

# Explainable Face Recognition

Jonathan R. Williford<sup>1[0000-0002-9178-2647]</sup>, Brandon B. May<sup>1[0000-0002-9914-2441]</sup>, and Jeffrey Byrne<sup>1,2[0000-0001-8973-0322]</sup>

<sup>1</sup> Systems & Technology Research, Woburn, MA 01801, USA

<https://www.stresearch.com>

{jonathan.williford,brandon.may}@stresearch.com

<sup>2</sup> Visym Labs, Cambridge, MA 02140, USA

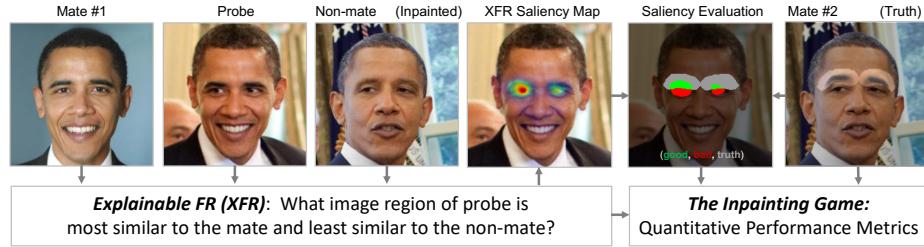
jeff@visym.com

**Abstract.** Explainable face recognition (XFR) is the problem of explaining the matches returned by a facial matcher, in order to provide insight into why a probe was matched with one identity over another. In this paper, we provide the first comprehensive benchmark and baseline evaluation for XFR. We define a new evaluation protocol called the “inpainting game”, which is a curated set of 3648 triplets (probe, mate, nonmate) of 95 subjects, which differ by synthetically inpainting a chosen facial characteristic like the nose, eyebrows or mouth creating an inpainted nonmate. An XFR algorithm is tasked with generating a network attention map which best explains which regions in a probe image match with a mated image, and not with an inpainted nonmate for each triplet. This provides ground truth for quantifying what image regions contribute to face matching. Finally, we provide a comprehensive benchmark on this dataset comparing five state-of-the-art XFR algorithms on three facial matchers. This benchmark includes two new algorithms called subtree EBP and Density-based Input Sampling for Explanation (DISE) which outperform the state-of-the-art XFR by a wide margin.

## 1 Introduction

Explainable AI [29] is the problem of interpreting, understanding and visualizing machine learning models. Deep convolutional network trained at large scales are traditionally considered blackbox systems, where designers have an understanding of the dataset and loss functions for training, but limited understanding of the learned model. Furthermore, predictions generated by the system are often not explainable as to why the system generated this output for that input. An explainable AI system would enable interpretation of what the ML model has learned [26][2], enable transparency to understand and identify biases or failure modes in the system [3][15][13][25] and provide user friendly visualizations to build user trust in critical applications [31][33][42].

Explainable face recognition (XFR) is the problem of explaining why a face matching system matches faces. Human adjudicators have a long history in explaining face recognition in the field of forensic face matching. Professional facial



**Fig. 1.** Explainable Face Recognition (XFR). Given a image triplet of (*probe, mate<sub>1</sub>, nonmate*), an explainable face recognition algorithm is tasked with estimating which pixels belong in a region that is discriminative for the mate - i.e. a region is more similar to the mate than the non-mate. These estimations are given as a saliency map. The nonmate has been synthesized by inpainting a given region (e.g. eyebrows) that changes the identity according to the given network. This provides ground truth for a quantitative evaluation of XFR algorithms using the “inpainting game” protocol.

analysts follow the FISWG standards [11] which leverage comparing facial morphology, measuring facial landmarks and matching scars, marks and blemishes. These features are used to match a controlled mugshot of a proposed candidate to an uncontrolled probe, such as a security camera image. However, these approaches require a candidate list for human adjudication, and a candidate list in a modern workflow is returned from a facial matching system [24]. Why did the face matching system return that candidate list for this probe? What facial features did the face matching system use, and are they the same as the FISWG standards? Is the face matcher biased or noisy? The goal of XFR is to explore such questions, and answer why a system matched a pair of faces. A successful explainable system would increase confidence in a face matching system for professional examiners, enable interpretation of the internal face representations by machine learning researchers and generate trust by the user community.

What is an “explanation” in face recognition? Explainable AI has explored various forms of explanation for machine learning systems in the form of: activation maximization [32], synthesizing optimal images [21], network attention [39][23][12], network dissection [2] or synthesizing linguistic explanations [16]. However, a key challenge in explainable AI is the lack of ground truth to compare and quantify explainable results across networks. XFR is especially challenging because the difference between near-mates or *doppelgangers* is subtle, the explanations are non-obvious, and differences are rarely well localized in a compact facial feature [6].

In this paper, we provide the first comprehensive benchmark for explainable face recognition (XFR). Fig. 1 shows the structure of this problem. An XFR system is given a triplet of (*probe, mate, nonmate*) images. The XFR system is tasked with generating a saliency map that best captures the regions of the probe image that increase similarity to the mate and decrease similarity to the nonmate. This provides an explanation for why the matcher provides a high ver-

ification score for the pair (*probe, mate*) and a low verification score for (*probe, nonmate*). This explanation can be quantitatively evaluated by synthesizing non-mates that differ from the mate only in specific regions (e.g. nose, eyes, mouth), such that if the saliency algorithm selects these regions, then it performs well on this metric. This paper makes the following contributions:

1. **XFR baseline.** We provide a baseline for XFR based on five algorithms for network attention evaluated on three publicly available convolutional networks trained for face recognition: LightCNN [35], VGGFace2 [5] and ResNet-101. These baselines include two new algorithms for network attention called subtree EBP (Sec. 3.1) and DISE (Sec. 3.2).
2. **Inpainting game protocol and dataset.** We provide a standardized evaluation protocol (Sec. 4.1) and dataset (Sec. 4.2) for fine grained discriminative visualization of faces. This provides a quantitative metric for objectively comparing XFR systems.
3. **XFR evaluation.** We provide the first comprehensive evaluation of XFR using the baseline algorithms on the inpainting game protocol to provide a benchmark for future research (Sec. 5.1). Furthermore, in the supplemental material, we show a qualitative evaluation on novel (non inpainted) images to draw conclusions about the utility of the methods for explanation on real images.

## 2 Related Work

The related work most relevant for our proposed approach to XFR can be broadly categorized into two areas: network attention models for convolutional networks and interpretable face recognition.

Network attention is the problem of generating an image based saliency map which visualizes the input regions that best explains a class activation output of a network. Gradient-based methods [31][33][42] attempt to compute the derivative of the class signal with respect to the input image, while other approaches [4] modify network architectures to capture these signals or localize attribution [17]. Excitation backprop [39], contrastive EBP [39] or truncated contrastive EBP [6] formulate the saliency map as marginal probabilities in a probabilistic absorbing Markov chain. Layerwise relevance propagation [18][28][1] provides network attention through a set of layerwise propagation rules theoretically justified by deep Taylor decomposition. Latent attention networks learn an auxiliary network to map input to attention, rather than exploring the network directly [14]. Inversion methods [19] seek to recover natural images that have the same feature representation as a given image. However, the same insights have not yet been applied to fine grained categorization for face recognition. Finally, black box methods have explored network attention for systems that do not have an exposed convolutional network [8][23][12][4]. The approaches to XFR explored in this paper are most closely related to EBP [39], RISE [23], and methods for network attention for pairwise similarity [34].

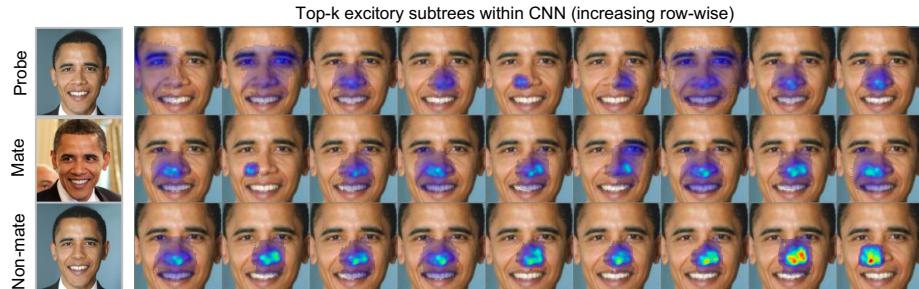
Recently, there has been emerging research on the interpretation of face recognition systems[37][38][6][36][27][9][41]. Visual psychophysics [27] have provided a set of tools for the controlled manipulation of input stimuli and metrics for the output responses evoked in a face matching system. This approach was inspired by Cambridge Face Memory Test [10], which involves progressively perturbing face images using a chosen transformation function (e.g. adding noise) to investigate controlled degradation of matching performance [27]. This approach enables detailed studies of the failure modes of a face matcher or exploring how facial attributes are expressed in a network [9][41]. In contrast, our approach generates controlled degradations using inpainting, to provide localized ground truth for evaluation of network attention models. In [37], the authors propose a novel loss function to encourage part separability during network dissection of parts in a convolutional network for face matching. This approach is primarily concerned with training new networks to maximize interpretability, rather than studying existing networks. In [38], the authors study pairwise matching of faces, to visualize features that lead towards classification decisions. This is similar in spirit to our proposed approach, however we provide a performance metric for evaluating a saliency approach as well as extending visualizations to mated and nonmated triplets. Finally, in [36], the authors visualize the features of shape and texture that underlie subject identity decisions. This approach uses 3D modeling to generate a controlled dataset, rather than inpainting. However, given the authors conclusions that texture has a much larger effect of matching than morphology, having a ground truth dataset that includes texture variation would be an appropriate metric for explainable face recognition.

### 3 Explainable Face Recognition (XFR)

XFR is the problem of explaining why a face matcher matches faces. Fig. 1 shows the structure of this problem. Given a triplet of (*probe*, *mate*, *nonmate*), the XFR algorithm is tasked with generating a saliency map that explains the regions of the probe image that maximize the similarity to the mate and minimize the similarity to the nonmate. This provides an explanation for why the matcher returns this image for the mated identity.

Why triplets? Previous work has shown that pairwise similarity between faces is heavily dominated by the periocular region and nose [6], as confirmed by the qualitative visualization study performed in the supplementary material. The periocular region and nose is almost always used for facial classification, but this level of XFR is not very helpful in explaining finer levels of discrimination. Our goal is to highlight those regions for a probe that are more similar to a presumptive mate and *simultaneously* less similar to a nonmate. This triplet of (*probe*, *mate*, *nonmate*) provides a deeper explanation beyond facial class activation maps for the *relative importance* of facial regions.

In this section, we describe five approaches for network attention in XFR. These approaches are all whitebox methods, which assume access to the underlying convolutional network used for facial matching. The objective of XFR is



**Fig. 2.** Subtree EBP. Given a triplet of (*probe*, *mate*, *nonmate*) subtree EBP explores the activations of individual nodes in a convolutional network that minimizes a triplet loss, which maximizes similarity to the mate and minimizes similarity to the nonmate. The excitatory regions for each node are visualized independently, sorted by loss and combined into a saliency map that best explains how to discriminate the probe.

to generate a non-negative saliency map, that captures the underlying image regions of the probe that are most similar to the mate and least similar to the nonmate. The XFR algorithm can use any property of the convolutional network to generate this saliency map. For our benchmark evalution, we selected three state-of-the-art approaches for network attention (excitation backprop, contrastive excitation backprop and truncated contrastive excitation backprop) following the survey and evaluation results in [6]. In this section, we introduce two new methods to improve upon these published approaches: subtree EBP (Sec. 3.1) and DISE (Sec. 3.2).

**Excitation Backprop (EBP).** Excitation backprop (EBP) [39] models network attention as a probabilistic winner-take-all (WTA) process. EBP calculates the probability of traversing to a given node in the convolutional network, with the probabilities being defined by the positive weights and non-negative activations. The output of EBP is a saliency map that localizes regions in the image that are excitatory for a given class.

In our approach, we replace the cross-entropy loss for EBP with a triplet loss [30]. The original formulation of EBP considers a cross-entropy loss to optimize softmax classification of a set of classes in the training set. In this new formulation, given three embeddings for a mate ( $m$ ), nonmate ( $n$ ) and probe ( $p$ ), the triplet loss function is a max-margin hinge loss

$$L(p, n, m) = \max(0, \|p - m\|^2 - \|p - n\|^2 + \alpha). \quad (1)$$

This uses the squared Euclidean distance between embeddings to capture similarity, such that the loss is minimized when the distance from the probe and mate is small (similarity is high) and the distance form the probe to the nonmate is large (similarity is low), with margin term  $\alpha$ . This loss function extends EBP to cases where a new subject is observed at test time that was not present in the training set, as is commonly the case with face matching systems.

**Contrastive EBP (cEBP).** Contrastive EBP was introduced [39] to handle fine-grained network attention for closely related classes. This approach discards activations common to a pair of classes, to provide network attention specific to one class and not another. In our approach, contrastive EBP [39] is combined with a triplet loss (eq 1).

**Truncated Contrastive EBP (tcEBP).** Truncated contrastive EBP was introduced [6] as an extension of cEBP that considers the contrastive EBP attention map only within the  $k$ th percentile of the EBP saliency map. This addresses an observed instability of cEBP [6] resulting noisy attention maps for faces.

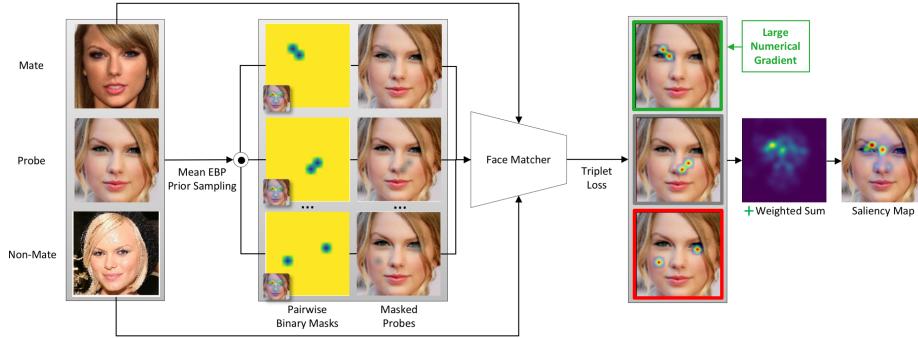
### 3.1 Subtree EBP

In this section, we introduce *Subtree EBP*, a novel method for whitebox XFR. This approach uses the triplet loss function (eq 1), with the following extension. Given a triplet (*probe*, *mate*, *nonmate*) images, we compute the gradient of the triplet loss function ( $\frac{\partial L}{\partial x_i}$ ) with respect to every node  $x_i$  in the network. This approach uses the standard triplet-based learning, where the mate and nonmate embeddings are assumed constant and the gradient is computed relative to the probe image. Next, we sort the gradients at every node  $x_j$  in decreasing order, and select the top- $k$  nodes with the largest positive gradients. These are the top- $k$  nodes in the network that most affect the triplet loss, to increase the similarity to the mate and simultaneously decrease the similarity to the nonmate. Finally, we construct  $k$  EBP saliency maps  $S_i$  starting from each of the selected interior nodes, then the  $S_i$  are combined in a weighted convex combination with weights  $w_i = \frac{\partial L}{\partial x_i}$  and

$$S = \frac{1}{\sum_j w_j} \sum_i \frac{\partial L}{\partial x_i} S_i \quad (2)$$

where the weights are given by the loss gradient ( $w_i$ ), normalized to sum to one. This forms the final subtree EBP saliency map  $S$ .

Fig. 2 shows an example of the subtree EBP method. This montage shows the top 27 nodes with the largest triplet loss gradient for the shown triplet. Each node results in a saliency map corresponding to the excitatory subtree rooted at this node. The weight of the saliency map is proportional to the gradient sorted rowwise, so that the nodes in the bottom right affect the loss more than the nodes in the upper left. Each of these saliency maps are combined into a convex combination (eq. 2) forming the final network attention map. In this example, the nonmate was synthesized to differ with the mate only in the nose region, and our method is able to correctly localize this region. The supplementary material shows a more detailed example of this selection process starting from the largest excitation node at each layer of a ResNet-101 network. This result shows that nodes selected close to the image will be well localized, nodes in the middle of the network correspond to parts and nodes selected close to the embedding correspond to the whole nose and eyes of the face.



**Fig. 3.** Density-based Input Sampling for Explanation (DISE). Our approach is an extension of RISE [23] for XFR. This approach occludes small regions in the probe image with grey (i.e. masked pixels), sampled according to a prior density derived from excitation backprop, and computes a numerical gradient for the triplet loss for (*probe*, *mate*, *nonmate*) given these masked probes. Masks with a large numerical gradient are more heavily weighted in the accumulated saliency map.

### 3.2 Density-based Input Sampling for Explanation (DISE)

Density-based Input Sampling for Explanation (DISE) is a second novel approach for whitebox XFR introduced in this paper. DISE is an extension of Randomized Input Sampling for Explanation (RISE) [23] using a prior density to aid in sampling. Previous work [23][12] has constructed a saliency map associated with a particular class by randomly perturbing the input image by masking selected pixels, evaluating it using a blackbox system, and accumulating those perturbations based on how confident the system is that the modified input image corresponds to the target class. However, these approaches generate masks to occlude the input image uniformly at random. This sampling process is inefficient, and can be improved by introducing a prior distribution to guide the sampling. In this section, we describe the extension to RISE [23] where the prior density for input sampling is derived from a whitebox EBP with triplet loss.

Fig. 3 shows an overview of this approach. Our approach extends RISE [23] for XFR as follows:

1. Using a non-uniform prior for generating the random binary masks
2. Restricting the masks to use a sparse, fixed number of mask elements
3. Defining a numerical gradient of the triplet loss to weight each mask

**Non-Uniform Prior.** Prior research on discriminative features for facial recognition showed that the most important regions of the face were generally located in and around the eyes and nose (Sec. 3). Fig. 3 shows an example of this saliency map computed for a probe image of Taylor Swift using the VGG-16 [22] network as the whitebox face matcher. Using this saliency map as our prior probability for generating random masks allows us to sample the space of most salient

masks that will affect the loss more efficiently than assuming a uniform probability across the entire image. Further limiting this prior to the upper 50th percentile of the mean EBP effectively eliminates the possibility of masking out unimportant background elements.

**Sparse Masks.** Next, we restrict the number of masked elements to be sparse. RISE considered random binary masks covering the entire input image. In contrast, we use a sparse mask to highlight the effect of a small localized region of the face on the loss. We used two mask elements per mask, upsampled by a factor of 12 (to avoid pixel level adversarial effects). We found that filling the masks with a blurred version of the image performed quantitatively better on the inpainting game than using grey masks.

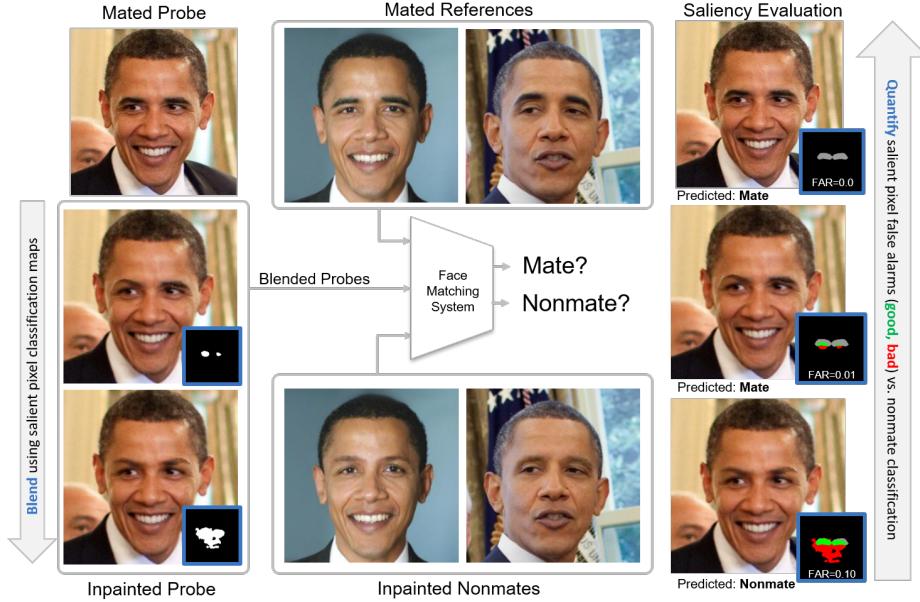
**Numerical gradient.** Finally, given the probe image which has been masked with the sparse mask sampled from the non-uniform prior, we can compute a numerical gradient of the triplet loss. Let  $p$  be an embedding of the probe,  $m$  the mated image embedding,  $n$  the nonmated image embedding, and  $\hat{p}$  the masked probe embedding. Then, the numerical gradient of the triplet loss (eq 1) can be approximated as:

$$\frac{\partial L_{dise}}{\partial p} \approx \max(0, L(p, m, n) - L(\hat{p}, m, n)) \quad (3)$$

The numerical gradient is an approximation to the true loss gradient computed by perturbing the input by occluding the probe with a pixel mask, and computing the corresponding change in the triplet loss. In other words, when the probe masks out a region that increases for the similarity between the probe and mate and decreases for the probe and nonmate, the numerical gradient should be large. This allows for a loss weighted accumulation of these masks into a saliency map. The final saliency is accumulated following (eq. 2), where saliency maps  $S_i$  are the pairwise binary masks, with non-negative gradient weights (eq. 3).

## 4 Experimental Protocol

Recent explainable AI research has focused on class activation maps [23][12][31] [33][42][39][4], which visualize salient regions used for classification. For facial recognition, prior work has shown this is almost always the eyes, nose, and upper lip of the face [6]. In facial identification, a probe image is given to a face matching system, which returns the top  $K$  identities from a gallery. A natural question is why the matching system picked the top match instead of the second top match (or remaining top K matches). One way to give an answer to this question is to highlight the region(s) that match a given identity more than the second identity or other identities. This saliency map should be larger for the regions that contribute the most to the identity and not others. In this paper,



**Fig. 4.** Inpainting game overview. The XFR algorithm is given triplets (probe, mate, nonmate) labeled in the figure as (mated probe, mated references, inpainted nonmates), and is tasked with estimating a discriminative saliency map that estimates the likelihood that a pixel belongs to a region that is discriminative for the mate. A threshold is applied to the saliency map to classify each pixel as being discriminatively salient (inset, blue squares, left). A high performing XFR algorithm will correctly classify the discriminative pixels within the inpainted region (green, right) while avoiding classifying the pixels that are identical between the mated references and the probes as being discriminative (red, right). See sec. 4.1 for more details.

our goal is to highlight the regions that are responsible for matching a given image to one identity versus a similar identity.

A key challenge for evaluating the performance of an XFR approach is generating ground truth. For XFR, ground truth not only depends on the selection of probes, mates, and nonmates, but can also depend on a target network for evaluation. We address this issue by synthesizing inpainted nonmates or *doppelgangers*, where a select region of the face is changed from the original identity. Only the inpainted region differs between the two images and therefore only the inpainted region can be used to discriminate between them. Furthermore, we synthesize doppelgangers based on their ability to reduce the match score for a target network. We call our overall strategy for quantitative evaluation the *inpainting game*.

#### 4.1 The Inpainting Game

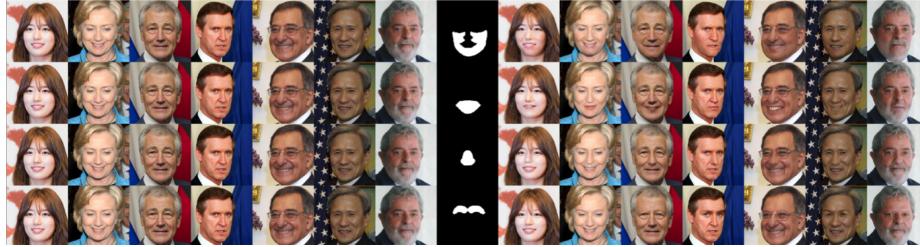
An overview of the inpainting game evaluation is given in Fig. 4. The inpainting game uses four (or more) images for each evaluation: a probe image, mate image(s), an inpainted probe and inpainted nonmate(s). The inpainted probe or *probe doppelganger* is subtly different from the probe in a fixed region of the face, such as the eyes, nose or mouth. Similarly, the inpainted nonmate or *mate doppelganger* is subtly different from the mate image, such that the doppelgangers are a different identity. The inpainted probe and inpainted nonmate are constrained to be the same new identity. Sec. 4.2 discusses the construction of this dataset.

The XFR algorithm is given triplets of probes, mates and nonmates, labeled in Fig. 4 as (“mated probe”, “mated references” and “inpainted non-mates”). For each triplet, the XFR algorithm is tasked with estimating the likelihood that each pixel belongs to a region that is discriminative for matching the probe to the mated identity over the nonmated/inpainted identity. These discriminative pixel estimations form a saliency map. Each pixel is classified as being discriminatively salient by applying a threshold, which forms a binary saliency map. For each binary saliency map, pixels in the probe are replaced with the pixels from an inpainted probe forming a blended probe. The inpainted probe is generating by inpainting the same facial region as the inpainted nonmates and is not provided to the XFR algorithm, which is sequestered and used for evaluation only. The saliency map is evaluated by how quickly it can flip the identity of the blended probe from the mate to the non-mate, while maximizing saliency (green) in ground truth (grey) while minimizing false alarms (red). See Sec. 4.3 for additional details, including the metrics for the inpainting game.

#### 4.2 Inpainting Dataset for Facial Recognition

The inpainting dataset for face recognition is based on the images from the IJB-C dataset [20]. The inpainting dataset contains 561 images of 95 subjects selected from IJB-C, for an average of 5.9 images per subject. We defined eight facial regions for evaluation: 1) cheeks and jaw, 2) mouth, 3) nose, 4) left eye, 5) right eye, 6) eyebrows, and 7) left face, 8) right face. Each image is inpainted for each of the eight regions forming a total of 4488 inpainted doppelgangers. From this set, we define 3648 triplets, such that each triplet is a combination of (probe, mate and inpainted doppelganger nonmate). The XFR algorithms should not be evaluated on triplets for networks that cannot distinguish the original and inpainted identities. Hence, the only the triplets that contain discriminable identities are included for the network the algorithm is being evaluated on.

The inpainted doppelgangers are generated as follows. In order to systematically mask the regions, we use the pix2face algorithm [7] to fit a 3d face mesh onto each facial image. We then projected the facial region masks onto the images. We use pluralistic inpainting [40] to synthesize an image completion in that masked region. Fig. 5 shown examples of these inpainted doppelgangers.

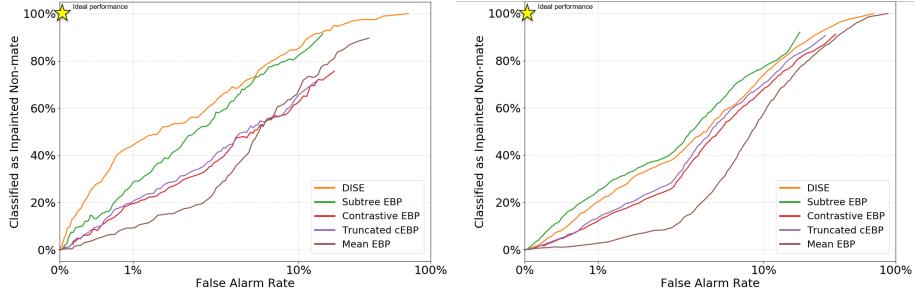


**Fig. 5.** Example facial inpainted images. This montage shows the first four of eight inpainting regions (cheeks, mouth, nose, eyebrows, left face, right face, left eye, right eye), and synthesized inpainted doppelgangers images. The first seven columns show seven subjects, with the same image repeated along the column. The middle column shows a binary inpainting mask that defines the inpainting region. The last seven columns show the inpainted doppelganger image using the mask region for that row, such that the inpainted image differs from the original image only in the mask region. Observe that the identity may change subtly while looking down a column.

A key challenge of constructing the inpainting dataset is to enforce that the inpainted nonmate is in fact a different identity. Most of our inpainted images are not sufficiently different in similarity from the original mated identity for a specific network. A given triplet of (*probe*, *mate*, and *inpainted nonmate*) is only included in the dataset if a given target network can distinguish the two identities for the mate/mate doppelganger and the probe/probe doppelganger. They are required to be able to distinguish these identities both using a nearest match protocol and an verification protocol, such that the verification match threshold for a target network is calibrated to a false alarm rate of 1e-4. Specifically, each triplet has to fulfill the following criteria in order to be included in the dataset for a given network:

1. The original probe must be: (i) more similar to the original/mated identity than the corresponding inpainted/nonmated identity and (ii) correctly verified as the original/mated identity at the calibrated verification threshold.
2. The inpainted probe must be: (i) more similar to the corresponding inpainted / nonmated identity than the original identity and (ii) correctly verified as the same identity as the inpainted/nonmated identity at the calibrated verification threshold.

The inpainting dataset is filtered for each target network according to the above criteria, resulting in a dataset specific to that target network. For example, for the ResNet-101 based system, the final filtered dataset includes 84 identities and 543 triplets, filtered down from 95 identities and 3648 triplets. Lower performing networks will generally have fewer triplets satisfying the selection criteria than higher performing networks, because they will not be able to discriminate as many of the subtle changes in the inpainted probes.



**Fig. 6.** (left) Inpainting game analysis using VGGFace2 ResNet-50. (right) Inpainting game analysis using Light-CNN. Refer to Table 1 for a summary of performance at fixed operating points on these curves.

### 4.3 Evaluation Metrics

The XFR algorithms estimate the likelihood that each pixel belongs to a region that is discriminative for matching the probe to the mated identity over the non-mated/inpainted identity. These discriminative pixel estimations form a saliency map, where the brightest pixels are estimated to be most likely to belong to the discriminative region. Fig. 4 shows an example and saliency predictions at two thresholds, where the saliency prediction is shown at different thresholds as a binary mask.

In order to motivate our proposed metric, let's first consider using a classic receiver operating characteristic (ROC) curve for evaluation of the inpainting game rather than our proposed metric. A ROC curve can be generated by sweeping a threshold for the pixel saliency estimations, and computing true accept rate and false alarm rate by using the inpainted region as the positive / salient region and the non-inpainted region as the negative / non-salient region (i.e. middle column in Fig. 5). However, not all pixels within the inpainted region contribute equally to the identity, and the saliency algorithm should not be either penalized or credited with this selection.

To address this key challenge, we use mean nonmate classification rate instead of true positive rate for saliency classification. We play a game where the pixels classified as being salient by sweeping the saliency threshold are replaced with the pixels from the “inpainted probe”, which is not provided to the saliency algorithm. These “blended probes” can then be classified as original identity or inpainted nonmate identity by the network being tested. High performing XFR algorithms will correctly assign more saliency for the inpainted regions that will change the identity of the blended probes without increasing the false alarm rate of the pixel salience classification. This is the key idea behind our evaluation metric. The false positive rate is calculated from salient pixel classification across all triplets, using the ground truth masks for the blended probe. The mean nonmate classification rate is weighted by the number of triplets within each

		Inpainting Game (Mean Nonmate Classification Rate)							
System	Saliency	All		Nose		Eyes		Eyebrows	
		FAR=1E-2	FAR=5E-2	FAR=1E-2	FAR=5E-2	FAR=1E-2	FAR=5E-2	FAR=1E-2	FAR=5E-2
ResNet-101	DISE	<b>0.438</b>	<b>0.816</b>	0.691	0.932	<b>0.627</b>	<b>0.931</b>	<b>0.723</b>	<b>0.981</b>
	Subtree EBP	0.274	0.792	<b>0.740</b>	<b>0.973</b>	0.503	<b>0.931</b>	0.065	0.942
	Mean EBP	0.143	0.626	0.208	0.904	0.565	<b>0.931</b>	0.010	0.781
	Contrastive EBP	0.132	0.454	0.178	0.589	0.310	0.517	0.040	0.543
	Truncated cEBP	0.167	0.582	0.247	0.699	0.276	0.573	0.066	0.642
VGGFace2 ResNet-50	DISE	<b>0.443</b>	<b>0.761</b>	<b>0.730</b>	0.902	<b>0.609</b>	<b>0.957</b>	<b>0.891</b>	<b>0.976</b>
	Subtree EBP	0.285	0.735	0.705	<b>0.984</b>	0.418	0.878	0.332	0.928
	Mean EBP	0.092	0.499	0.148	0.705	0.388	0.831	0.108	0.821
	Contrastive EBP	0.195	0.520	0.343	0.705	0.484	0.652	0.323	0.850
	Truncated cEBP	0.205	0.536	0.361	0.754	0.539	0.696	0.329	0.868
LightCNN NiN+ MFM-28	DISE	0.202	0.587	<b>0.729</b>	<b>0.961</b>	<b>0.409</b>	<b>0.847</b>	<b>0.641</b>	<b>0.927</b>
	Subtree EBP	<b>0.250</b>	<b>0.643</b>	0.699	0.914	0.294	0.778	0.489	0.921
	Mean EBP	0.027	0.307	0.048	0.477	0.085	0.600	0.026	0.558
	Contrastive EBP	0.121	0.526	0.286	0.821	0.211	0.533	0.287	0.842
	Truncated cEBP	0.135	0.557	0.294	0.820	0.209	0.576	0.298	0.892

**Table 1.** Inpainting game evaluation results. This table summarizes the performance at two operating points of false alarm rate (1E-2, 5E-2) for the performance curves in Fig. 6 (ResNet-50 and LightCNN) and in the supplementary material. Mean nonmate classification rate is the proportion of triplets where the identity of the blended image was correctly “flipped” to the doppelganger. Results show that our new methods (DISE, Subtree EBP) outperform the state of the art by a wide margin on three matchers. Detailed subprotocol results and curves are provided in the supplemental material.

facial region for a filtered dataset, to avoid bias for subprotocols with more examples. Example of the output curves for this metric is shown in Fig. 6.

## 5 Experimental Results

### 5.1 Inpainting Game Quantitative Evaluation

We ran the inpainting game evaluation protocol on the inpainting dataset using three target networks: LightCNN [35], VGGFace2 ResNet-50 [5] and a custom trained ResNet-101. We considered the five XFR algorithms described in Sec. 3 forming the benchmark for XFR evaluation.

The evaluation results are summarized in Table 1 and plotted in Fig. 6 and in the supplementary material. The summary table shows for each combination of network and XFR algorithm, at two false alarm rates (1E-2, 5E-2) for the full protocol and three subprotocols: eyes, nose and eyebrows only. Additional results in the supplementary material show the results for the individual facial region subprotocols.

Overall, results show that for deeper networks (ResNet-101, ResNet-50), the top performing XFR algorithm is DISE. However, for shallower networks (LightCNN) then top performing algorithm is Subtree EBP. Both of these new approaches outperform the state of the art (EBP, cEBP, tcEBP) by a wide margin. We believe that DISE outperforms Subtree EBP since subtree EBP cannot

localize image regions any better than the underlying network represents faces. For example, consider the eyebrows subprotocol result in the supplementary material, which shows that subtree EBP cannot represent eyebrows independently from the eyes. DISE can mask image regions independently from the underlying target network and correctly localize eyebrow effects.

## 6 Conclusions

In this paper, we introduced the first comprehensive benchmark for XFR. We motivated the need for XFR and describe a new quantitative method for comparing XFR algorithms using the inpainting game. The results show that the DISE and subtree EBP methods provide a significant performance improvement over the state of the art, which provides a new baseline for visualizing discriminative features for face recognition. This evaluation protocol provides a means to compare different approaches to network saliency, and we believe this form of quantitative evaluation will help encourage research in this emerging area of explainable AI for face recognition. All software and datasets for reproducible research are available for download at <http://stresearch.github.io/xfr>.

**Acknowledgement.** This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under contract number 2019-19022600003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

## References

1. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Muller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015. [3](#)
2. D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, 2017. [1](#), [2](#)
3. J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. [1](#)
4. C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, and Others. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015. [3](#), [8](#)

5. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. [3](#), [13](#), [18](#), [33](#)
6. G. Castanon and J. Byrne. Visualizing and quantifying discriminative features for face recognition. In *International Conference on Automatic Face and Gesture Recognition*, 2018. [2](#), [3](#), [4](#), [5](#), [6](#), [8](#)
7. D. Crispell and M. Bazik. Pix2face: Direct 3d face model estimation. In *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 2512–2518, Oct. 2017. [10](#)
8. P. Dabkowski and Y. Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017. [3](#)
9. P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and R. Chellappa. How are attributes expressed in face dcnn? *ArXiv*, abs/1910.05657, 2019. [4](#)
10. B. Duchaine and K. Nakayama. The cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44:576–585, 2006. [4](#)
11. Facial Identification Scientific Working Group. FISWG guidelines for facial comparison methods:. In *FISWG standards version 1.0 - 2012-02-02*, 2012. [2](#)
12. R. Fong and A. Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *arXiv preprint arXiv*, 2017. [2](#), [3](#), [7](#), [8](#)
13. C. Garvie, A. Bedoya, and J. Frankle. The perpetual line-up: Unregulated police face recognition in america. In *Technical report, Georgetown University Law School*, 2018. [1](#)
14. C. Grimm, D. Arumugam, S. Karamcheti, D. Abel, L. L. Wong, and M. L. Littman. Latent attention networks. In *arXiv:1706.00536v1*, 2017. [3](#)
15. P. Grother, M. Ngan, and K. Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. In *NISTIR 8280*, 2019. [1](#)
16. R. Hu, J. Andreas, T. Darrell, and K. Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, 2018. [2](#)
17. P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*, 2017. [3](#)
18. H. Li, K. Mueller, and X. Chen. Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation. *Image Vision Comput.*, 83-84:70–86, 2017. [3](#)
19. A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. [3](#)
20. B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018. [10](#)
21. A. M. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *NIPS*, 2016. [2](#)
22. O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. [7](#)
23. V. Petruik, A. Das, and K. Saenko. Rise: Randomized input sampling for explanation of black-box models. *British Machine Vision Conference (BMVC)*, 2018. [2](#), [3](#), [7](#), [8](#)

24. P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J.-C. Chen, C. P. del Castillo, R. Chellappa, D. White, and A. J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. In *Proceedings of the National Academy of Sciences of the United States of America*, 2018. [2](#)
25. I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AIES '19*, 2019. [1](#)
26. M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *KDD '16*, 2016. [1](#)
27. B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *European Conference on Computer Vision (ECCV)*, 2018. [4](#)
28. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2660–2673, 2015. [3](#)
29. W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019. [1](#)
30. F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. [5](#)
31. 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2016. [1, 3, 8](#)
32. K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013. [2](#)
33. K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Iclr*, pages 1—, 2014. [1, 3, 8](#)
34. A. Stylianou, R. Souvenir, and R. Pless. Visualizing deep similarity networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2029–2037. IEEE, 2019. [3](#)
35. X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. [3, 13, 18, 26, 27, 28, 29, 30, 31](#)
36. T. Xu, J. Zhan, O. G. B. Garrod, P. H. S. Torr, S.-C. Zhu, R. A. A. Ince, and P. G. Schyns. Deeper interpretability of deep networks. *ArXiv*, abs/1811.07807, 2018. [4](#)
37. B. Yin, L. Tran, H. Li, X. Shen, and X. Liu. Towards interpretable face recognition. In *In Proceeding of International Conference on Computer Vision*, Seoul, South Korea, October 2019. [4](#)
38. T. Zee, G. Gali, and I. Nwogu. Enhancing human face recognition with an interpretable neural network. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. [4](#)
39. J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation Backprop. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9908 LNCS:543–559, 2016. [2, 3, 5, 6, 8](#)
40. C. Zheng, T.-J. Cham, and J. Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. [10](#)

41. Y. Zhong and W. Deng. Exploring features and attributes in deep face recognition using visualization techniques. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019. 4
42. B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016. 1, 3, 8



**Fig. S1.** Qualitative visualization study. This figure shows the XFR saliency maps generated using the LightCNN Subtree EBP method for 16 probes (columns) of 7 subjects (rows), each with 16 mates (not shown) and a common set of 8000 nonmates (not shown) all sampled from VGGFace2 [5]. Results show that the discriminative features used to distinguish a subject from the entire nonmate population are inconsistent, but are primarily the nose and mouth for frontal probes, including the eyes for non-frontal probes. See supplemental Fig. S15 for additional examples.

## A Supplementary Material

### A.1 Qualitative Visualization Study

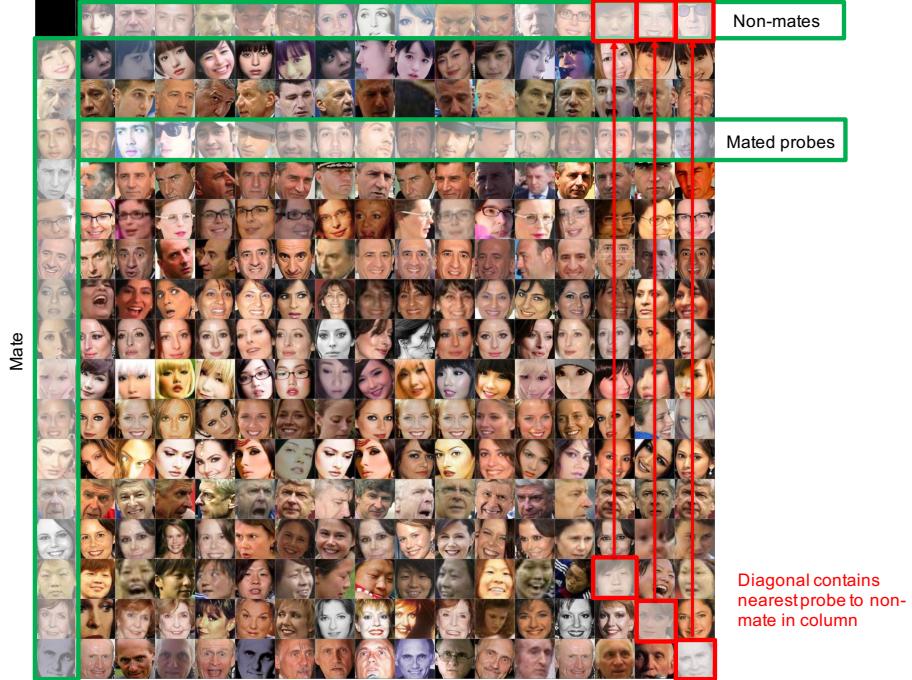
The inpainting game provides a quantitative comparison of XFR algorithms, however it does not provide insight as to how useful these XFR algorithms are on novel face images. In this section, we provide a qualitative study of XFR algorithms visualized on a standard set of triplets. We consider two target networks: ResNet-101 and Light-CNN [35], and provide visualizations for the whitebox methods referenced in the main submission. This analysis includes the following figures showing qualitative visualization results for combinations of (target network, XFR method): (ResNet-101, EBP, Fig. S3), (ResNet-101, cEBP, Fig. S4), (ResNet-101, tcEBP, Fig. S5), (ResNet-101, Subtree, Fig. S6), (Light-CNN, EBP, Fig. S8), (Light-CNN, cEBP, Fig. S9), (Light-CNN, tcEBP, Fig. S10), (Light-CNN, Subtree EBP, Fig. S11). Finally, we show results for the Light-CNN using only single probes (Fig. S12), or repeated probes (Fig. (S13) to highlight the effect of non-mates in the triplet visualization.

From this visualization study, we draw the following conclusions:

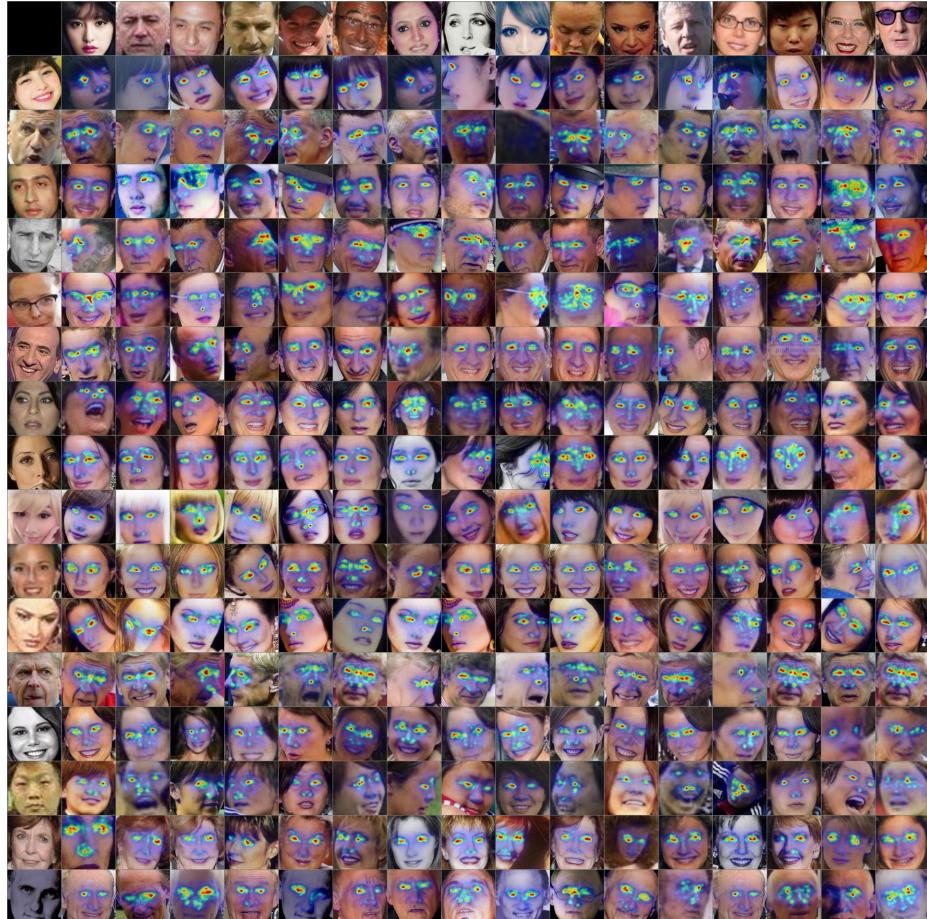
1. **Non-localized.** Unlike facial examiners which leverage the complete FISWG standards for facial comparison, there is no evidence that modern face matchers leverage localized discriminating features such as scars, marks and blemishes. All visualizations are centered on the facial interior, and almost no activation is on the shape of the head. Also, the systems tend to overgen-

eralize to represent all faces in a standard manner using the eyes and nose, brow and mouth, ignoring localized features such as moles or facial markings.

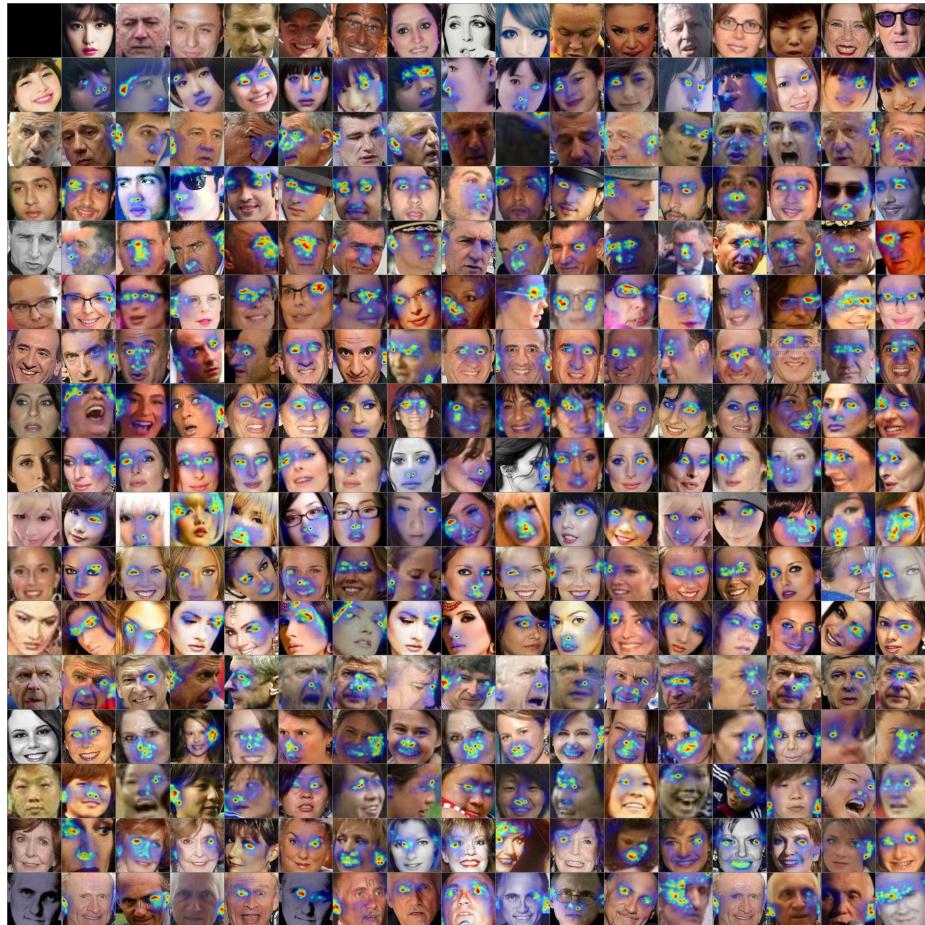
2. **Pose variant.** The target networks tested are not truly pose invariant. When considering different probes of the same subject, where the probe differs in pose, the whitebox systems can generate different visualizations. This suggests that the underlying network is still pose variant.
3. **Triplet specific.** The features that are used for recognition depend on the selection of the triplet, notably the selection of the non-mate for comparison. The visualized features are more consistent when considering a larger non-mate set (Fig. S1).
4. **Network specific.** The visualized features are dependent on the selected target network for visualization. A higher performing network (light-CNN) tends to use more facial features of the brow and mouth in addition to the eyes and nose, than a lower performing network (ResNet-50). No networks yet tested use the hair or chin.



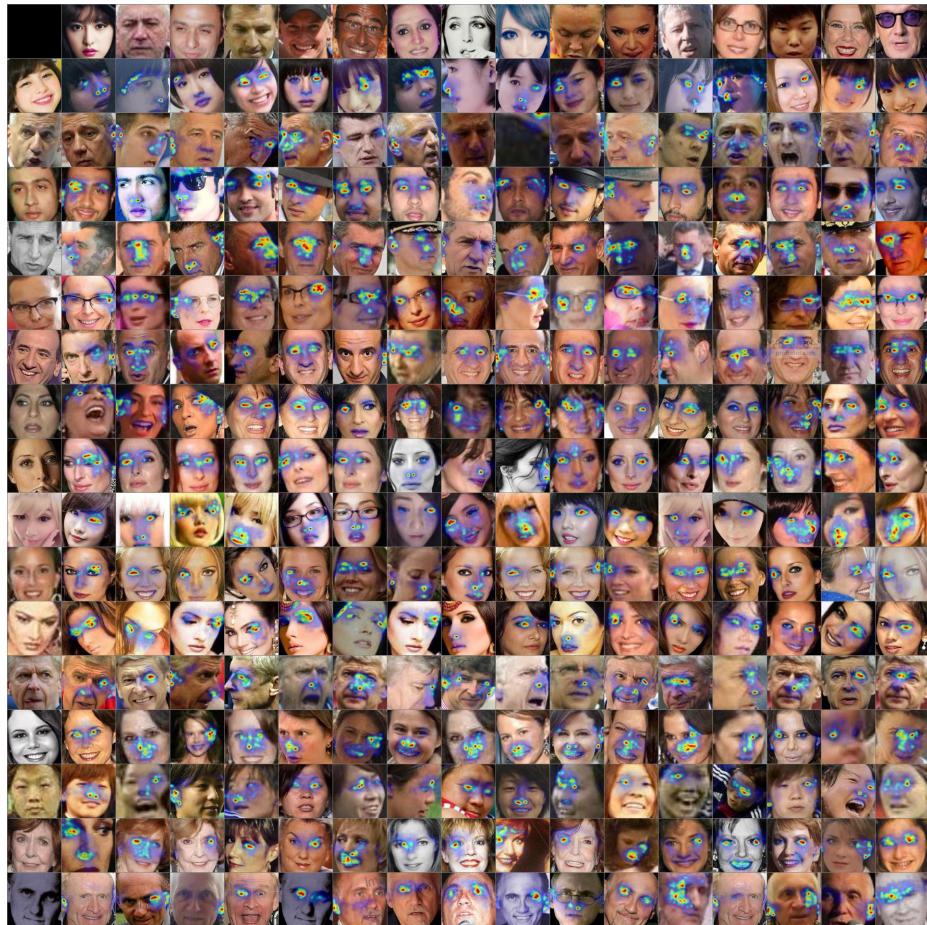
**Fig. S2.** Whitebox visualization overview. This montage shows a set of 16 randomly selected subjects from IJB-C, such that every row has the same identity. Images  $(i, j)$  in this montage define a triplet  $(m_i, p_{ij}, n_j)$  for probe  $p_{ij}$ , mate  $m_i$  in the first entry of column  $i$  and non-mate  $n_j$  in the first entry in row  $j$ . Non-mates are ordered such that on the diagonal are the nearest non-mated subject in IJB-C. In other words, for triplet  $(m_i, p_{ii}, n_i)$ , non-mate  $n_i$  is more similar to  $m_i$  than any other nonmate  $n_j$ , using a ResNet-101 matching system. This montage is used to visualize how the whitebox saliency map changes when considering different triplets.



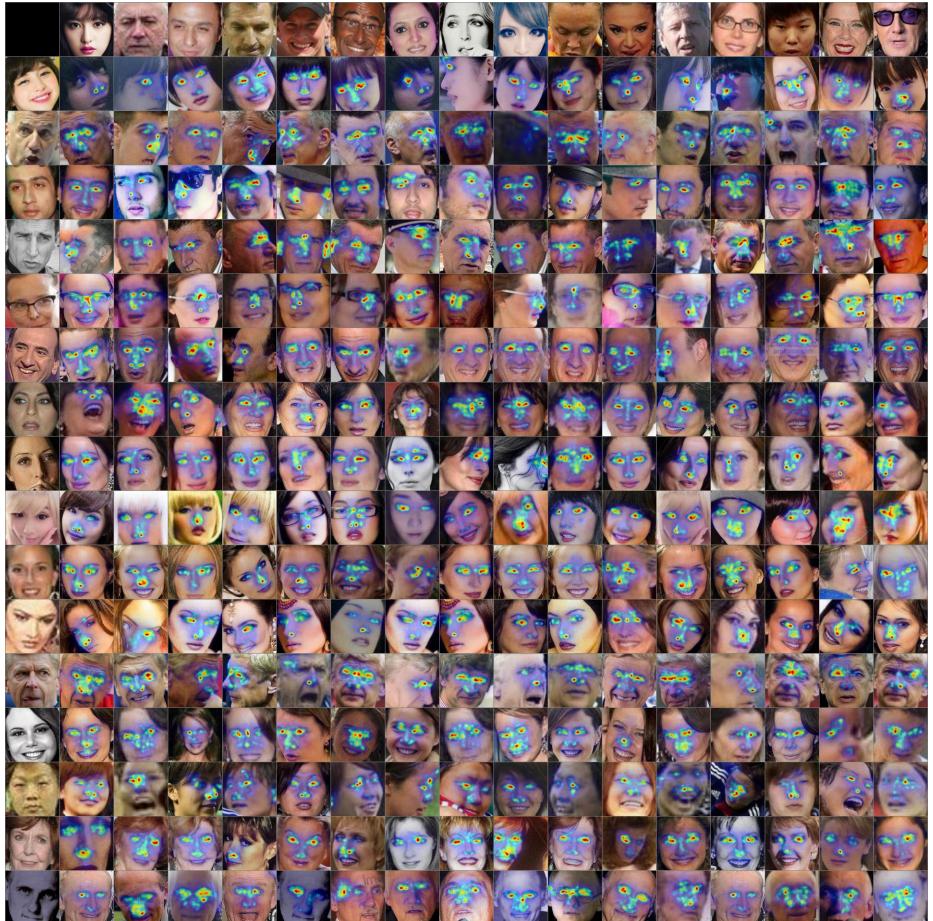
**Fig. S3.** EBP (ResNet-101). This montage is the same images as in Fig. S2, but with a whitebox saliency map derived from excitation backprop for a whitebox ResNet-101 system. Observe that EBP always selects the eyes and nose no matter what non-mated subject is being considered. This does not provide subtle distinctions between the regions that are discriminative for a mate vs. a non-mate, but it does provide a visualization of the regions of the probe that are used for classification. This visualization should be compared with Fig. S8 for the same subjects and whitebox method, but a different underlying trained network (light-cnn).



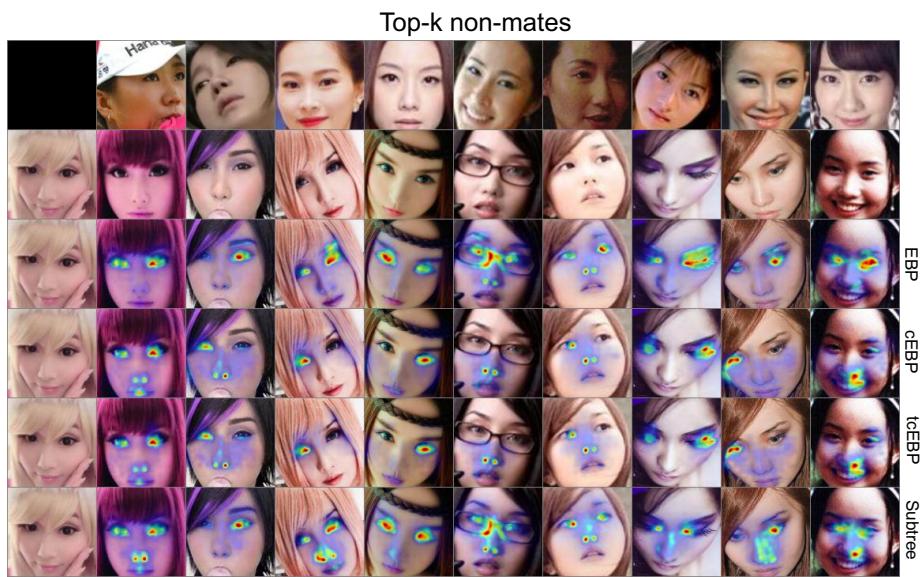
**Fig. S4.** Contrastive triplet EBP (ResNet-101). This montage shows the contrastive EBP for a ResNet-101 whitebox. Observe that this saliency map is unstable, and at times generates saliency maps on the background of the image (e.g. probe (16,15)). This is a known challenge of contrastive EBP, which led towards the development of truncated contrastive EBP.



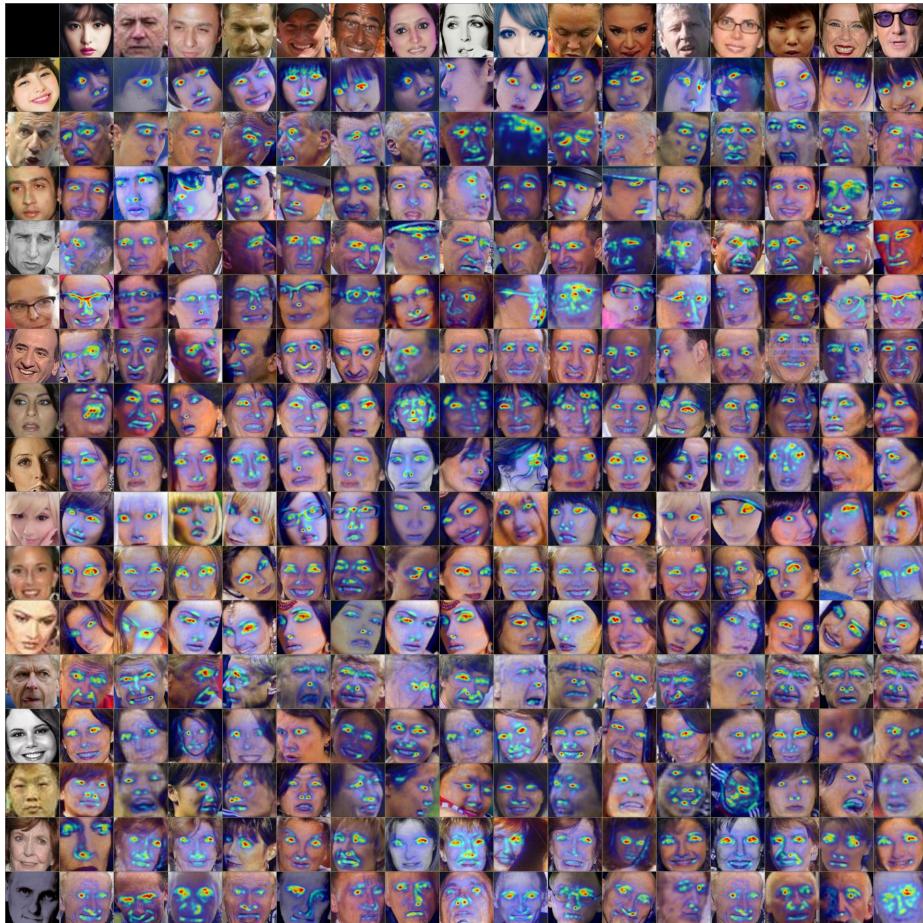
**Fig. S5.** Truncated contrastive triplet EBP (ResNet-101). This montage shows truncated contrastive triplet EBP.



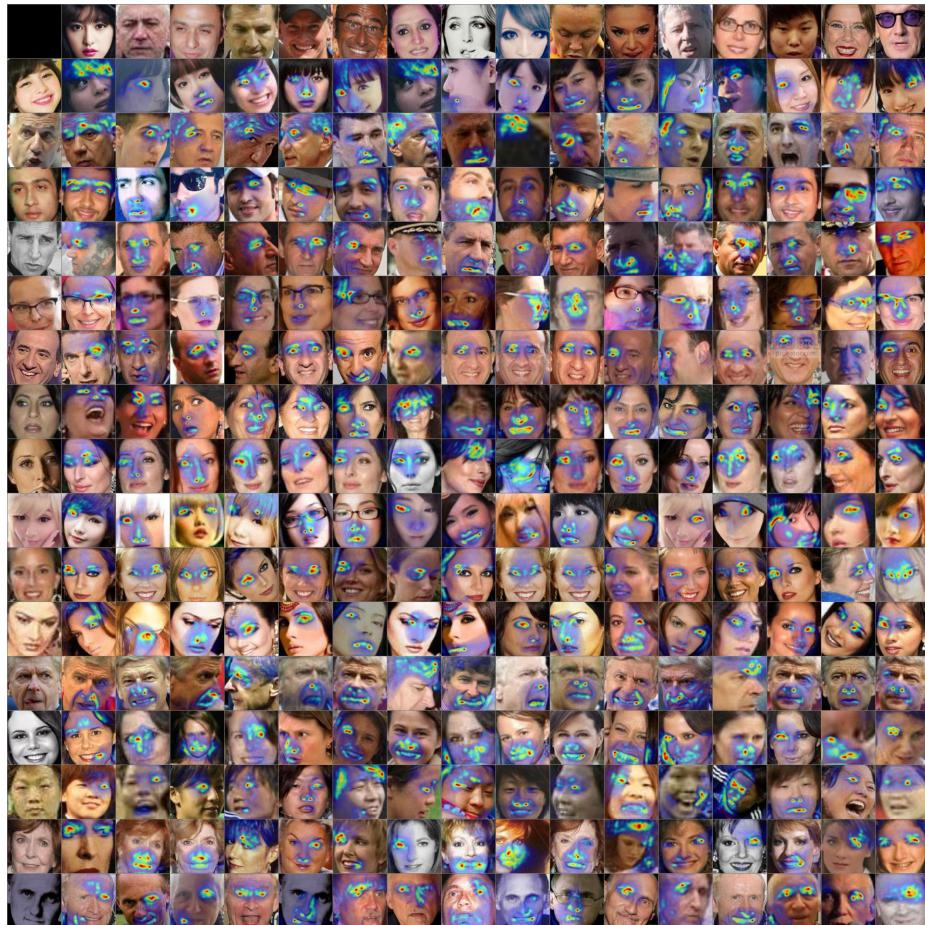
**Fig. S6.** Subtree Triplet EBP (ResNet-101). This montage shows subtree triplet EBP.



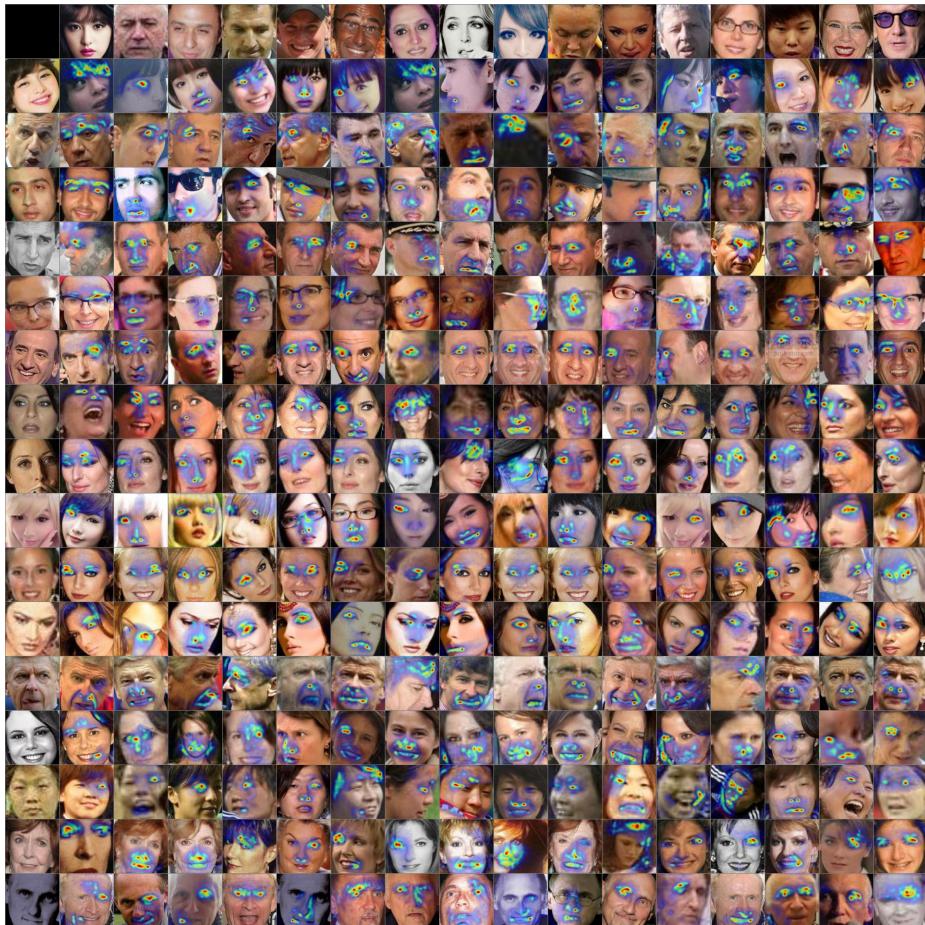
**Fig. S7.** Single probe montage (ResNet-101). This montage compares four white box methods on a common set of probes, such that the non-mates are now ordered in decreasing similarity with the mate. This shows how the different methods compare for real-world doppelgangers.



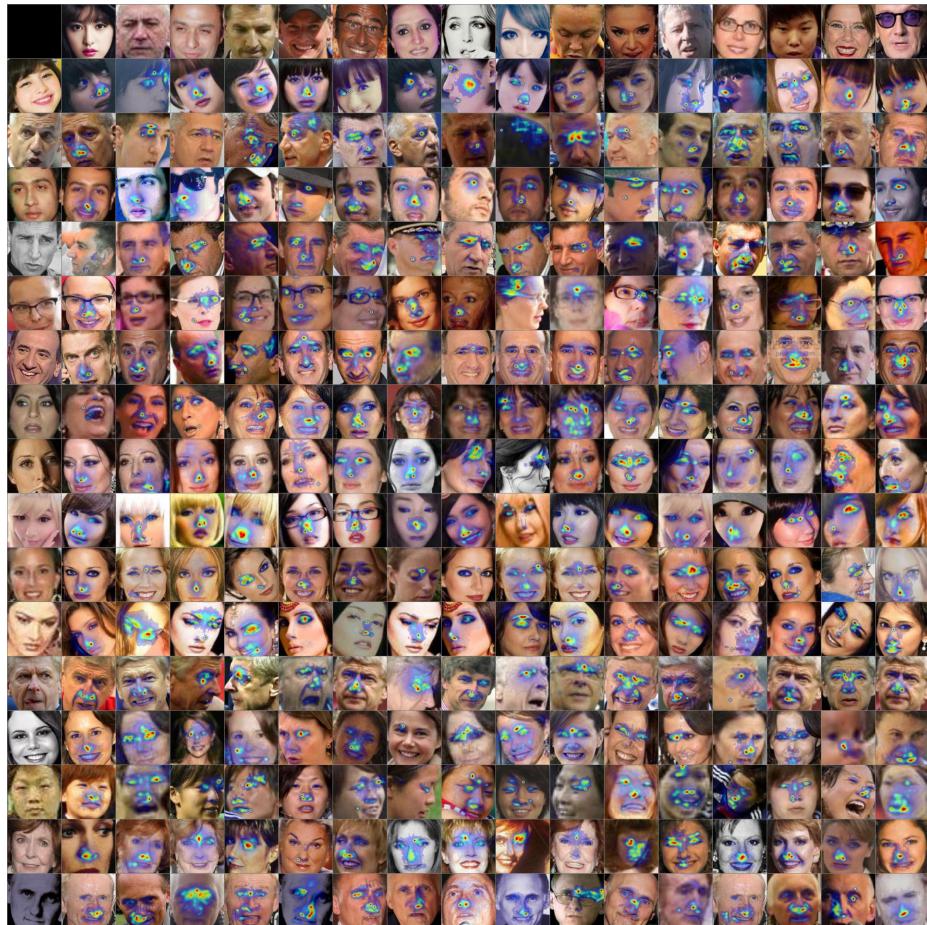
**Fig. S8.** EBP (Light-CNN [35]). This montage generates the EBP saliency map for the Light-CNN network. This should be compared with Fig. S3, which shows that this network exhibits more saliency around the mouth and brow than the ResNet-101 network.



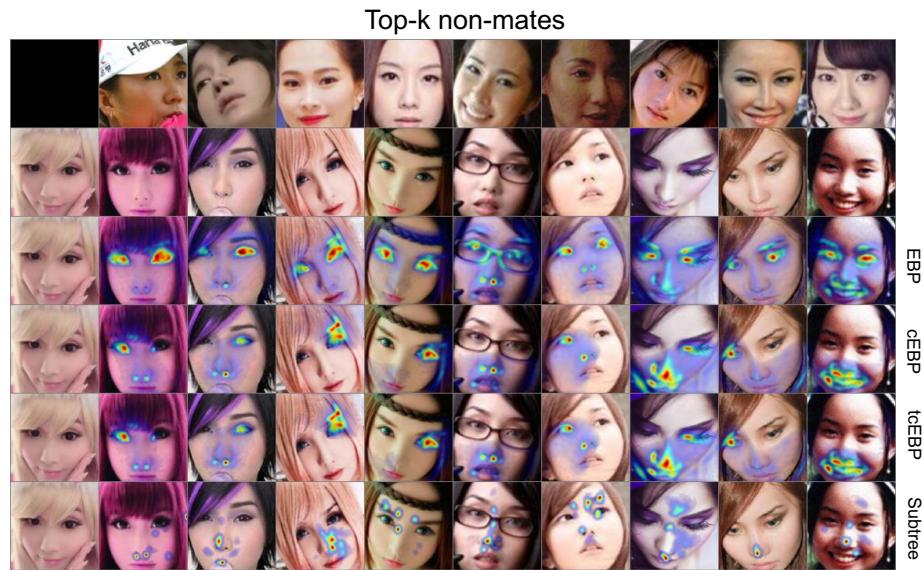
**Fig. S9.** Contrastive triplet EBP (Light-CNN [35]). This montage should be compared with Fig. S4 to show the differences for contrastive triplet EBP comparing ResNet-101 with light-CNN.



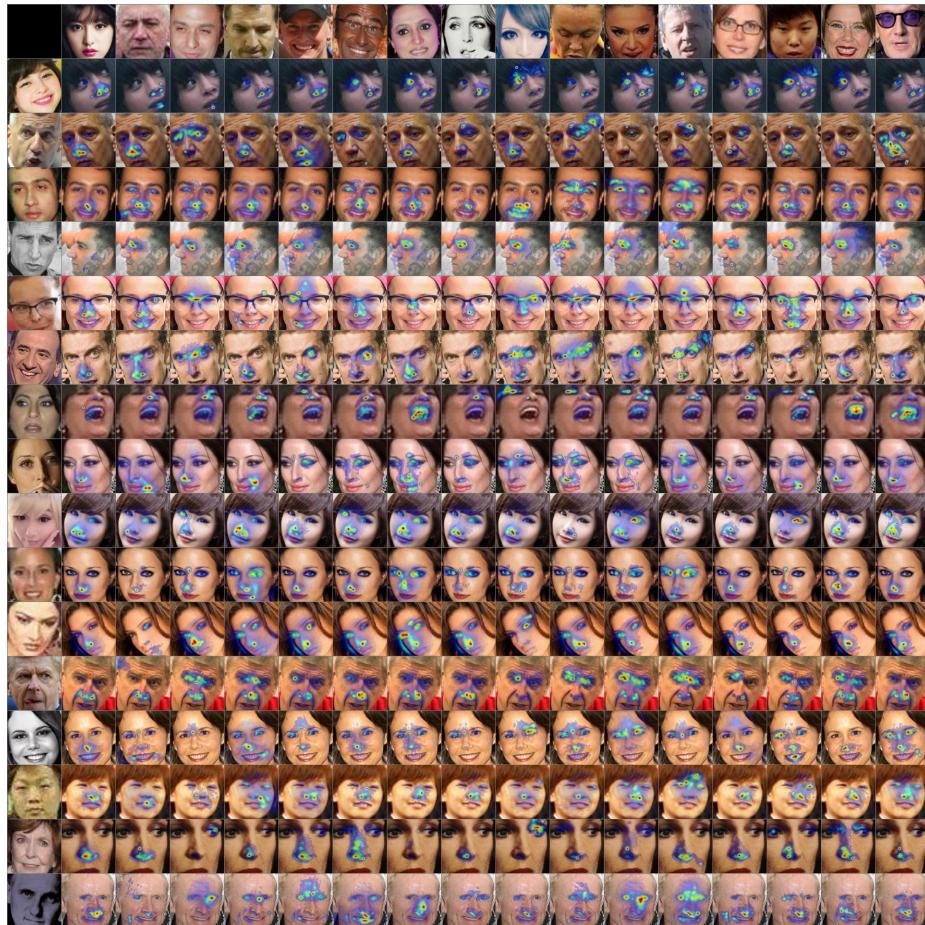
**Fig. S10.** Truncated contrastive triplet EBP (Light-CNN [35]). This montage should be compared with Fig. S5 to show the differences for tcEBP comparing ResNet-101 with light-CNN.



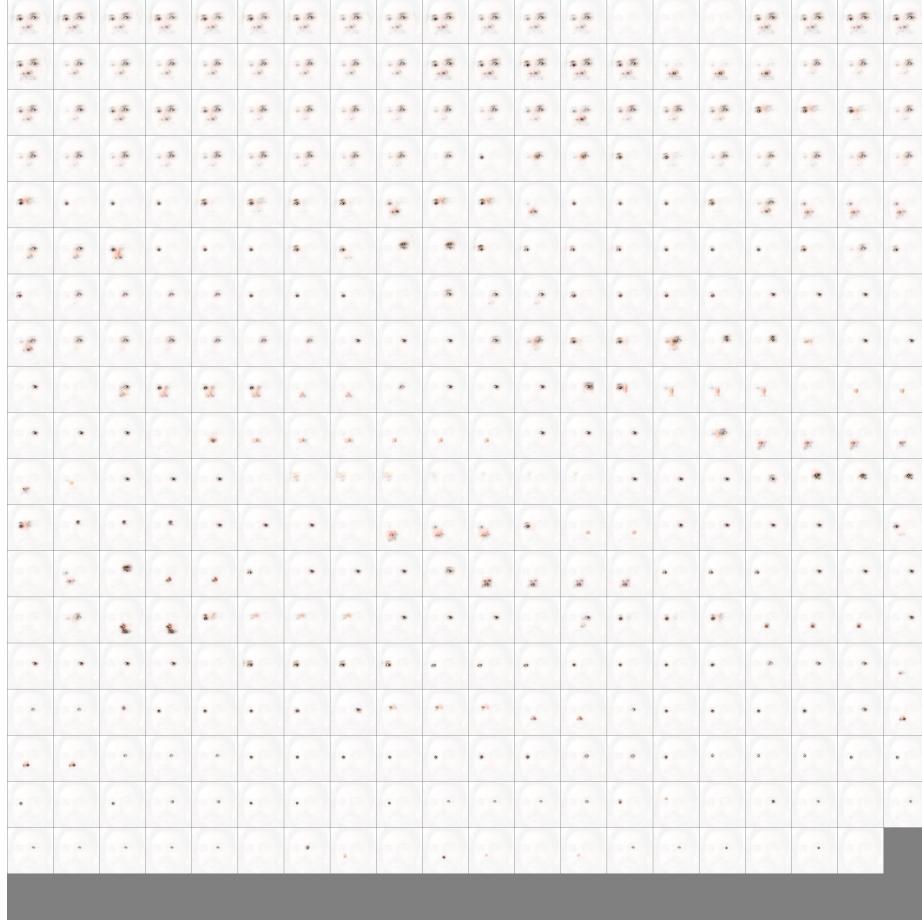
**Fig. S11.** Subtree triplet EBP (Light-CNN [35]). This montage should be compared with Fig. S6 to compare the differences for subtree triplet EBP for ResNet-101 vs. light-CNN.



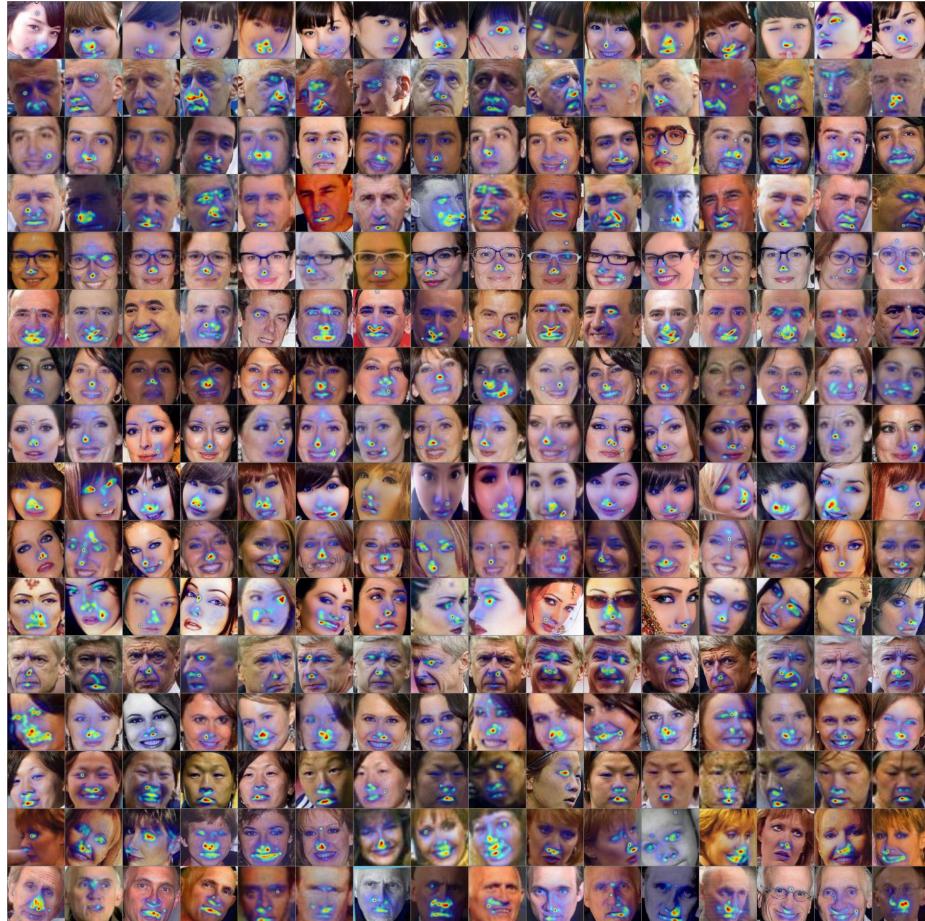
**Fig. S12.** Single probe montage (Light-CNN [35]). This montage should be compared with Fig. S7 to compare the effect of top- $k$  non-mates for ResNet-101 vs. light-CNN.



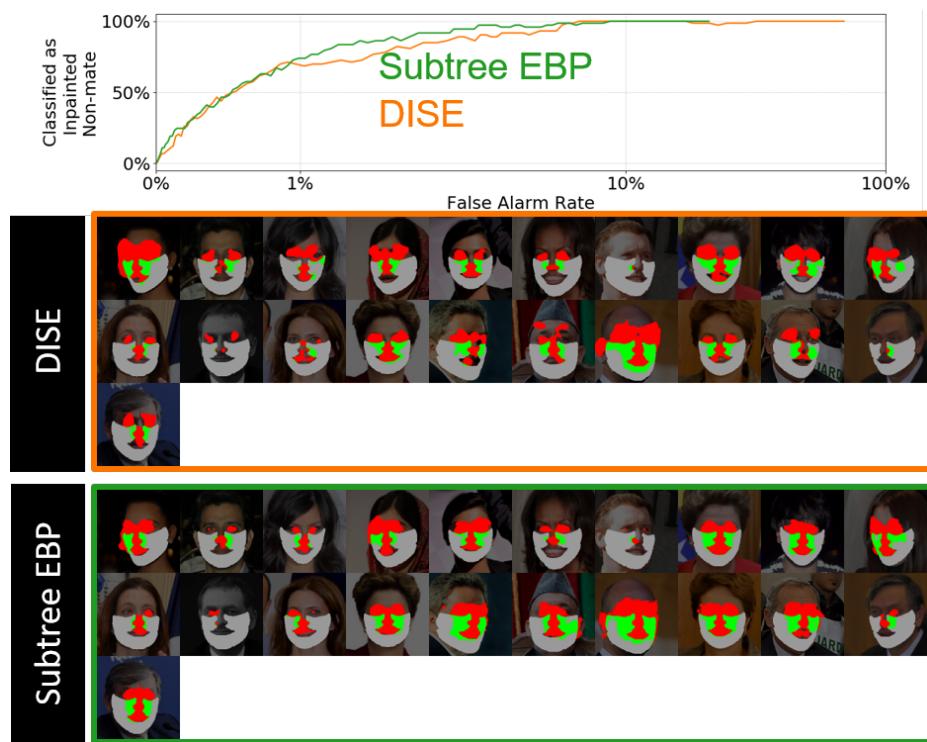
**Fig. S13.** Repeated probe montage (Light-CNN [35]). This montage shows the same probe repeated across each row to highlight the effect of the non-mate in the triplet on the resulting saliency map.



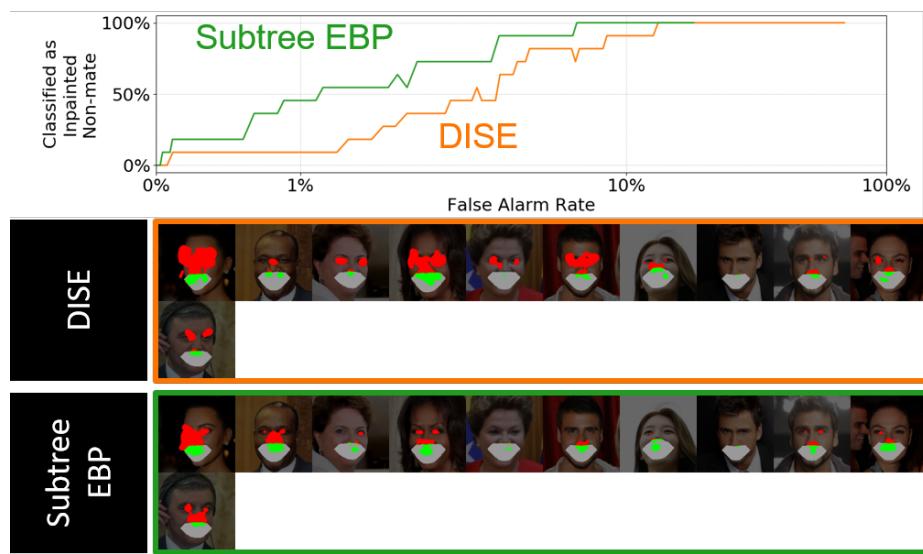
**Fig. S14.** Layerwise EBP. This montage shows the EBP saliency map generated starting from the maximum excitation for each layer in a ResNet-101 network. The layers are ordered rowwise, starting from the embedding layer in the upper left down to the image layer in the bottom right. The visualization shows the saliency map encoded as the alpha channel of a cropped face image, so that non-zero saliency results in a more opaque (less transparent) region. This visualization style is useful to accentuate small activations. This result shows that saliency maps starting from the layers closer to the embedding result in holistic regions covering the eyes and nose, layers in the middle show parts such as the eyes, nose and mouth, layers closer to the image are highly localized on specific regions of the image, and some layers provide no excitation at all.



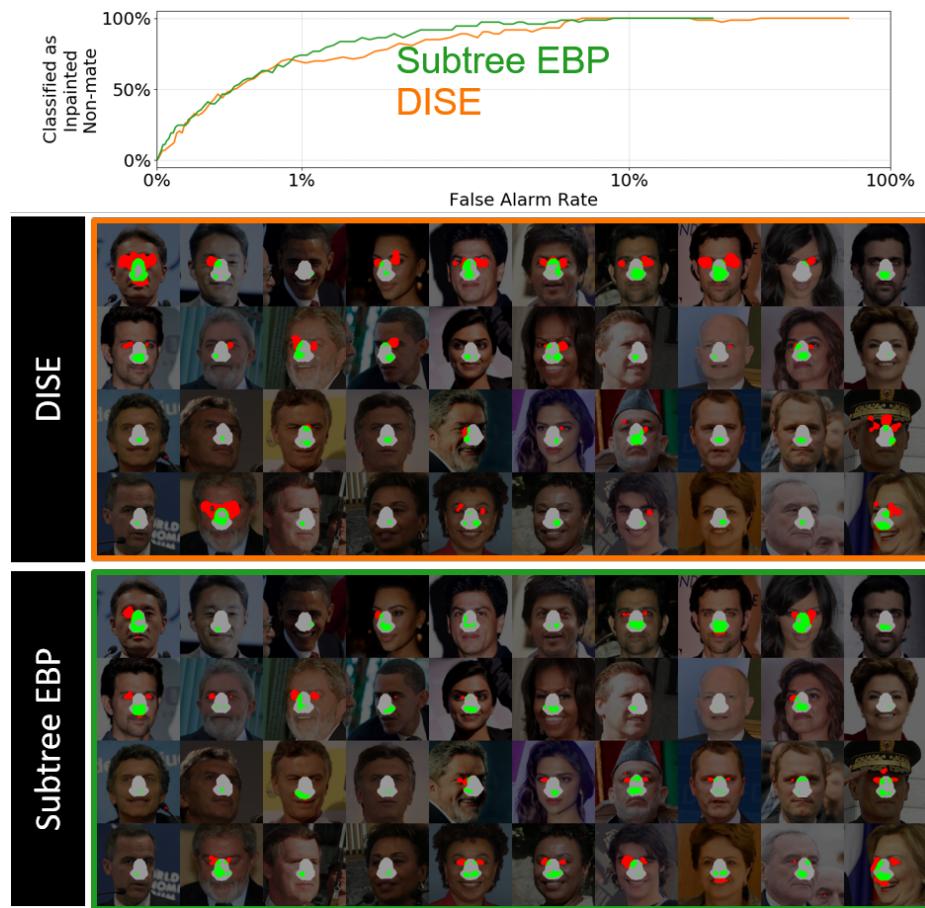
**Fig. S15.** Qualitative visualization study. This figure shows the XFR saliency maps generated using the LightCNN Subtree EBP method for 16 probes (columns) of 16 subjects (rows), each with 16 mates (not shown) and a common set of 8000 nonmates (not shown) all sampled from VGGFace2 [5]. Results show that the discriminative features used to distinguish a subject from the entire nonmate population are primarily the nose and mouth for frontal probes and eyes for non-frontal probes. These network attention maps are remarkably consistent across probes and provide insight into the features that a network uses to distinguish a subject from a large set of nonmates (i.e. What makes you unique?).



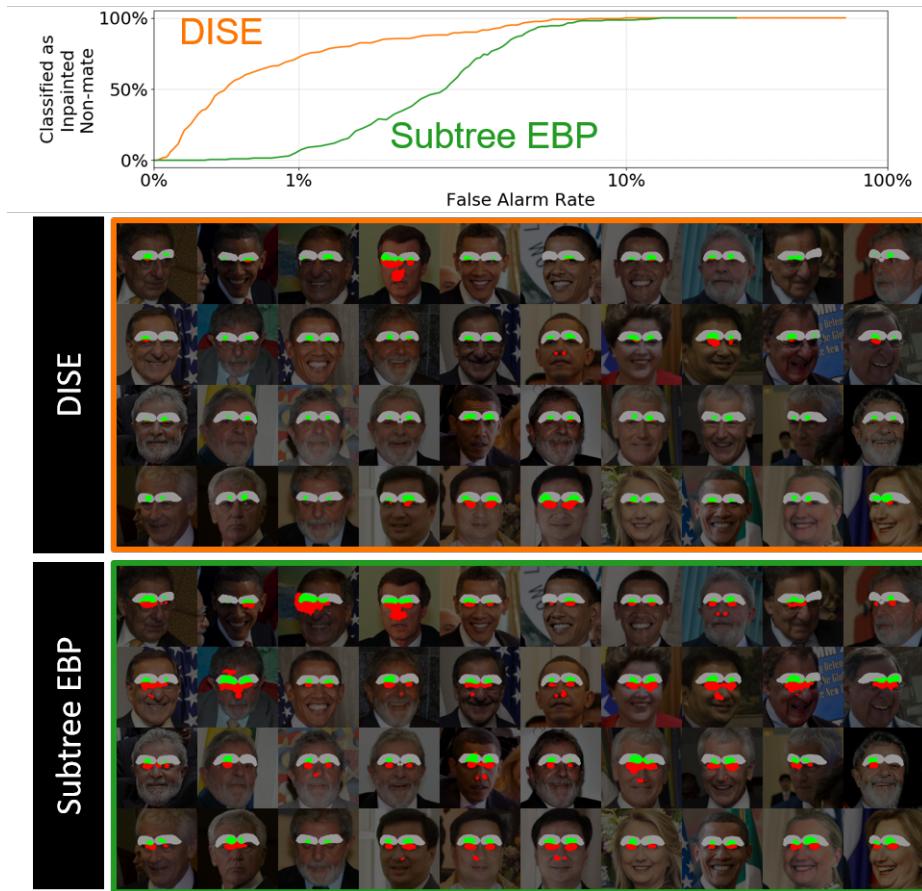
**Fig. S16.** Cheek/Chin Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.



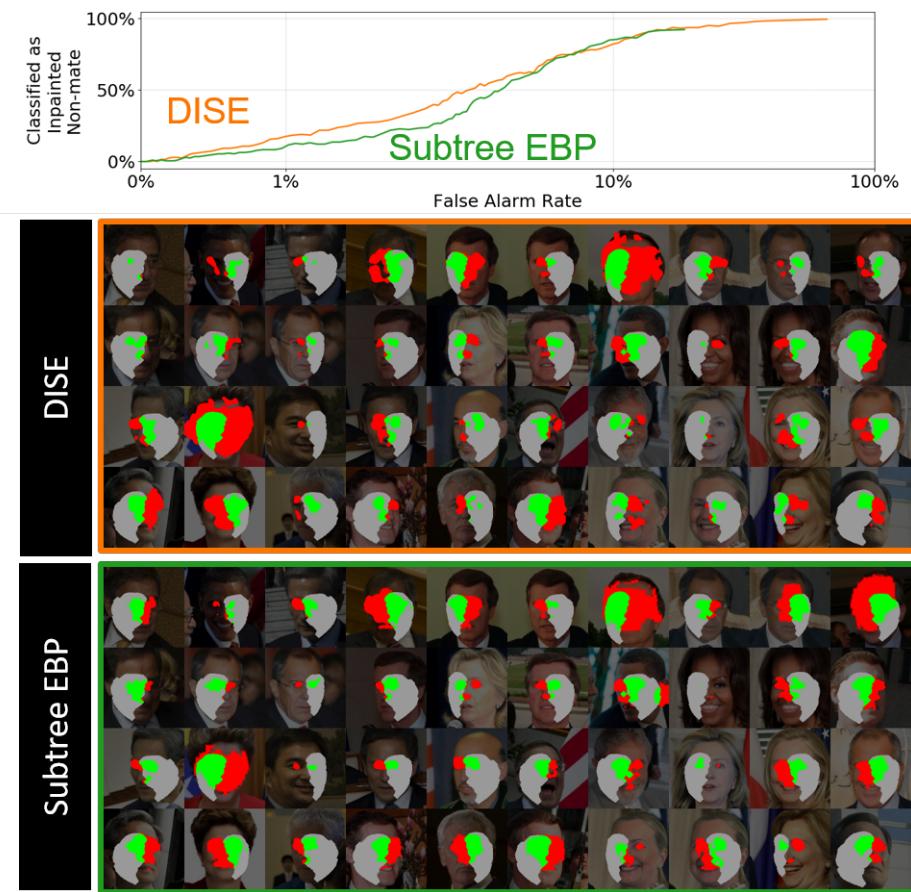
**Fig. S17.** Mouth Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.



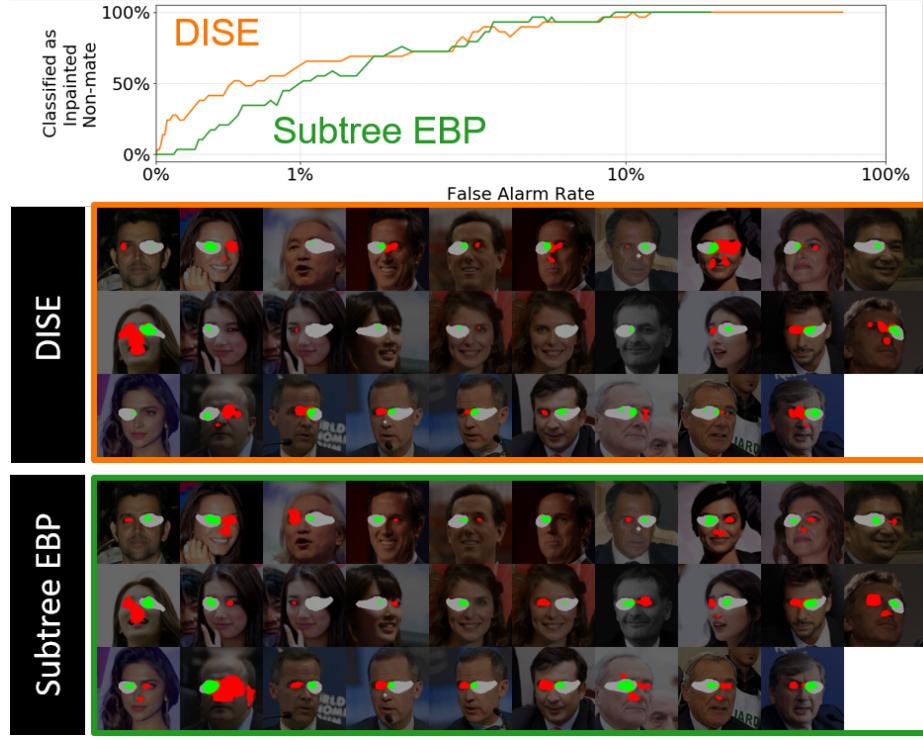
**Fig. S18.** Nose Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.



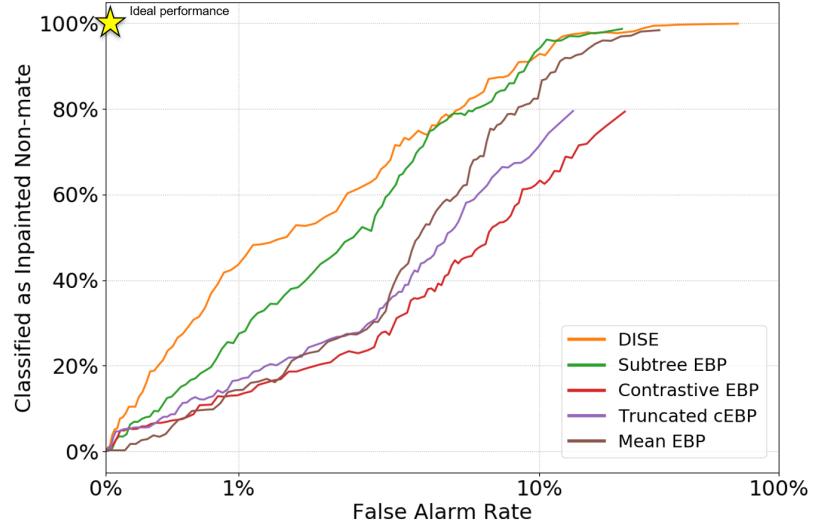
**Fig. S19.** Eyebrow Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.



**Fig. S20.** Left-/Right-Face Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.



**Fig. S21.** Left-/right- eye Mask (ResNet-101): Evaluation plot and classification on saliency maps from Subtree EBP and DISE at identity flip.



**Fig. S22.** Inpainting game analysis using the ResNet-101.