benchmark [2], a dataset of face images in unconstrained settings (DeepFace: 97.35% vs. Human: 97.53%), using a training dataset of 4 million images. In recent years, the accuracy has increased up to 99.8% [3], thereby surpassing human performance on the benchmark. Deep face models are now deemed to be real-world ready. They are used in many critical areas by government agencies including law enforcement and access control. Currently, models for face tasks are available from major companies like Microsoft, IBM and Amazon who claim that their models are highly accurate. In this scenario, two crucial questions arise: Do pre-trained models perform as well as they claim, and how do we find the weaknesses existing in these models and improve them.

Failures of face models in critical areas have far-reaching and devastating consequences. Inaccuracies in facial recognition technology can result in an innocent person being misidentified as a criminal and subjected to unwarranted police scrutiny. Big Brother Watch U.K. released the Face-Off report [4] highlighting false positive match rates of over 90% for the facial recognition technology deployed by the Metropolitan police. A recent study [5] demonstrated that although commercial software solutions report high accuracies (Amazon’s Rekognition reports an accuracy of 97%), they demonstrate skin-type and gender biases that go unreported as the benchmarks themselves are skewed. When performance is reported on public or private databases, they are always subject to the biases inherent in these databases. The algorithms may be then used in the real world in conditions that differ wildly from the ones that they are tested in, causing the algorithms to produce erroneous results. How do we catch such issues at an early stage? High reported accuracy is not enough to determine how an algorithm will perform under real-life conditions. We need to be able to peek inside the algorithms and understand how they work. The opaqueness of deep models restricts its usefulness in highly regulated environments (e.g., healthcare, autonomous driving), which may require the reasoning of the decisions taken by the deep models to be provided. To build trust in deployed intelligent systems, they need to be transparent, i.e., they should be able to explain why they predict what they predict [6]. Interpretable algorithms allow us to responsibly deploy deep face models in the real world, as they help end users be aware of these models’ characteristics and shortcomings.

Several visualization methods have been proposed to increase the interpretability and transparency of deep neural networks. So far, most neural network visualization methods have been created with the task of object recognition in mind. There have been very few works that applied these algorithms exclusively to the face domain [7], [8]. The saliency methods of object recognition do not readily translate to the face domain, as the images used for face tasks have different properties from those used for generic object recognition. Face images are highly structured forms of input. The

Manuscript received May 2, 2021; revised August 8, 2021; accepted October 9, 2021. This work was supported in part by Ministry of Education and Department of Science and Technology, Government of India, and in part by Honeywell through the UAY Program under Grant UAY/IITH005. This article was recommended for publication by Associate Editor Y. Wang upon evaluation of the reviewers’ comments. *(Corresponding author: Thrupthi Ann John.)*

Thrupthi Ann John and C. V. Jawahar are with the Center for Visual Information Technology, International Institute of Information Technology Hyderabad, Hyderabad 500032, India (e-mail: thrupthi.ann@research.iiit.ac.in).

Vineeth N. Balasubramanian is with the Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Hyderabad 502285, India.

This article has supplementary downloadable material available at https://doi.org/10.1109/TBIOM.2021.3120758, provided by the authors.

Digital Object Identifier 10.1109/TBIOM.2021.3120758

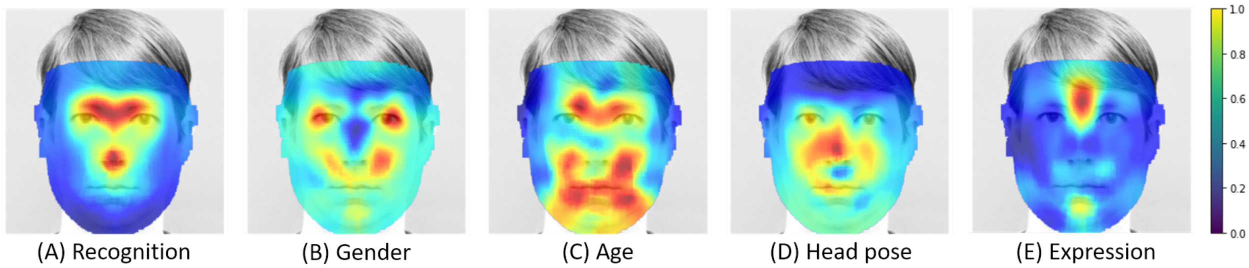


Fig. 1. Are all parts of the face of equal importance for different face classification tasks? In this work, we show that deep models do not give equal importance to the entire face. Canonical Model Saliency (CMS) maps show parts of the face that play a significant role in decisions made by the deep model. CMS maps reveal how deep face models work and allow us to detect and diagnose problems inherent to the models, such as biases. For heatmaps, red indicates a high value while blue indicates a low value (*Best viewed in color*).

intra-class difference is very small and face tasks are a form of fine-grained classification. Input images to face classification models are usually pre-processed so that they are centered around the face of interest and there is only one face per image. Examples of current saliency methods applied to faces are given in Figure 2. We observe that most methods highlight a large area in the center of the face. This type of heatmap may be useful for object recognition when there are multiple objects in a single image, but shows only trivial information for face images. Since faces are centered in the input image, the question ‘where in the image’ is not as relevant as ‘where on the face’. In this work, we introduce a simple yet effective ‘standardization’ process for visualization of deep learning models for face processing, that converts image coordinates to face coordinates and thus makes the resultant saliency maps more effective in practice. We utilize the structure of faces and project the saliency maps onto a standard frontal face to obtain *‘Canonical Saliency Maps’* that are independent of image coordinates. These canonical saliency maps can be further processed to compare images or observe trends.

To this end, we propose two types of canonical maps: Canonical Image Saliency (CIS) maps and Canonical Model Saliency (CMS) maps. CIS maps are detailed attention maps of input faces projected onto a standard frontal face. CMS maps, on the other hand, globally visualize the characteristic heatmap of an entire face network, as opposed to an input image. This shows the general trend of facial features a network fixates on while making decisions. Such a model-level saliency map can only be generated using a canonical approach, and not by currently available saliency maps. CMS maps highlight areas that are of most significance for a given face task across a dataset. Since we need only the confidence of the classifier for this purpose, these can be generated for any available model or architecture, even if the implementation details are not available. Thus, this approach may even be used for analyzing commercial models that may not reveal their architecture designs.

In order to validate our contributions comprehensively, we study our canonical maps on five different face processing tasks: face recognition, gender recognition, age recognition, head pose estimation and facial expression recognition (Figure 1). We use well-known architectures in our studies and also compare the fixation patterns of the models for human recognition of faces. We also show that our visualization method helps discovers a bias in gender recognition models which rely on eye make-up to make decisions.

Our key contributions can be summarized as follows.

- We present a method to standardize face saliency images and project them from image coordinates to face coordinates. This ‘standardization’ produces canonical heatmaps that show the relevance of different facial parts to a deep face task. The new maps are more insightful than the saliency maps produced by current methods and can be used for comparison and observation of trends.
- We introduce two types of canonical heatmaps: (i) Canonical Image Saliency maps which highlight the significant facial areas of a specific input image pertinent to a prediction; and (ii) Canonical Model Saliency maps, which capture global characteristics of an entire deep face model while making predictions across data points, which allows us to understand the network and potentially diagnose problems.
- Our algorithms can be performed on any face model even if the implementation is not available. We demonstrate the superior performance of our method using extensive suite of experiments.
- We explore the working of deep face models trained for various face tasks having different architectures. We illustrate how to interpret the canonical maps and demonstrate their diagnostic utility by detecting a bias that arises from using a celebrity face dataset to train a deep network to classify gender.

II. Related Work

There has been extensive research dedicated to saliency visualization methods in recent years. One of the first efforts to obtain image saliency was by Simonyan *et al.* [10] which used the magnitude of the gradients to obtain a noisy and scattered saliency map. Zeiler and Fergus [15] and Springernberg *et al.* [12] subsequently introduced methods to highlight the important details of the image. These visualizations were not class-sensitive. Zeiler and Fergus [15] also proposed a method to obtain coarse class-specific saliency maps by occluding parts of the input image and monitoring the output of the classifier. Recent works such as CAM [16], GradCAM [6], GradCAM++ [13] and ScoreCAM [14] proposed gradient-based methods to produce coarse, class-sensitive saliency maps that highlights areas of the input image that were influential in the classifier output. Smilkov *et al.* proposed a technique called ‘SmoothGrad’ [11]

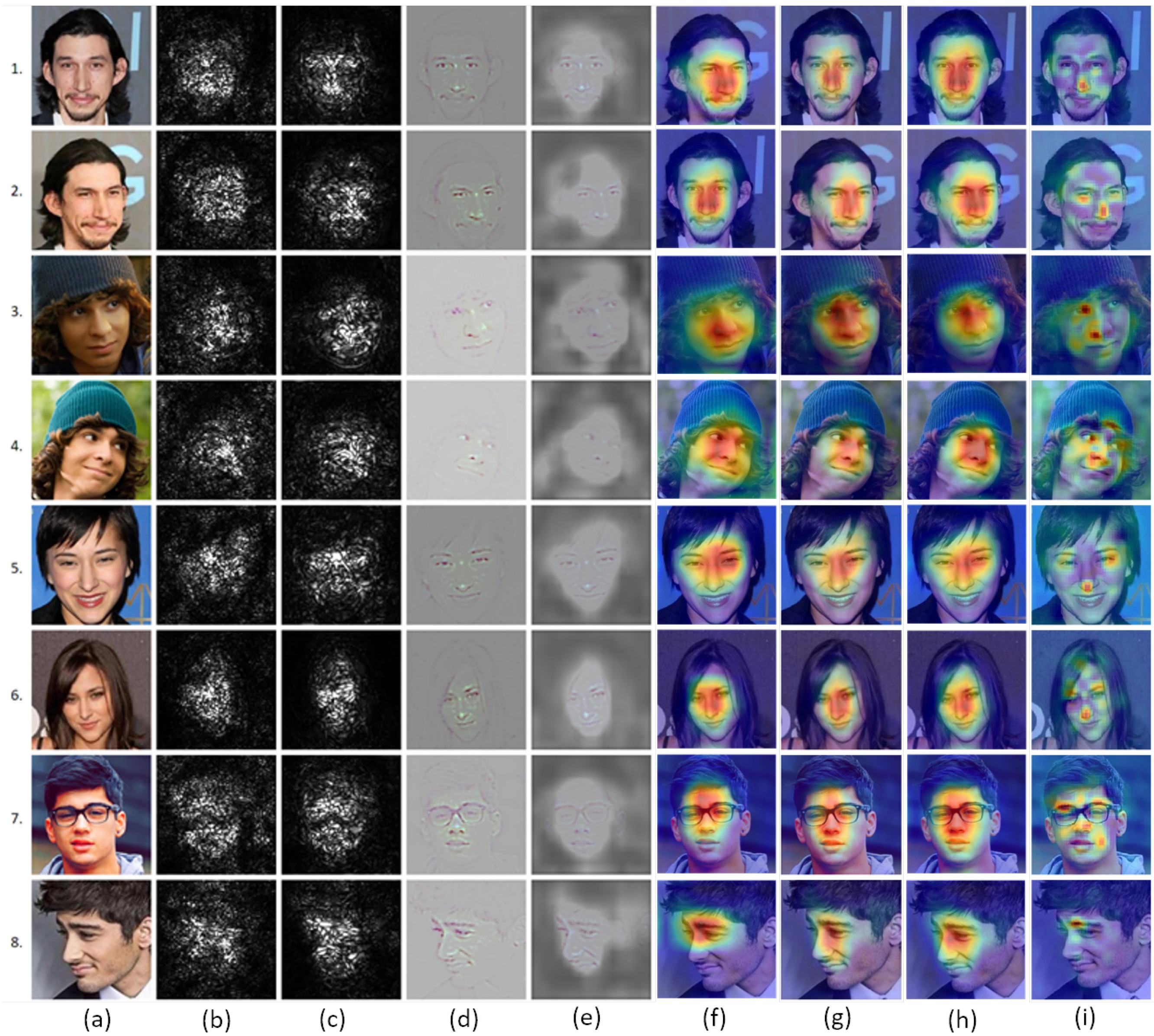


Fig. 2. A comparison of various saliency visualization methods on the VGG-Face model [9] for the task of face recognition. For each image, the target class of the visualization is the ground truth class. (a) Original image; (b) Vanilla gradients [10]; (c) Smooth-grad [11]; (d) Guided Backpropagation [12]; (e) Guided GradCAM++ [13]; (f) GradCAM [6]; (g) GradCAM++ [13]; (h) ScoreCAM [14]; (i) Occlusion map [15]. Images are taken from the VGG-Face dataset [9]. Rows (1, 2), (3, 4), (5, 6) and (7, 8) have the same identity. (Best viewed in color).

which produced a smooth version of such maps by averaging gradient maps after perturbing the input image with noise.

Although there have been many methods introduced for saliency visualization for general image classification settings, such methods do not explicitly address non-trivial fine-grained details when used on face images, as shown in Figure 2. Columns (b) and (c) in the figure show results of methods that use the magnitude of gradients to produce a heatmap. These heatmaps are scattered and it is difficult to see the details and interpret classification results using them. Guided backpropagation, shown in column (d), shows the finer details of the face, but is not class-sensitive, thus reducing their utility for interpretation. Columns (f), (g) and (h), corresponding to GradCAM [6], GradCAM++ [13] and ScoreCAM [14], are class-specific, but most commonly highlight the central area of a face making them uninformative across different face processing tasks. Column (e) represents the results of Guided GradCAM++, obtained by multiplying the output of guided backpropagation with the GradCAM++ heatmap, shows fine details while highlighting the class-discriminative area of the face. Occlusion maps in column (i) of Figure 2 seem to give the most informative results for our use case. This method maps the impact that each region of the image has on the classification, in effect mapping out how representative of the class each region is. It produces a more non-trivial heatmap showing finer details than the other heatmaps. The heatmap resolution can also be adjusted by changing the size of the occlusion and the stride, and the method can be used with any architecture and loss function. Our visualization method is hence built on occlusion maps given this inference from our studies on face images. The closest method to ours is [8], which uses occlusion maps generated between pairs of similar-looking face images to assist humans in telling them apart. They do this by aligning two faces using keypoints and systematically occluding patches of both faces and recording the change in cosine similarity between the faces on a heatmap. The resulting heatmaps reflect the degree of difference between the face pairs. Unlike this work, our method works on multi-class classification tasks and introduces the face canonicalization procedure.

Our proposed Canonical Model Saliency Maps visualize saliency of face networks w.r.t. different regions of the face for different face processing tasks. These maps allow us to conduct useful analysis by comparing the facial areas important to the network to the areas that are expected to be important to classify the task. However, the challenge herein is - how do we obtain the 'correct' expectations to compare the network's saliency map to? One may look at human cognition as a benchmark for what a deep network should see.

Extensive research exists on how humans recognize faces; important results have been presented recently in [17]. For, e.g., humans are known to be good at recognizing low-resolution and degraded faces, when compared to machines. There is a marked difference in the recognition rate of humans when seeing familiar faces when compared to unknown faces. The face's top part, especially the eyebrows, is known to be an important cue for human face recognition [17]. Comparing our face saliency maps with such insights can tell us when the obtained saliency maps of trained networks point to wrong cues for classification (see Section V-B for examples.). Our results on gender and age agree with some of the earlier conclusions [18], [19], [20], [21] on the usefulness of eyes and lips for gender or the eyes and mouth corners for age, but also provide new insights such as the importance of eye corners for gender due to make-up, in addition to providing a methodology for such analysis. We now describe our methodology.

III. Canonical Saliency Maps: Methodology

The key aim of our methodology is to create a visualization which highlights the discriminative parts of a face for a given task. Our method is based on the assumption that the discriminative importance of a part of an input image is proportional to the drop in confidence of the classifier when the part is occluded [15], however on a canonical face representation. Like other occlusion-based saliency map methods, given an image $I \in \mathbb{R}^{W_I \times H_I}$ and the coordinates (i, j), the importance of a patch ($|i-x| < \frac{sz}{2}\ \forall x < W_I$, $|j-y| < \frac{sz}{2}\ \forall y < H_I$) is given as follows:

$$S_{i,j} = \phi(I, c) - \phi\big(I \odot B_{i,j}, c\big) \tag{1}$$

where $\phi(I, c)$ is the confidence of class c for image I and $B_{i,j} \in \{0,1\}^{W_I \times H^I}$ is a mask such that:

$$B_{i,j}[x]\big[y\big] = 0 \text{ if } |i-x| < \frac{sz}{2} \text{ and } |j-y| < \frac{sz}{2} \tag{2}$$

$$= 1 \text{ otherwise} \tag{3}$$

and sz is the size of the patch, which is a hyperparameter.

A. Alignment to a 'Canonical' Face

In order to capture the finer details of the parts of an image a trained DNN model looks at, we compute our saliency map on a standard neutral frontal face image $F \in \mathbb{R}^{W_F \times H_F}$ called

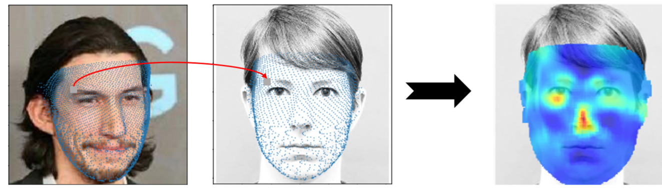


Fig. 3. Procedure of computing Canonical Image Saliency (CIS) map. First, the input face is densely aligned. Each part of the input face is occluded with a small patch and the classification confidence is obtained. The drop in confidence is plotted on the same face location on a neutral face image to obtain the Canonical Image Saliency map.

the *canonical face*, which helps compare saliency maps on a standardized platform.

We find an one-to-one mapping between the input face image and the canonical face image by fitting a 3D modular morphable model (3DMMM) [22] using the procedure used by PR-Net [23]. In particular, we use a convolutional neural network to regress a UV positional map from the input image, which gives the depth for a set of fixed points on the UV map of the face. For details of this procedure, please see [23]. Let $M \in \mathbb{R}^{N \times 3}$ be a set of N 3D points representing the 3DMMM. We fit it on the input image I and the canonical image F to obtain the set of 2D points $M_I \in \mathbb{R}^{N \times 2}$ and $M_F \in \mathbb{R}^{N \times 2}$ as the projection of M on I and F respectively. Thus, we have a 1:1 dense mapping of points from I to F such that $I[M_I[n,1]][M_I[n,2]]$ refers to the same facial feature as $F[M_F[n,1]][M_F[n,2]]\ \forall n \in \{1.2, \ldots, N\}$.

B. Mapping Discriminative Areas

The Canonical Image Saliency (CIS) map is generated by accumulating the drop in confidence at each point of the dense alignment matrix M_I and recording it on the corresponding location of F on an intermediate matrix $P^* \in \mathbb{R}^{W_F \times H_F}$ as follows:

$$\begin{aligned} P^*_{M_F[n,1],M_F[n,2]} = {} & P^*_{M_F[n,1],M_F[n,2]} \\ & + S_{M_I[n,1],M_I[n,2]} \\ & \forall n < N \end{aligned} \tag{4}$$

where $P^*_{M_F[n,1],M_F[n,2]}$ is the patch around the point $(M_F[n,1], M_F[n,2])$ on the heatmap P, and $S_{M_I[n,1],M_I[n,2]}$ is the drop in confidence in the patch around point $(M_I[n,1], M_I[n,2])$ calculated according to Equation (1). Note that self-occluded areas cannot be mapped to the canonical faces. This is acceptable however, as self-occluded regions of a face are not class-discriminative for a given image and face task.

C. Density Normalization

Note that an equi-spaced grid on a 3-dimensional face may not correspond to equi-spaced grid on a 2D projection of the face. For example, on a frontal face image, the points on the sides of the face may be more spatially concentrated due to the curvature of the face. The heatmap values in these regions will hence be higher due to the concentration. We hence introduce a normalization step that keeps track of the number of times a pixel on an image is occluded, when performing the occlusion heatmap on the mesh. Let $N \in \mathbb{R}^{W_F \times H_F}$ be a matrix

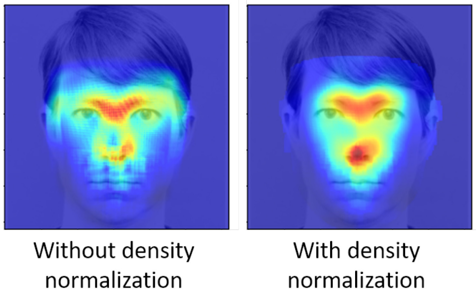


Fig. 4. Effect of applying density normalization to the heatmap. Without density normalization, the nose is not highlighted despite it being a discriminative feature, mainly because the density of points on the nose is low.

which stores the count of times each pixel of P^* was updated. The final CIS map is calculated as follows:

$$P = P^* \oslash N \tag{5}$$

where $\oslash$ represents element-wise division. Figure 4 shows the effect of density normalization on the CIS map.

D. From Image Saliency to Model Saliency

We now discuss how the CIS maps are used to understand facial features that are important across all images for a given model trained for a specific task (for, e.g., the part of the face that may be important for gender recognition vs another part that may be important for age recognition). We call these Canonical Model Saliency (CMS) Maps, which are model-level saliency visualizations to highlight facial areas that influence the model across all test images.

Given a test set D consisting of images $\{I_1, I_2, I_3, \ldots\}$ with variations in factors such as pose, lighting, or expressions, we consider the average CIS map across these test images as the CMS map, i.e.,

$$V = \frac{1}{N}\sum_i P_i \quad \forall I \in D \tag{6}$$

where P_i is the CIS map of $I_i \in D$. It is possible to combine the CIS maps in other ways, but we found that simple averaging worked well in practice for model-level analysis. Learning CMS maps in other ways could be an interesting direction of future work. Furthermore, in practice, we observe that it requires only a few images to generate a stable CMS map for a complete trained model. This suggests that face networks consistently rely on a few facial features and the canonical visualizations are stable across images. This is shown in Figure 5 where we see that the trends become obvious from the first random 100 images. After 1000 images, the CMS is practically unchanged with the addition of more images.

Figure 6 shows a comparison between occlusion heatmaps of [15] and our CIS maps. Our methodology is summarized as follows.

E. Model Saliency for Non-Classification Tasks

CMS maps can be generated for any face model which has a measure of confidence associated with each input image. Our method can be adapted to non-classification models by defining an appropriate confidence function. Here, we

Algorithm 1 Canonical Image Saliency Map

Input: nosep,leftmargin=0.5in,topsep=0pt
- input image I of size $W_I \times H_I$
- input mesh M_I of size $N \times 3$
- frontal image F of size $W_F \times H_F$
- frontal mesh M_F of size $N \times 3$
- model ϕ: deep model to find saliency where $\phi(I, c)$ gives the confidence of I for class c
- target class C of the input image I
- sz: size of occlusion square

Output: heatmap P of size $W_F \times H_F$

procedure CIS$(I, M_I, F, M_F, \phi, C, sz)$
- $P \leftarrow \{0\}^{W_F \times H_F}$
- $N \leftarrow \{0\}^{W_F \times H_F}$
- $fsz \leftarrow fsz \times \frac{H_F}{H_I}$
- **for** $i \leftarrow 0$ **to** n **do**
  - $I^* \leftarrow I$
  - $I^*[M_I[i,0] - \frac{sz}{2} : M_I[i,0] + \frac{sz}{2}][M_I[i,1] - \frac{sz}{2} : M_I[i,1] + \frac{sz}{2}] \leftarrow 0$
  - $x_F, y_F \leftarrow M_F[i,0], M_F[i,1]$
  - $P[x_F - \frac{fsz}{2} : x_F + \frac{fsz}{2}][y_F - \frac{fsz}{2} : y_F + \frac{fsz}{2}] += \phi(I, C) - \phi(I^*, C)$
  - $N[x_F - \frac{fsz}{2} : x_F + \frac{fsz}{2}][y_F - \frac{fsz}{2} : y_F + \frac{fsz}{2}] += 1$
- $P[N = 0] \leftarrow 0$
- $N[N = 0] \leftarrow 1$
- $P \leftarrow P \oslash N$
- return P

define the confidence function for two commonly-used face tasks: zero-shot recognition using nearest neighbor and face verification.

1) Zero-Shot Face Recognition: Here, the query image q is assigned the label of the image from the training set whose features have the highest cosine similarity with the features of the query image [24]. We define the confidence of classification in this setting as follows:

$$S_{q,c} = \frac{A.Q}{\|A\|\|Q\|} \tag{7}$$

where c is the ground truth label of q, Q is the feature of q and A is the feature of the closest training set image with label c. This new confidence function can be used in place of the class confidence ϕ in Equation (1).

2) Face Verification: Here, a pair of face images is considered to have the same identity if the cosine similarity between their features is more than a threshold calculated on the training set [24]. We define the confidence in this setting as follows:

$$S_{q_1,q_2,c} = c \times \left(\tau - \frac{Q_1.Q_2}{\|Q_1\|\|Q_2\|}\right) \tag{8}$$

where $c \in \{-1, 1\}$ is the verification ground truth label, τ is the verification threshold, and Q_1 and Q_2 are the features of the image pair q_1 and $q2$. Using this function, we generate an IMS map for each pair of images and calculate the CMS map using Equation 6.

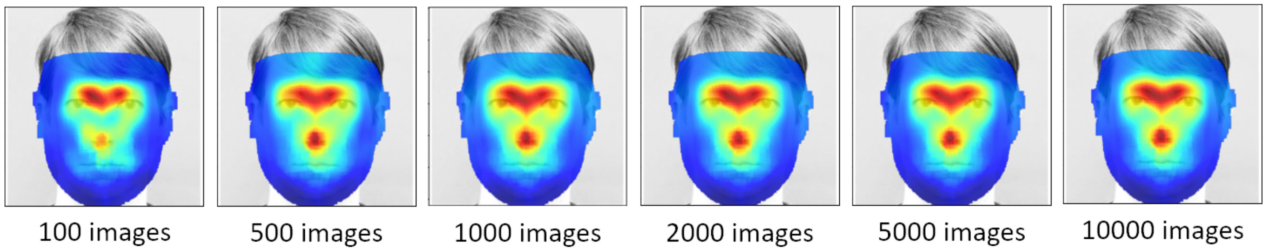


Fig. 5. Ablation study to study the effect of the number of images used to create a CMS map. CMS maps for recognition using 100, 500, 1000, 2000, 5000 and 10000 random CIS maps.

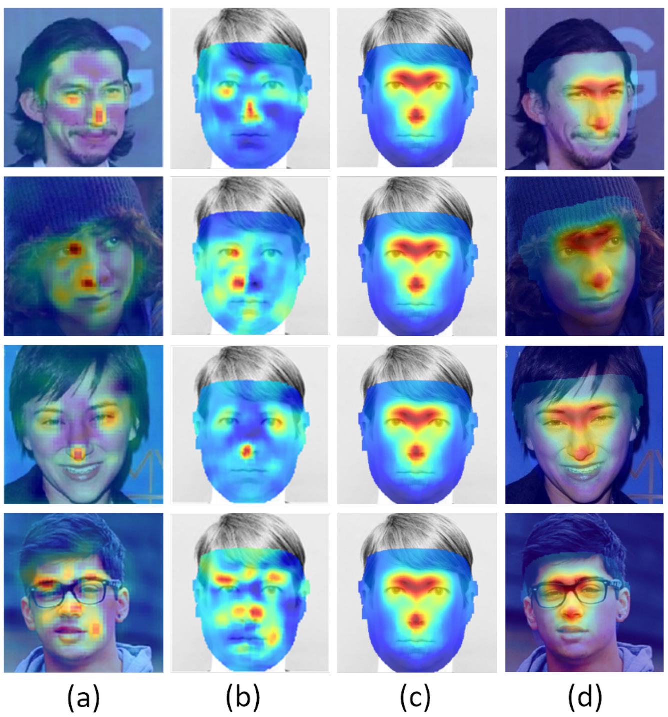


Fig. 6. Column (a) shows Occlusion Maps used for saliency visualization (see Section II); Column (b) shows Canonical Image Saliency (CIS) maps. CIS maps are a projection of occlusion maps onto a canonical frontal face; Column (c) shows Canonical Model Saliency (CMS) maps. These maps are generated for a model as a whole and hence do not vary with input; Column (d) shows the CMS maps reprojected back onto the input face.

IV. Experiments and Results

We now present our comprehensive experimental results, that analyze the effectiveness of canonicalizing saliency maps for face processing tasks. First, we explore our saliency maps through visual examples in Section IV-A. Second, we objectively assess the ability of our visualization to highlight discriminative parts of the face in Section IV-B. Third, we present the results of a user survey which shows that the parts of the face highlighted by our algorithm are important for the human perception of facial attributes in Section IV-C. Finally, we present extensive ablation experiments and discussions on our method in Section V. Unless otherwise mentioned, our experiments are conducted using the VGG-Face pre-trained model [9] based on the VGG-16 architecture [25]. We use a random subset of the CelebA dataset [26] consisting of 22,000 images (henceforth called CelebA-subset) for all our experiments. (Note that these images are only used in the model's test phase, the model by itself is trained on all the training images in the CelebA benchmark). See the Supplementary Section for more details.

A. Qualitative Results

We compare the saliency maps produced by various methods in Figure 2. As in Section II, most visualizations are not practically useful, and highlight a vague central portion of the face. In Figure 6, we display the visualization methods introduced in this work. From simple occlusion maps in column (a), we obtain Canonical Image Saliency (CIS) maps by projecting the occlusion maps onto a neutral frontal face, as shown in column (b). This 'canonicalizing' allows us to collate the CIS maps to create Canonical Model Saliency (CMS) maps as shown in column (c). In column (d), we show that when the CMS maps are reprojected onto the input images, the saliency maps become meaningful for analysis.

1) Evaluation of Canonical Model Saliency Maps on Various Face Tasks: For this experiment, we used our algorithm on five models trained for the tasks of classification, expression, head pose, age and gender. We used the VGG-Face [9] pre-trained model, and finetuned it for each of the aforementioned tasks on the CelebA [26] dataset. The ground truth labels for gender are provided with the CelebA dataset. The head pose ground truth was obtained by using PRNet [23], and the age ground truth was obtained using the DEX method [27]. For expression, the ground truth for CelebA was obtained from a model trained on the FER 2013 data set [28]. Since both head pose and age are real-valued, we grouped the values into discrete bins to convert them into classification tasks. For pose, the yaw and pitch values were binned into 9 bins ranging from top-left to bottom-right (see Figure 19). Similarly, the real-valued ages obtained from the DEX model were grouped into 10 bins, each having 10 years. More details of the networks used are given in the Supplementary Section S1. The generated CMS maps are shown in Figure 1. We notice how models of the same architecture trained on different tasks focus on different face areas. For recognition, the eye-nose triangle is important and there is less focus on the mouth or the chin. Gender models surprisingly find the corners of the eyes to be the most discriminative facial features. We discuss the implications of this in Section V-B. The nose is a crucial feature for the head pose model and the area between the eyebrows for the expression model. The age model looks at many different facial features. We see that CMS maps are a valuable asset to understand the nature of face tasks and the characteristics of various deep

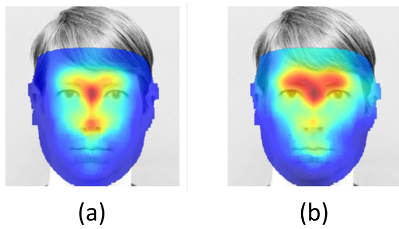


Fig. 7. Calculating CMS maps for non-classification tasks on the LFW dataset: (a) CMS map for zero-shot learning of identity using nearest neighbor; (b) CMS map for face verification.

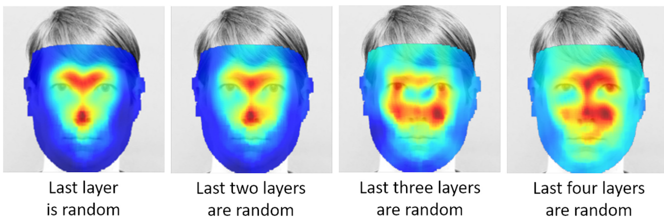


Fig. 8. Sanity check on our visualization method. We progressively randomized the layers of the VGG-16 face model starting with the output layer as described in [29]. We observe that the CMS map gets progressively randomized; our method passes the sanity check. (a) Last layer randomized; (b) Last two layers randomized; (c) Last three layers randomized; (d) Last four layers randomized.

models when addressing these tasks. We discuss some of these results in more detail in Section V.

2) Canonical Model Saliency Maps on Non-Classification Tasks: In this experiment, we show that CMS maps can be generated for non-classification face tasks. We generated CMS maps for zero-shot learning of face identities using nearest neighbor and face verification of VGG-Face fc1 features on the LFW [2] dataset. For the zero-shot learning task, we occluded parts of the query image while using Equation (7) as the confidence function. For the verification task, we occluded the same region of both images in a verification pair and used Equation (8) as the confidence function. The results are shown in Figure 7. In both cases, we see the highlighted facial areas are similar to the classification task of recognition in Figure 1.

3) Sanity Check Using Randomization: Reference [29] proposed a sanity check for saliency maps, where the layers of a trained model are progressively randomized starting from the output layer, and the changes in generated saliency maps are observed. A method is said to pass the sanity check if progressive randomization increases the randomization of the corresponding visualization. We perform a sanity check on our visualization using the same procedure, and reports the results of this experiment in Figure 8. We observe that as more layers get randomized, the visualization gets more randomized. Thus, our method passes the sanity check.

B. Quantitative Results

We conduct an objective evaluation of the faithfulness of our method on two datasets: CelebA and LFW [2] and compare it with three popular saliency visualizations: GradCAM [6], GradCAM++ [13] and ScoreCAM [14]. Similar to [13], [14], we measure the confidence drop of explanation images produced by pixel-wise multiplication of the saliency heatmap

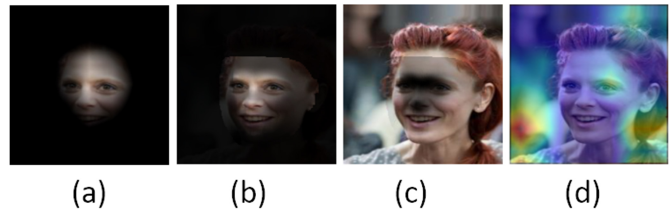


Fig. 9. Since face models are trained to look holistically at the face, they have more confidence in figure (a) than in figure (b), even though figure (b) highlights more relevant features. Thus, we use negative saliency maps where darkening relevant features should cause a larger drop in confidence. This also ensures that there is enough context for the model to interpret the face holistically. Another reason for using negative saliency maps is to take care of cases where a visualization method does not interpret the face correctly, as in figure (d). Here, the heatmap completely misses the face and is focused on disparate parts of the image. Using normal explanation maps will result in almost the original image, which will give a high score in the metrics used. This is avoided by using negative explanation maps and normalizing the sum of pixels.

with the base image. In particular, we utilize a 'negative explanation image' by darkening the relevant areas of the base image. Unlike the task of object recognition, face images have a single object at the center of the image, and models trained on face images focus on different parts of the face image. In this process, saliency maps at times fail to detect the face completely (see Figure 9). Using negative explanation maps addresses such concerns. The negative explanation image E is given by:

$$E = (1 - H) \otimes I \tag{9}$$

where H is the heatmap, I is the base image and $\otimes$ represents pixel-wise multiplication. The heatmaps are first normalized to a range of [0, 1] and the heatmaps for all the methods are standardized to have the same sum of pixels for each image:

$$H' = \frac{h - \min(h)}{\max(h) - \min(h)}; H = \frac{s}{\Sigma H'}H' \tag{10}$$

where h is the original heatmap, s is a scalar which is the same for all heatmaps of the same image, and H is the final heatmap which is used to create negative explanation maps. Normalizing the heatmaps in this way ensures that no visualization method gets an advantage of highlighting a large area of the input image, as only the discriminative parts should be highlighted.

We adopt the three metrics used in [13] with negative explanation images.

Average Drop %: The confidence of an image when passed through a model is expected to decrease when the most discriminative parts are covered. We measure the drop of confidence when compared to the unmodified image as:

$$\frac{1}{N}\sum_{n=1}^{N} \max\left(0, \frac{M(I_n) - M(E_n)}{M(I_N)}\right) \times 100 \tag{11}$$

where $M(E_n)$ and $M(I_n)$ are the confidence values of the n^{th} explanation image and original image respectively. A high value of Average Drop % indicates that the heatmap accurately highlights the most discriminative parts of the image.

% Increase in confidence: In some images, covering the highlighted parts may result in an undesired increase in confidence with respect to the original image. We measure the

number of such images using this measure as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{M(E_n)>M(I_n)} \times 100 \tag{12}$$

where $\mathbb{I}$ is the indicator function which returns 1 if $M(E_n) > M(I_n)$ and 0 otherwise. A low score in this metric is better.

Win %: Here, we compare all the four methods and measure which method produces the greatest drop in confidence for a given test image. For example, Win % of CMS is calculated as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{M}\left(E_n^{CMS}\right) < \left(M\left(E_n^{GradCAM}\right),\right.$$
$$\left. M\left(E_n^{GradCAM++}\right), M\left(E_n^{ScoreCAM}\right)\right) \times 100 \tag{13}$$

where the indicator returns 1 if the explanation map produced by CMS has the lowest confidence. The sum of Win % across all the visualization methods for a single task should add up to 100.

We conduct three experiments for quantitative evaluation. First, we calculate the above metrics on VGG-16 for the tasks of recognition, gender, age, head pose and expression on the CelebA dataset. For fair comparison, we use our maps projected back onto the input image (Col (d) of Figure 6). Figure 10 shows our results and a comparison with other visualization methods. Our method outperforms all other methods in all metrics. The Win % shows that for most images, removing the explanation map given by our method causes the highest drop in confidence (higher the better).

Secondly, we repeat the experiment on the LFW [2] dataset using the VGG-Face network, using the same experimental settings as above. We show the results in Figure 11. Here too, our method outperforms all other methods by a large margin in all quantitative metrics, showing that our method generalizes across datasets.

We also compare our saliency methods on various off-the-shelf gender models. We use pretrained models from [27], [30], [31] and evaluate our metrics on CelebA-subset. More details about these models are given in the Supplementary Section S1. Our results are shown in Figures 12. Once again, we see that our method outperforms all other methods on all metrics. We show the CMS maps obtained using the various networks in Figure 13.

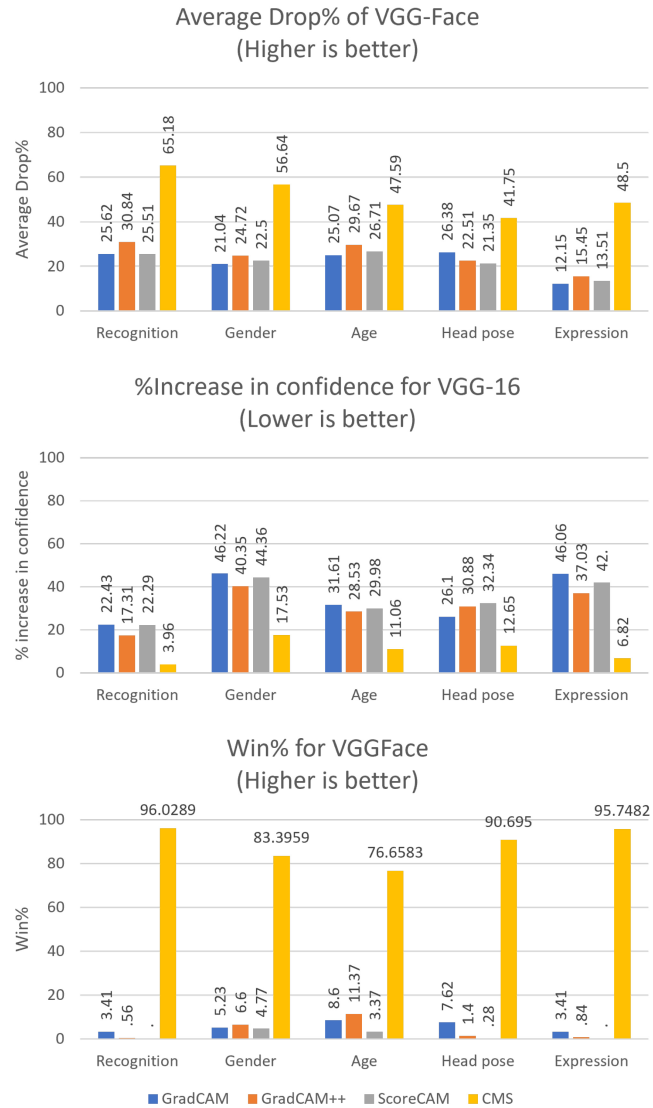


Fig. 10. Results for Average Drop %, % Increase in Confidence and Win % of VGG-16 on Celeb-A for the tasks of recognition, gender, age, head pose and expression.

C. User Survey on Perception of Facial Attributes

We conducted a user survey to evaluate the human interpretability of our saliency maps as compared to other visualization methods. In particular, we explored whether the discriminative facial areas found by Canonical Model Saliency Maps are vital for human perception of facial attributes. We focused on the tasks of gender and expression for this study. The survey used a total of 96 images, each of which were evaluated by 154 participants not involved in this work. Twelve base images for each task were used, for which we generated four negative explanation maps corresponding to the four saliency visualization methods GradCAM, GradCAM++, ScoreCAM and reprojected CMS maps using the Gender and Expression models mentioned in Section IV-B. We also applied a vignette to each of the explanation images to hide the context information (see Figure 15 for sample images). Each participant was given a binary choice for each image (male-female or happy-sad, depending on the task). Since a better visualization algorithm hides crucial information and makes it more difficult to interpret an image, we use the percentage of wrong answers as a measure of the goodness of the visualization method. We show some sample survey images in Figure 14. See the Supplementary section for more examples.

The results of our survey are given in Figure 15. We see that the percentage of wrong answers marked by the respondents is higher for our method than other methods, indicating that our method performed better at hiding the most crucial and discriminative facial areas.

V. Analysis and Discussion

In this section, we present analysis of the proposed method including ablation studies and discussions.

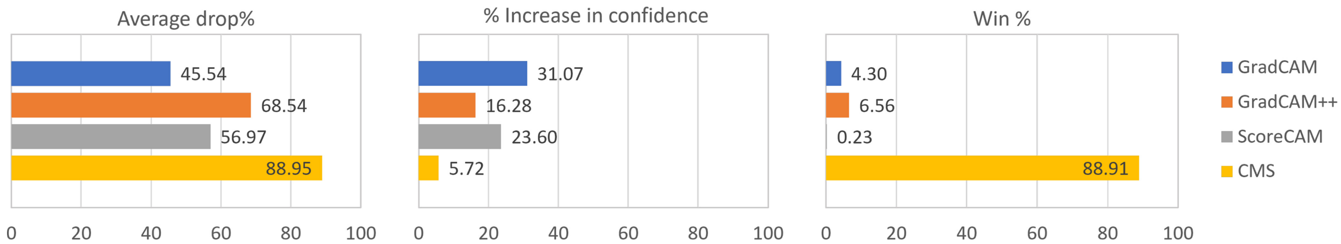


Fig. 11. Results for Average Drop %, % Increase in Confidence and Win % of the explanations generated by Grad-CAM, Grad-CAM++, ScoreCAM and CMS on LFW for the VGG-16 model.

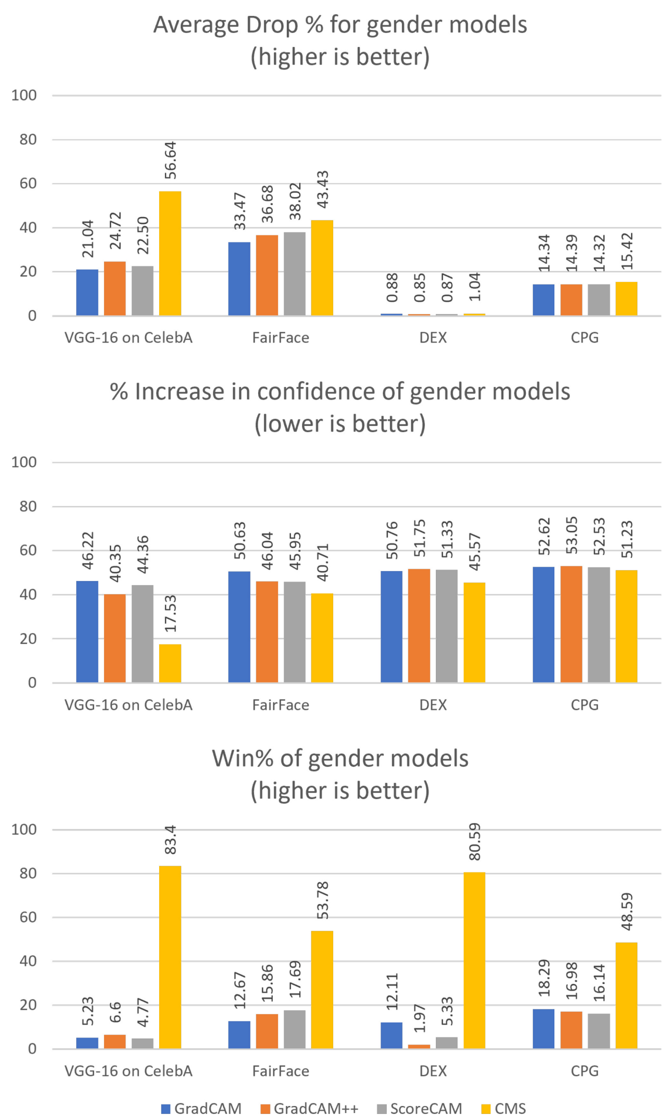


Fig. 12. Results for Average Drop %, % Increase in Confidence and Win % of the explanations generated by Grad-CAM, Grad-CAM++, ScoreCAM and CMS on CelebA for various deep face gender models.

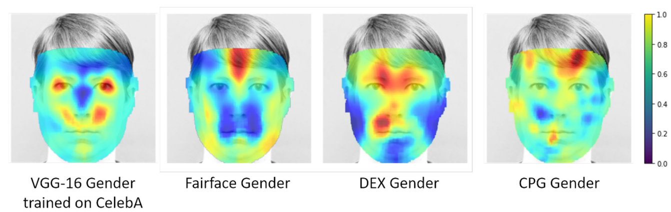


Fig. 13. We compare CMS maps obtained from various off-the-shelf deep gender models.

Fig. 14. Samples of figures used in our survey (see Section IV-C.

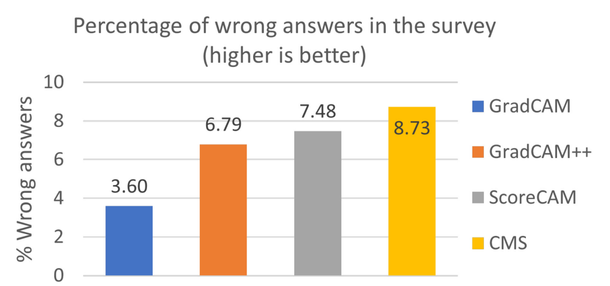


Fig. 15. Results for user survey on the perception of gender and emotion on explanation maps. We used 12 base images modified using GradCAM, GradCAM++, ScoreCAM and CMS. The users had to pick binary labels for each image (male-female, happy-sad). Each question was answered by 143 people who were not involved in this project.

A. *Why Model-Level Saliency Maps?*

Canonical Model Saliency (CMS) maps allow us to observe patterns and trends in the functioning of deep face models by adding the simple yet powerful step of alignment of occlusion-based saliency maps to a canonical face model. For example, using CMS maps, we observed that the corners of the eyes are important for gender classification (Section V-B). This is not directly apparent by observing individual, unaligned occlusion maps, as seen in Figure 16. The advantage of this alignment process is in allowing comparison and aggregation of saliency maps. A single occlusion map may contain variations caused by differences in the image setting such as pose, occlusion and lighting, thus not allowing us to understand the whole picture. The process of aggregation averages out the effects of variations in individual images, showing us the parts of the face that are truly important.

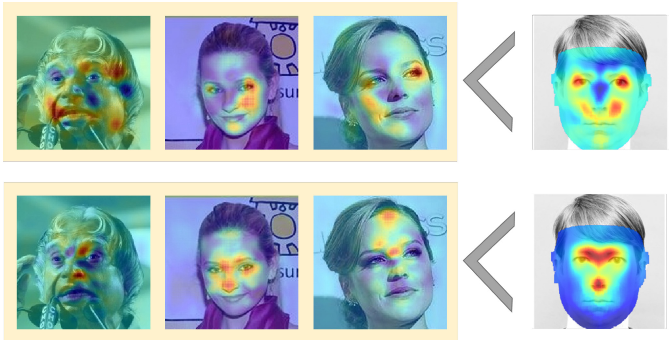


Fig. 16. In this figure, we compare individual occlusion maps of gender (first row) and recognition (second row) to the respective cumulative model saliency maps on the right. Individual occlusion maps vary widely and may have slightly different areas highlighted due to differences in pose, occlusion and lighting. Thus, it is hard to compare these images and get the big picture from them. Aggregating heatmaps gets rid of tiny differences caused due to the conditions in which the photo is taken, allowing us to gain valuable insights.

B. Effect of Make-Up on Gender Classification

The CMS maps for the gender model provided interesting insights using our method (Figure 1E). We expected the heatmap to highlight the areas around the mouth, jaw and cheeks, as they contain facial hair cues and different bone structure for different genders. However, the map showed that the model fixated mostly on eye corners. We hypothesize that this is because the model was finetuned on the CelebA dataset [26], which consists of images of celebrities who use make-up extensively. The model picked up on the cue of eye make-up to classify gender. We presume that such a model will not work well for a different demographic distribution. This may be the reason why many commercial face models fail in detecting gender for females and different races [5]. This indicates the importance of detecting dataset biases as they can have a significant impact on the performance of deep models. We test our hypothesis with the following qualitative experiment. We collect a few images of people with and without eye make-up from the Internet. These images were passed through the gender model and the confidence for 'male' and 'female' classification was observed. Our results are presented in Figure 17. We observed that in all cases, there was a drop of confidence in 'male' classification when the men wore make-up and a smaller drop in confidence of 'female' classification for women without make-up. In some cases, the drop in confidence was large enough to flip the original classification result. This was especially true for males of Asian origin, especially those from the far East. We conclude that eye make-up has a significant effect on the performance of such a gender model, which is skewed towards people of a certain ethnicity.

C. Head Pose Model Relies on the Nose

The shape of the nose changes according to the pose of the face (Figure 19A). Generally, the nose is positioned at the centre of the face, and its placement on the face changes consistently with the 3D orientation of the face. The head pose can be detected quite accurately from the shape of the nose and the quadrant of the face in which the nose tip resides (along with the jawline), especially when there are only nine classes, as shown in Figure 19. The nose thus provides the strongest cue for the head pose. This is reflected in the CMS map shown in Figure 1D.

D. Age Model Uses the Whole Face

The CMS map for age (Figure 1F) shows that the cues for age are present in multiple areas of the face. Some of the distinctive features for age may be the tightness of skin around the eyes and jaws, wrinkles and receding hairline. Pre-deep learning methods used the geometry or texture of the face for age prediction [32], thus corroborating our finding on why age-related cues are found all over the face.

E. How Occlusion Size Affects Saliency Maps

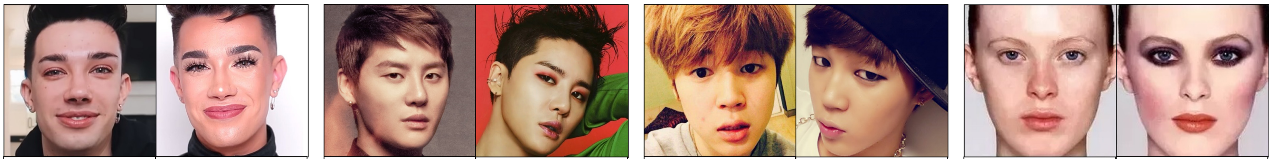
We present a qualitative ablation study to explore the effect of the size of the occluding patch on the generated CIS map. The number of vertices provided by the dense face alignment algorithm is very high and the time required to compute the heatmap at each vertex is large. Hence we use a tunable 'stride' parameter to omit vertices at regular intervals. As the size of the occluding patch decreases, a smaller stride is chosen so that gaps don't appear in the visualization. The stride can be larger for bigger occluding patches without affecting the visualization quality. In Figure 18, we show the result of changing the patch size on the CIS maps generated from the same input image. We observe that as the patch size increases, the map becomes fuzzier but general patterns do not change. Our method provides useful information regardless of the size of the occluding patch, although smaller patches give better resolution. We used a patch of size 15×15 for generating other saliency maps in this work, as it provides a good balance between heatmap resolution and computation time.

F. Why Align to Canonical Face?

Here we examine the need for a canonical face instead of using keypoint-based alignment or the image pixel positions. The main advantage of canonical face alignment is that it ensures that the model saliency maps remain accurate while aggregating the individual image saliency maps. If we do not align the heatmaps precisely, the changes in position add up to produce an inaccurate model saliency map.

We conduct an ablation study to demonstrate this effect. We use three types of alignment and generate model saliency maps on the LFW dataset: 1) no alignment; 2) keypoint-based alignment; and 3) canonical face alignment. For the first case, we create image saliency maps by sliding an occlusion window over the entire input image. We repeat the procedure for the second case, but we used LFW images aligned with keypoint alignment [33] as the input instead of the raw LFW images. The third case used the same setting as previous CMS experiments. We create the model saliency maps for each case by averaging individual image saliency maps. We generate explanation maps and calculate quantitative metrics. The results are shown in Figure 20. Canonical alignment performs better than keypoint-based alignment or no alignment in all cases. We show all three model saliency maps in Figure 21.

Using canonical faces also results in lower computation cost, as we know exactly which parts of the image we need to occlude, as opposed to sliding the occlusion patch over the whole image.



Female	2.28×10^{-5}	0.46	0.15	**0.96**	1.61×10^{-3}	**0.99**	0.06	**0.98**
Male	**0.99**	**0.54**	**0.84**	0.04	**0.99**	2.54×10^{-3}	**0.94**	0.02
	Ground truth: **male**		Ground truth: **male**		Ground truth: **male**		Ground truth: **female**	

Fig. 17. Make-up matters! The figure shows the classification confidence of a gender model on the same person with and without eye make-up. The top row shows the confidence for 'female' classification and the bottom row shows the confidence for 'male' classification. The ground truth label is given below each pair of images.

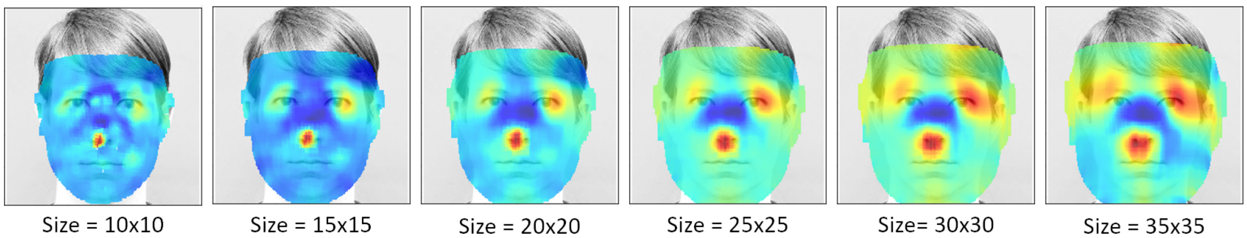


Fig. 18. Canonical Image Saliency maps generated when the size of the occluding patch is varied. We used a patch size of 15×15 in all other experiments in this work.

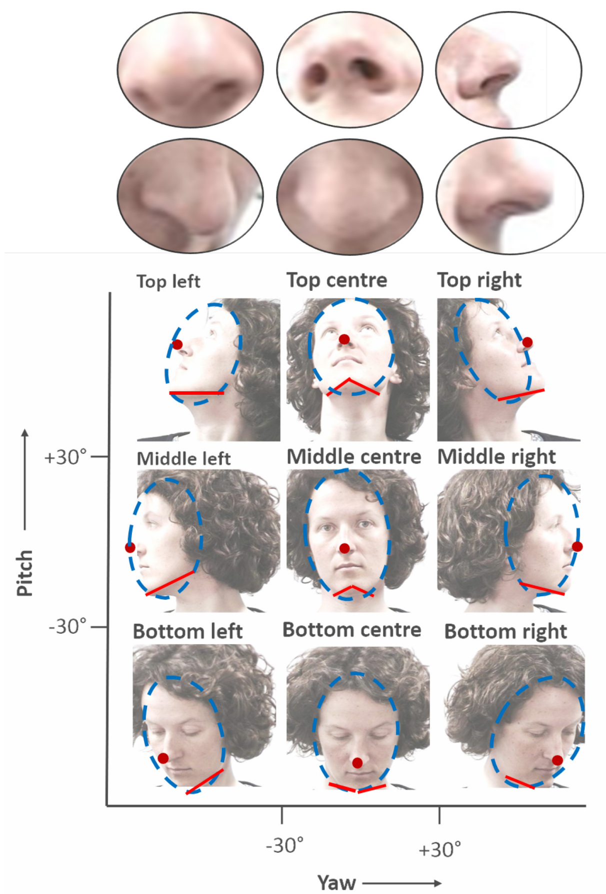


Fig. 19. Look at the close-ups of the nose tip in this figure. Can you tell the 3D orientation of the face with this information? The nose, along with jawline, provide a good cue for the face pose. We also observe that the quadrant of the face area in which the nose tip is found is consistent for the same 3D orientation.

G. Robustness in Deep Models

Robustness refers to the property of a model wherein small deviations in input images, due to noise or natural variations, do not affect the correctness of the model. If a model relies on a small set of cues, it is more likely to go wrong due to input image diversity. Instead, if the model looks at many cues, small variations are less likely to confuse the model. The CMS maps indicate the areas from which deep models pick up cues. The maps thus also allow us to obtain an estimate of the model's robustness. A model that concentrates on a few facial areas is likely to be less robust than one that focuses on many facial areas. Less robust models are more prone to mistakes when presented with extreme cases of occlusion, lighting and other deviations. We see an example with our trained gender model (Section V-B), where the model is not robust to changes in the face due to make-up.

VI. Conclusion

In this work, we showed that standardization of saliency maps via Canonical Saliency Maps provides usable and interpretable results in the face domain when compared to current saliency methods which give trivial outputs for face images. Canonical Saliency Maps highlight the facial areas of importance by projecting occlusion-based heatmaps onto a neutral face. Computing model-level canonical saliency maps enable us to perceive which facial features are important for different face tasks, thereby revealing the strengths and weaknesses of face models. These observations can be compared to human perception, which can show us if the model is behaving in unexpected ways. The maps aid in detecting problems and biases inherent in the model. In particular, by utilizing Canonical Model Saliency maps, we identified a bias in a gender model, wherein the model was wrongly using make-up as a cue to classify gender. We confirmed the presence of the bias with additional studies. Such models can cause problems when used in demographics unlike the training dataset, where the patterns of applying make-up are different.

Nowadays, deep face models are deployed in critical applications like security and law enforcement – the proposed

Fig. 20. Ablation study on the effect of different types of alignment. Shown are the Average Drop%, % Increase in confidence and Win % for three different types of alignment on the LFW dataset: 1. Canonical face 2. Keypoint-based alignment 3. No alignment.

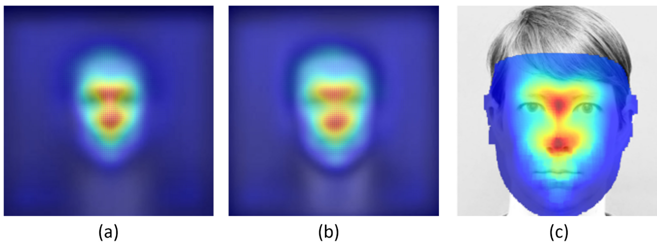


Fig. 21. Ablation study on the effect of different types of alignment. Shown are the model saliency maps for three different types of alignment on the LFW dataset: (a) No alignment, superimposed on the average image of LFW; (b) Keypoint-based alignment, superimposed on the average image of LFW-funneled; and (c) CMS superimposed on the canonical face.

Canonical Saliency Maps allow such systems to be critically analyzed before deployment, and thus increase trust. They can also be used to predict failures during development and help improve the models. We hope that the tools presented in this work, while simple, can be very effective in practical use for deeper understanding of face models, their biases and failures. In future work, we aim to study methods of mitigating the problems and biases detected by our visualization methods.

References

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in *Proc. CVPR*, 2014, pp. 1701–1708.

[2] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Comput. Vision Lab, Univ. Massachusetts, Amherst, MA, USA, Rep. 07-49, 2007.

[3] M. Wang and W. Deng, “Deep face recognition: A survey,” *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021.

[4] *Face Off*, Big Brother Watch, London, U.K., 2018. [Online]. Available: https://bigbrotherwatch.org.uk/campaigns/stop-facial-recognition/report/

[5] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proc. Conf. Fairness Accountability Transparency*, 2018, pp. 77–91.

[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. ICCV*, 2017, pp. 618–626.

[7] Y. Zhong and W. Deng, “Exploring features and attributes in deep face recognition using visualization techniques,” in *Proc. AFGR*, 2019, pp. 1–8.

[8] Y. Zhong and W. Deng, “Deep difference analysis in similar-looking face recognition,” in *Proc. ICPR*, 2018, pp. 3353–3358.

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *BMVC*, 2015, pp. 1–12.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *ICLR Workshop*, 2014.

[11] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, “Smoothgrad: Removing noise by adding noise,” 2017, *arXiv:1706.03825*.

[12] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *Proc. ICLR Workshop*, 2015.

[13] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in *Proc. WACV*, 2018, pp. 839–847.

[14] H. Wang *et al.*, “Score-CAM: Score-weighted visual explanations for convolutional neural networks,” in *Proc. CVPR*, 2020, pp. 111–119.

[15] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. ECCV*, 2014, pp. 818–833.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. CVPR*, 2016, pp. 2921–2929.

[17] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, “Face recognition by humans: Nineteen results all computer vision researchers should know about,” *Proc. IEEE*, vol. 94, no. 11, pp. 1948–1962, Nov. 2006.

[18] H. Han, C. Otto, X. Liu, and A. K. Jain, “Demographic estimation from face images: Human vs. machine performance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[19] T. Ezure, E. Yagi, N. Kunizawa, T. Hirao, and S. Amano, “Comparison of sagging at the cheek and lower eyelid between male and female faces,” *Skin Res. Technol.*, vol. 17, no. 4, pp. 510–515, 2011.

[20] K. Tsukahara *et al.*, “Comparison of age-related changes in wrinkling and sagging of the skin in caucasian females and in Japanese females,” *Int. J. Cosmetic Sci.*, vol. 55, no. 4, pp. 351–371, 2004.

[21] K. Tsukahara *et al.*, “Comparison of age-related changes in facial wrinkles and sagging in the skin of Japanese, Chinese and Thai women,” *J. Dermatol. Sci.*, vol. 47, pp. 19–28, Jul. 2007.

[22] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3D face model for pose and illumination invariant face recognition,” in *Proc. AVSS*, 2009, pp. 296–301.

[23] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, “Joint 3D face reconstruction and dense alignment with position map regression network,” in *Proc. ECCV*, 2018, pp. 557–574.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proc. CVPR*, 2017, pp. 6738–6746.

[25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR*, 2015.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. ICCV*, 2015, pp. 3730–3738.

[27] R. Rothe, R. Timofte, and L. V. Gool, “DEX: Deep expectation of apparent age from a single image,” in *Proc. ICCVW*, 2015, pp. 252–257.

[28] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, “Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order,” *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[29] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” in *Neural Information Processing Systems (NeurIPS)*. Red Hook, NY USA: Curran, 2018.

[30] K. Kärkkäinen and J. Joo, “FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,” in *Proc. WACV*, 2021, pp. 1547–1557.

[31] S. C.-Y. Hung, C.-H. Tu, C.-E. Wu, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, “Compacting, picking and growing for unforgetting continual learning,” in *Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., Inc., 2019.

[32] M. Georgopoulos, Y. Panagakis, and M. Pantic, “Modeling of facial aging and kinship: A survey,” *Image Vis. Comput.*, vol. 80, pp. 58–79, Dec. 2018.

[33] G. B. Huang, V. Jain, and E. Learned-Miller, “Unsupervised joint alignment of complex images,” in *Proc. ICCV*, 2007, pp. 1–8.

[34] P. Huber *et al.*, “A multiresolution 3D morphable face model and fitting framework,” in *Proc. VISIGRAPP*, 2016, pp. 79–86.

[35] P.-L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, and Y. Bengio, “FER-2013 face database,” Univ. Montreal, Montreal, QC, Canada, Rep. 1365, 2013.