2024 IEEE International Joint Conference on Biometrics (IJCB) | 979-8-3503-6413-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCB62174.2024.10744464

Grains of Saliency: Optimizing Saliency-based Training of Biometric Attack Detection Models

Colton R. Crum, Samuel Webster, Adam Czajka
University of Notre Dame
Notre Dame, IN, USA
{ccrum,swebster,aczajka}@nd.edu

Abstract

Incorporating human-perceptual intelligence into model training has shown to increase the generalization capability of models in several difficult biometric tasks, such as presentation attack detection (PAD) and detection of synthetic samples. After the initial collection phase, human visual saliency (e.g., eye-tracking data, or handwritten annotations) can be integrated into model training through attention mechanisms, augmented training samples, or through human perception-related components of loss functions. Despite their successes, a vital, but seemingly neglected, aspect of any saliency-based training is the level of salience granularity (e.g., bounding boxes, single saliency maps, or saliency aggregated from multiple subjects) necessary to find a balance between reaping the full benefits of human saliency and the cost of its collection. In this paper, we explore several different levels of salience granularity and demonstrate that increased generalization capabilities of PAD and synthetic face detection can be achieved by using simple yet effective saliency post-processing techniques across several different CNNs.

1. Introduction

An ongoing challenge across biometric presentation attack detection (PAD) involves obtaining sufficient model generalization to unknown attack types, or unknown variants of known attacks. For real-world biometric system implementations, this model generalization is crucial, as new attack types show up frequently and cannot be meaningfully represented with sufficient training samples, if at all. Despite the accelerated gains made in recent years for closed-set recognition tasks (all attack types, or their variants known during training), these trends have failed to fully materialize in biometrics due to their appetite for mammoth amounts of training data. As a consequence, state-of-the-art (SOTA) biometric PAD performance is lackluster at best, as evidence from the recurrent Liveness Detection (LivDet) competitions for face [31], iris [41, 11, 35], and fingerprint [43, 28, 6].

The incompetence of models on attack types trivial to a human (e.g., doll eye for iris) can be mitigated by incorporating human-perceptual intelligence into model training through attention mechanisms [26], human saliency-guided data augmentations [3], or loss function components penalizing divergence of model's saliency from human saliency [5]. Despite their success, human saliency-based methods are significantly diminished by the sheer costs associated with acquisition (monetary, time, availability of human subjects), often curtailing its practical use. However, a seemingly obvious but universally unanswered question at the cornerstone of all saliency-based methods is the level of detail, or granularity, necessary to achieve the desired generalization. Salience granularity can be explored through the initial acquisition phase (e.g., high-resolution eye fixation, or simply bounding boxes roughly approximating human salience), or through post-processing measures (e.g., averaging annotation maps from multiple experts, or using a single expert).

Understanding the optimal level of salience granularity leads to a better assessment of the costs associated with salience acquisition. In addition, salience granularity validates the ability of saliency-based methods to incorporate salient information meaningfully into model training. Without this necessary foundation, how the collection and assembly of saliency impacts the generalization performance of saliency-based training downstream remains unclear. We find that for iris PAD and synthetic face detection, single saliency maps provide a sufficient amount of human-sourced information, which suggests that the collection of fine-grained salience may not be necessary. Additionally, it answers the previous unrealized trade-off between the quality (granularity) and quantity of human salient-information necessary for successful saliency-based training. Our results suggest that the quantity of saliency contributes more to model generalization more than its quality (depending on

the biometric modality). Furthermore, we explore salience granularity across several different sources, including human subjects, models trained to mimic human saliency, and domain-specific segmentation models. We find that substantial performance gains can be made within saliency-based training by using optimal salience granularity with no additional overhead.

We organized our paper around the following research questions:

- **RQ1**: What is the optimal level of granularity of human saliency maps for saliency-based training of models detecting biometric spoofs?
- **RQ2**: Does the optimal level of granularity generalize across different biometric presentation attack instruments?
- **RQ3**: Do models trained to mimic human saliency offer image annotations which – when used in saliency-guided trained – lead to better generalization?
- **RQ4**: Can saliency be sourced from domain-specific segmentation models instead of humans?

We release our source codes, model weights and saliency maps to allow others to replicate all the experiments.[1]

2. Related Work

Incorporating human-perception into the training of deep learning models has shown to increase generalization performance [5, 4, 39], create more human-interpretable outputs [9, 39], increase overall training efficiency by requiring less samples [38], and even help ease regulatory tensions within high-risk AI [8]. Domains such as medical imaging and biometrics largely benefit from incorporating human-salient information into training since these scenarios pose challenging constraints not common within general vision tasks. For biometric presentation attack detection, models often operate within open-set contexts where it is unrealistic or otherwise impossible to obtain training samples for every possible attack type. As a result, human-perceptual components have proven invaluable towards increasing generalization capabilities by augmenting training samples [4], attention modules [26, 39], or through loss function components [5, 30, 38].

While methods on how to effectively incorporate human saliency into model training are important, arguably a more crucial aspect of saliency-based training is how to effectively source and assembly the raw saliency, which is often overlooked. Many saliency-based rely upon eye-tracking [38], gaze patterns [39], or written annotations [5, 4]. After the initial collection phase, most post-processing methods include averaging correctly classified samples together into a single saliency map. However, this vastly reduces the number of available saliency and leaves unanswered questions as to how this might affect models trained downstream.

[1] https://github.com/CVRL/GrainsOfSaliency

3. Methodology

In this section, we first describe the training, validation, and testing datasets. Second, we define three levels of salience granularity and describe several sources of salience used within saliency-based training. Finally, we describe the training and evaluation procedure used to evaluate the research questions presented in the Introduction.

3.1. Datasets

We evaluate the affect of salience granularity from several sources for iris-PAD and synthetic face detection tasks. These tasks were selected due to their availability of human annotations necessary to explore granularity across consistent amounts of salient-information, and have proven useful using the CYBORG loss function [5].

Training & Validation Set For the **iris-PAD** task, training and validation images were sampled from a superset composed of various live iris and iris PAD datasets [1, 24, 14, 22, 44, 37, 23, 40, 36, 42, 12]. The training set consisted of 765 samples comprising of bona fide (live) and seven spoof attack types (artificial, diseased, post mortem, paper print outs, synthetic, textured contact lens, textured contact lens & printed), offered by [4]. The validation set comprised of 23,312 samples, completely disjoint from training and testing sets.

For the **synthetic face detection** task, we follow the training splits introduced in [5][2]. The training set consisted of 1821 (919 real and 902 synthetic), and the validation set consisted of 20,000 samples (10,000 real and 10,0000 synthetic) extracted from the FRGC-Subset [29], SREFI [2] and StyleGAN2-generated acquired from *thispersondoesnotexist.com*.

Each training set was accompanied by human saliency maps (*i.e*. every training sample has a corresponding saliency map), as described in the next section. No saliency accompanied the validation set.

Test Set For the **iris-PAD** task, we use LivDet-2020 to benchmark model generalization performance [11]. This edition of LivDet is particularly useful within saliency-based training as a variety of attack types can be meaningfully annotated by human subjects (*i.e*. post mortem), which translate downstream to raise generalization performance.

[2] The authors of this paper would like to thank the authors of [5] for sharing their data and training splits with us.

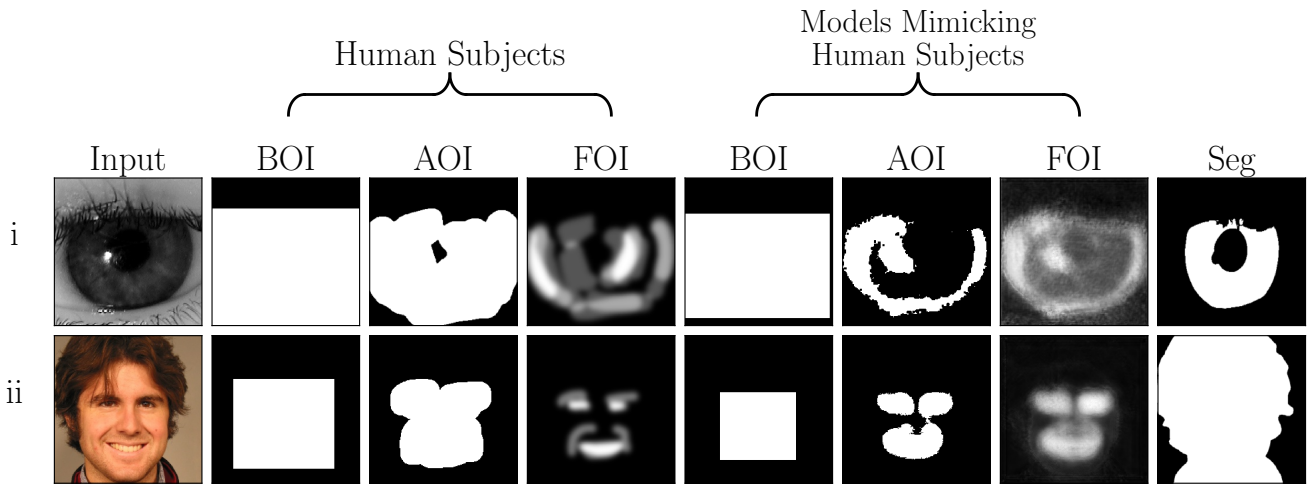


Figure 1. Examples of salience granularity used in saliency-based training defined within this paper: Boundary of Interest (BOI), Area of Interest (AOI), Features of Interest (FOI), sourced from either human subjects or models that were trained to mimic the human subjects. "Seg" indicates segmentation masks sourced from domain-specific segmentation models; (i) iris presentation attack detection task and (ii) synthetic face detection task.

Our paper aims to explore the optimal salience granularity (by first using configurations that have already proven to work), and not introduce attack-type specifics that distract from our analysis. For the **synthetic face detection task**, we sub-sampled the test set provided by [5] to reduce computational overhead. 500 images were randomly without replacement from two live (FFHQ [18] and CelebA-HQ [18]) and 1000 images from six synthetic, GAN-generated sources (ProGAN [18], StarGANv2 [7], StyleGAN [18], StyleGAN2 [21], StyleGAN2-ADA [19], and StyleGAN3 [20]. In total, the synthetic face detection test set comprised of 7,000 (1000 live and 6000 synthetic) test images.

Finally, no saliency accompanied the testing sets.

3.2. Acquisition of Salience

In this section, we first define three levels of salience granularity (Boundary of Interest, Area of Interest, and Features of Interest), aimed to provide a varied amount of salience. Next, we describe three sources of salience (human subjects, models trained to mimic human subjects, and domain-specific segmentation models). These configurations allow for a cross-sectional analysis surrounding the optimal salience granularity and from which particular source. For our experiments, we use the human saliency provided by previous works for iris-PAD [4] and synthetic face detection [5], for its history of success in implementation, despite not explicitly including expert annotators.

Granularity of Human Salience As described in Sec.2, previous work spent little time exploring effective means of assembling salience from human annotations. Most works simply averaged annotations from correctly classified samples, aggregating these samples into a single saliency map [4, 5]. As a consequence, the number of training samples with accompanying saliency maps was pruned to a meager few hundred samples. In Boyd *et al*., 10,750 saliency maps collected were reduced to only 1,821 [5]. This configuration prioritizes fine-grained, human-salient features at the expense of an abundant supply. However, to our knowledge, no prior work has viewed human saliency under a variable approach, or within different levels of granularity.

In this work, we evaluate salience (sourced by human subjects and models trained to mimic human subjects) under the following levels of granularity (see Fig. 1):

- **Boundary of Interest (BOI)**: a rectangular box that indicates a general boundary pertaining to human saliency.
- **Area of Interest (AOI)**: a binarized region that indicates human saliency *uniformly*.
- **Features of Interest (FOI)**: regions that indicate fine-grained, *variable* amounts human saliency within specific regions of the input sample.

Since previous works average saliency maps together to obtain fine-grained detail within each sample, we describe this as Features of Interest (FOI). This assembly prioritizes sample-level features pertinent towards solving the larger task. In an effort to standardize the amount of salient-information contained within each sample while maintaining a varied degree of granularity, Boundary of Interest (BOI) and Area of Interest (AOI) salience was derived from the FOI salience. AOI salience was generated by first binarizing FOI salience, wherein pixels with values greater than

0 were set to 255. For BOI salience, a minimally enclosing rectangle was drawn that encompasses all salient regions found within the FOI salience. The same process was conducted from salience generated by both human subjects and models trained to mimic human subjects (described in the section below).

Models Mimicking Human Subjects Given the constraints associated with collecting salience from human subjects, a more efficient use of human saliency is to train a model to generate or emulate human saliency. Under this scenario, an autoecoder-type model is trained to learn human-salient features given a cooresponding input image. Once trained, the model can generate salience easily at scale (unlike human saliency). To explore the potential sources of salience, autoencoders were trained using human-sourced (FOI) saliency maps (described above) and used to generate salience for a second training split. To make fair comparisons, this generated split was comprised of the same size (764 samples for iris-PAD, 1821 samples for synthetic face) and sources as detailed above. AOI and BOI saliency were derived from the FOI saliency sourced by the autoencoder for both respective tasks. For iris-PAD, AOI saliency was derived by binarizing the FOI saliency with a threshold of 0.5, wherein pixels with values greater than 127 were set to 255 and lesser or equal values were set to 0. BOI saliency was drawn with a minimally enclosing rectangle over the salient regions found within the AOI saliency eroded by a 3×3 kernel, initialized uniformly at 1.0 for a single iteration. For the synthetic face detection task, AOI saliency was likewise derived by binarizing the FOI saliency with a threshold of 0.5. BOI saliency was also generated from the AOI saliency, except no erosion was applied prior to computation. For convenience in training the autoencoders, a lightweight parameter sweep was performed using RayTune for efficient traning of the autoencoder for each respective task[25]. For iris-PAD, a DenseNet-161-based UNET [16, 32] was trained using the Adam optimizer ($lr = 0.0001$) for 50 epochs and a batch size of 20. For the synthetic face detection task, an Inception-V4-based UNET [33, 32] was trained using the Adam optimizer ($lr = 0.0001$) for 50 epochs and a batch size of 10. Both models were trained using a sigmoid activation function, initialized with ImageNet weights [13], and using a Mean Squared Error (MSE) loss, sourced from [17].

Segmentation Models Domain-specific segmentation models offer feature-level masks that may be useful in saliency-based training, requiring no overhead and essentially for free. Additionally, if domain-specific segmentation models can achieve comparable performance to human subjects, human salience collections may be antiquated and completely unnecessary. We explore the validity of SOTA segmentation-sourced saliency for iris-PAD using CC-NET [27, 10], and the BiSeNet-based face parser [46, 45] for synthetic face. Models were evaluated off the shelf (without any fine-tuning or training), and segmentations were generated using the original training split described in Sec.3.1 (see column "Seg" in Fig. 1).

3.3. Training & Evaluation Configurations

The various configurations of salience granularity and salience sources was evaluated across three CNN architectures, ResNet50 [15], DenseNet-121 [16], and Inception-V3 [34] using the CYBORG loss function [5]. The CYBORG loss parameter weighting human-guidance and classification performance was given equal weight ($\alpha = 0.5$) for all salience granularity experiments (BOI, AOI, and FOI), as in [5]. Baseline training without the use of any salience was performed with cross-entropy loss ("None" in Tab. 1 and Tab. 3) under the same training configurations for additional comparison. All models were initialized with ImageNet weights [13], trained with a batch size of 20 using Stochastic Gradient Descent (SGD) for 50 epochs, with learning rate of 0.005, modified by a scheduler, which reduced the learning rate by a factor of 0.1 every 12 epochs. We evaluate the generalization performance on the test set using Area Under the ROC Curve (AUC). For the **iris-PAD task**, the means and standard deviations of AUC are reported across 3 independent runs in Tab. 1, whereas the **synthetic face detection** models are reported across 5 independent runs (Tab. 3). Since our analysis is focused solely on measuring granularity, we do not include generalization results from previous work as they use the same granularity of salience for all experiments (FOI granularity).

4. Results

This section describes the results, organized by the research questions presented in the Introduction. Results for all variants are summarized in Tab. 1 and Tab. 3, and ROC curves are displayed in Fig. 2 and Fig. 3.

4.1. RQ1: What is the optimal level of granularity?

Focusing first on salience sourced from human subjects, granularity has a sizable impact on generalization performance for iris-PAD (see Tab. 1). The conventional configuration of salience, FOI, common previous works, translated downstream to sub-optimal model generalization (average AUC=0.898), compared to AOI (average AUC=0.910) across all three architectures. More specifically, DenseNet had the largest improvement over the conventional saliency selection. These results suggest that strong generalization can be achieved through simpler salience granularity indicating merely Areas of Interest (AOI), which significantly reduces the overhead associated with human subject collection. More specifically, it suggests that detailed

Table 1. Three CNN architectures (ResNet50, DenseNet-121, and Inception-V3) generalization performance with saliency-based training across various sources of saliency (Human Subjects, Models Mimicking Human Saliency, Segementation Models, or None) across three different salience granularities (only applicable for Human Subjects and Models Mimicking Human Saliency) for **iris-PAD task**. Means and standard deviations of Area Under the Curve (AUC) are reported across **3 independent runs**. Optimal saliency configuration for each backbone is **bolded**, and the optimal average configuration across backbones is underlined.

Source of Saliency	ResNet	DenseNet	Inception	Average
	Backbones Used in Saliency-Based Training			
Human Subjects				
Boundary of Interest (BOI)	0.886±0.015	0.903±0.010	0.873±0.023	0.887±0.016
Area of Interest (AOI)	0.909±0.006	0.921±0.013	0.900±0.005	0.910±0.008
Features of Interest (FOI)	0.908±0.005	0.895±0.018	0.890±0.015	0.898±0.013
Models Mimicking Human Subjects				
Boundary of Interest (BOI)	0.939±0.008	0.933±0.016	0.953±0.007	0.942±0.010
Area of Interest (AOI)	**0.956±0.006**	**0.962±0.005**	**0.962±0.013**	0.960±0.008
Features of Interest (FOI)	0.945±0.007	0.955±0.003	0.958±0.007	0.953±0.006
Segmentation Models				
Iris Segmentations	0.894±0.010	0.884±0.004	0.878±0.022	0.885±0.012
None				
Baseline	0.875±0.013	0.893±0.019	0.889±0.006	0.886±0.010

Table 2. Same as Tab. 1, except the Equal Error Rate is reported.

Source of Saliency	ResNet	DenseNet	Inception	Average
	Backbones Used in Saliency-Based Training			
Human Subjects				
Boundary of Interest (BOI)	0.197±0.013	0.182±0.011	0.202±0.023	0.194±0.018
Area of Interest (AOI)	0.174±0.005	0.163±0.017	0.183±0.005	0.173±0.013
Features of Interest (FOI)	0.178±0.007	0.189±0.017	0.196±0.016	0.188±0.015
Models Mimicking Human Subjects				
Boundary of Interest (BOI)	0.140±0.009	0.143±0.018	0.120±0.008	0.134±0.016
Area of Interest (AOI)	**0.115±0.006**	**0.107±0.008**	**0.102±0.021**	0.108±0.014
Features of Interest (FOI)	0.131±0.013	0.114±0.005	0.111±0.012	0.119±0.013
Segmentation Models				
Iris Segmentations	0.187±0.009	0.199±0.003	0.203±0.017	0.196±0.013
None				
Baseline	0.206±0.015	0.191±0.025	0.193±0.004	0.197±0.017

saliency methods (*i.e*. averaging saliency maps from multiple annotators, or collecting salience through eye-tracking) does not reap additional generalization benefits for iris-PAD. Furthermore, simply collecting a boundary of salience (BOI in 1) proves insufficient in generalization performance (AUC=0.887), and is on par with using no saliency at all (AUC=0.886). Thus, **the answer to RQ1 is Area of Interest (AOI) salience granularity for iris-PAD**.

4.2. RQ2: Does optimal granularity generalize across different presentation attack instruments?

Tab. 3 indicates that optimal salience granularity generalizes differently for synthetic face detection compared to iris-PAD. More specifically, the results suggest that salience granularity for synthetic face detection tasks are more dependent on the architectures themselves. The optimal salience granularity using human subjects-sourced salience for ResNet architectures was BOI (AUC=0.604), whereas optimal granularity for both DenseNet and Inception architectures was FOI (0.643 and 0.641, respectively). However, it's worth noting the wide standard deviations reported indicate the differences in salience granularity may not be statistically significant for synthetic face detection. These results could also indicate that the human annotations collected for this specific synthetic face detection may not be as valuable as for iris-PAD. However, unlike iris-PAD, including any type of salience during training (BOI, AOI, or FOI) for synthetic face detection on average boosted generalization performance across all architectures (compared to the Baseline, bottom row Tab. 3), suggesting that saliency-based gains can be made with quite simple salience over the baseline.

The answer to RQ2 is negative: optimal salience granularity is different across biometric modalities (iris-PAD and synthetic face detection), and the use of saliency-based training necessitates thoughtful consideration pertaining to salience granularity and the corresponding architectures used within saliency-based training.

4.3. RQ3: Do models trained to mimic human saliency offer image annotations which, when used in saliency-guided trained, lead to better generalization?

Arguably the most interesting finding from our experiments is that models trained to mimic human saliency offer impressive gains in generalization performance, including over the human subjects. The majority of model back-

Table 3. Same as Tab. 1, except for **synthetic face detection task**, where results are reported across **5 independent runs**. **Per class accuracy** (Real / Synthetic) is also reported for additional comparisons.

Source of Saliency	ResNet	DenseNet	Inception	Average AUC	Average Accuracy	
	Backbones Used in Saliency-Based Training				Real	Synthetic
Human Subjects						
Boundary of Interest (BOI)	0.604±0.048	0.546±0.059	0.617±0.062	0.589±0.056	0.001±0.002	1.0 ±0.001
Area of Interest (AOI)	0.579±0.035	0.577±0.045	0.639±0.029	0.598±0.036	0.003±0.002	0.999±0.001
Features of Interest (FOI)	0.590±0.023	**0.643±0.033**	**0.641±0.046**	0.629±0.037	0.01±0.009	0.997±0.001
Models Mimicking Human Subjects						
Boundary of Interest (BOI)	0.584±0.031	0.583±0.054	0.539±0.034	0.569±0.040	0.001 ± 0.001	1.0±0.0
Area of Interest (AOI)	**0.614±0.056**	0.640±0.046	0.608±0.071	0.621±0.058	0.0±0.0	1.0±0.0
Features of Interest (FOI)	0.600±0.025	0.619±0.033	0.632±0.019	0.617±0.026	0.001 ± 0.001	1.0 ± 0.0
Segmentation Models						
Face Segmentations	0.548±0.048	0.451±0.050	0.579±0.040	0.526±0.046	0.002±0.001	1.0±0.0
None						
Baseline	0.572±0.047	0.535±0.075	0.540±0.037	0.549±0.053	0.057±0.03	0.971±0.015

Table 4. Same as Tab. 3, except the Equal Error Rate (EER) is reported.

Source of Saliency	ResNet	DenseNet	Inception	Average
	Backbones Used in Saliency-Based Training			
Human Subjects				
Boundary of Interest (BOI)	0.429±0.035	0.462±0.043	0.413±0.048	0.435±0.045
Area of Interest (AOI)	0.440±0.022	0.442±0.032	0.390±0.025	0.424±0.035
Features of Interest (FOI)	0.433±0.023	0.391±0.029	0.390±0.035	0.405±0.034
Models Mimicking Human Subjects				
Boundary of Interest (BOI)	0.594±0.147	0.458±0.066	0.468±0.032	0.507±0.109
Area of Interest (AOI)	**0.408±0.039**	**0.388±0.043**	0.405±0.059	0.400±0.045
Features of Interest (FOI)	0.515±0.123	0.449±0.097	**0.384±0.018**	0.449±0.101
Segmentation Models				
Iris Segmentations	0.462±0.037	0.537±0.037	0.440±0.032	0.480±0.054
None				
Baseline	0.444±0.034	0.491±0.045	0.453±0.006	0.463±0.037

bones across both tasks (except DenseNet and Inception for synthetic face detection) performed best when models were trained using saliecy sourced from models mimicking human saliency. For iris-PAD, all models held a substantial gains across all granularities (cross model AUC averages include BOI=0.942, AOI=0.960, FOI=0.953) compared to human subjects (BOI=0.887, AOI=0.910, FOI=0.898). Similar to the findings of RQ2, the synthetic face models offered a mixed result of improvement, largely depending on the architecture. Models mimicking human subjects achieved the best average performance out of all salience configurations for ResNet (AOI=0.614), whereas DenseNet and Inception benefited best from fined-grained saliency sourced from human subjects (0.643 and 0.641, respectively). The performance gains using saliency offered by models mimicking human subjects over the human subjects directly can be a result of supplemental salient-information incorporated by the autoencoder (see the differences between BOI, AOI, and FOI between human subjects and models mimicking human subjects in Fig. 1). Although the autoencoder was trained to mimic the human annotations, it still has agency in deciding which salient regions to annotate while satisfying the initial human annotation. This process allows an interweaving of complementary human and model salient-information to be encoded directly within the generated saliency map, which can boost model generalization performance downstream with saliency-based training. Finally, these results suggest that the availability of human saliency collections can be expanded without the scaling limitations associated with collection from human subjects. More specifically, these models largely benefited from AOI granularity, which strikes a necessary balance between fine-grained, feature level saliency (FOI) and the unfocused, blanket saliency of BOI.

4.4. RQ4: Can saliency be sourced from domain-specific segmentation models instead of humans?

Given our findings in RQ3, a question arises whether saliency collection from human subjects is necessary at all, and whether domain-specific segmentation models can be used instead of human-sourced (*i.e.* by human subjects or models mimicking human subjects). We found that saliency-based training using segmentation models significantly falls short of human-sourced saliency, but is occasionally an improvement from no salience use at all. For iris-PAD, using iris segmenter-based saliency offered the worst average model performance (AUC=0.885), including the baseline with no saliency use at all (AUC=0.886). These results bolster the need for human subjects within the saliency-generation pipeline (either sourced directly from human subjects, or training models to mimic human subjects) for presentation attack detection and related synthetic detection tasks. Domain-specific segmentation models are

trained to simply locate basic features of the input sample (*i.e*. annular iris, or nose for faces), which is insufficient information required to solve these tasks. Often the information necessary to correctly classify the PAD sample goes beyond simple feature matching, and requires models to look elsewhere (*i.e*. corners of the sample for post mortem attack types). Unlike domain-specific segmentation masks, human subjects locate these anomalous regions (as do models trained to mimic the human subjects), guiding the models towards salient-regions necessary to solve the PAD task.

The answer to RQ4 is negative: saliency is best sourced directly from humans (human subjects or models mimicking human subjects). However, our findings suggest that using saliency from segmentation models may provide generalization gains over traditional configurations where no saliency is used during training.

5. Conclusion

Efforts to raise generalization in challenging biometric tasks have often incorporated human-perceptual information into the training of CNN models, most commonly through saliency-based training. Despite their ability to improve generalization in iris-PAD and synthetic face detection tasks, obtaining fine-grained salience from human subjects remains an ongoing obstacle. However, our results indicate this challenge may be an illusion for some biometric modalities. In this paper, we find that model generalization can be improved through more manageable collection and assembly. First, we define three levels of salience granularity: Boundary of Interest (BOI), Area of Interest (AOI), and Features of Interest (FOI), which all have varying degrees of detail and associated acquisition costs. Second, we show that traditional salience granularity methods (FOI) is often inferior to more simpler methods (AOI) for iris-PAD tasks. For synthetic face detection, we found that optimal granularity is largely architectural dependent, though models benefited from the use of any level of salience (BOI, AOI, FOI) over no salience use at all during training. Third, we showcase how substantial generalization gains can be made using salience generated by models that mimic human subjects, which combine the complementary information between human subjects and the model. Finally, we show the ineffectiveness of salience sourced from domain-specific models within saliency-based training, encouraging for a human to be involved in the salience curating process.

We acknowledge a performance gap between the biometric domains explored, correlating with human performance; iris-PAD anomalies are easier for both humans and CNNs to detect compared to synthetic faces, whose fidelity is near to that of real faces. Despite this, our results illustrate that optimizing saliency granularity is an effective path to raising generalization in real-world biometric applications. Our findings suggest saliency-based training is limited by the necessity of salience collection, the domain of application, and the architecture of implementation. Our paper calls attention to an important, but remarkably missed component to all saliency-based training methods and suggests that generalization performance can be improved through less taxing means of acquisition.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2237880. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Chinese Academy of Sciences Institute of Automation. http://www.cbsr.ia.ac.cn/china/Iris%20Databases%20CH.asp. Accessed: 03-12-2021.

[2] S. Banerjee, J. S. Bernhard, W. J. Scheirer, K. W. Bowyer, and P. J. Flynn. SREFI: Synthesis of realistic example face images. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 37–45, 2017.

[3] A. Boyd, K. W. Bowyer, and A. Czajka. Human-aided saliency maps improve generalization of deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA*, pages 2735–2744, 2022.

[4] A. Boyd, Z. Fang, A. Czajka, and K. W. Bowyer. Iris presentation attack detection: Where are we now? *Pattern Recognition Letters*, 138:483–489, 2020.

[5] A. Boyd, P. Tinsley, K. Bowyer, and A. Czajka. Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6097–6106, 2023.

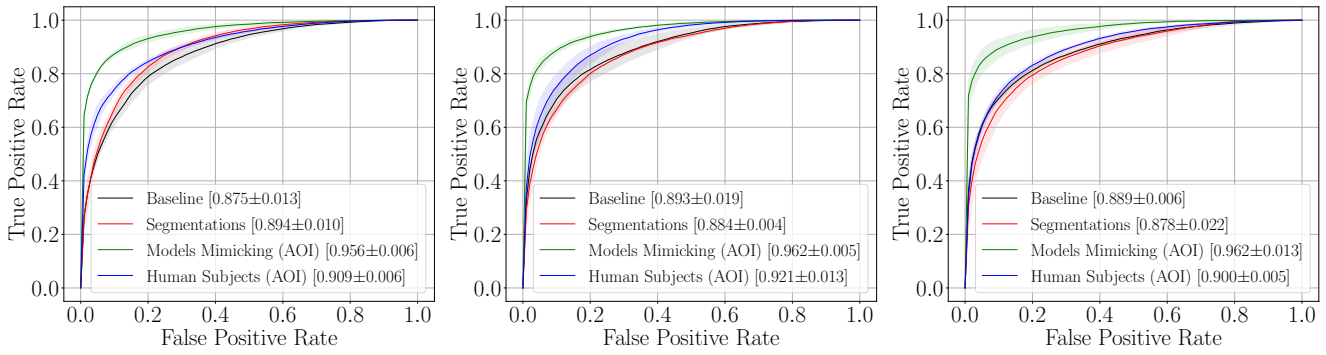
[6] R. Casula, M. Micheletto, G. Orrù, R. Delussu, S. Concas, A. Panzino, and G. L. Marcialis. Livdet 2021 fingerprint liveness detection competition-into the unknown. In *2021 IEEE international joint conference on biometrics (IJCB)*, pages 1–6. IEEE, 2021.

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[8] C. R. Crum and C. Coglianese. Taking training seriously: Human guidance and management-based regulation of artificial intelligence. *arXiv preprint*, 2024.

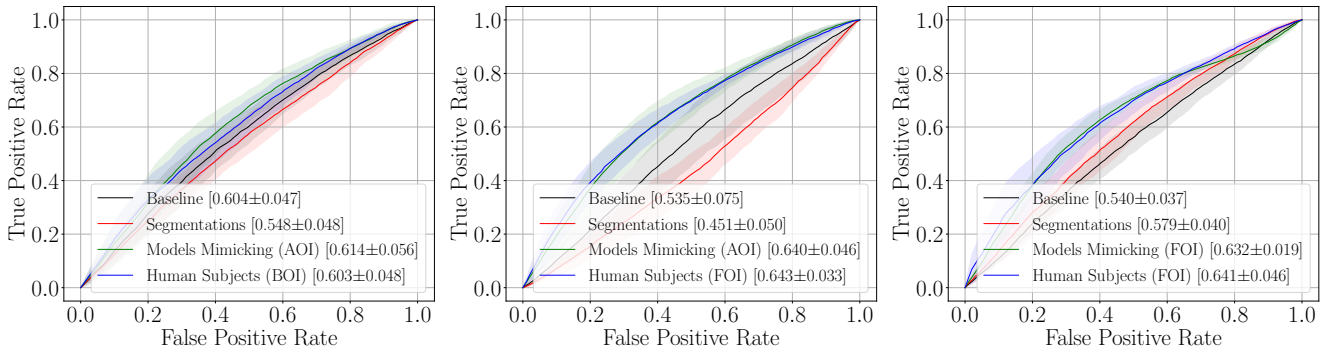
[9] C. R. Crum, P. Tinsley, A. Boyd, J. Piland, C. Sweet, T. Kelley, K. Bowyer, and A. Czajka. Explain to me: Salience-based explainability for synthetic face detection models. *IEEE Transactions on Artificial Intelligence*, 2023.

[10] A. Czajka. Iris recognition designed for post-mortem and diseased eyes. `https://github.com/aczajka/iris-recognition---pm-diseased-human-driven-bsif`, 2023.



(a) ResNet (b) DenseNet (c) Inception

Figure 2. Mean ROC curves and bands representing standard deviations (along the True Positive Rate axis) for all backbones used in saliency-based training with varied configurations of saliency for **iris-PAD (top row)** and **synthetic face detection (bottom row) tasks**. For human subjects and models mimicking human subjects, the optimal granularity (BBOI, AOI, FOI) is selected, indicating that generalization performance improves having human subjects within the saliency generation pipeline. Means and standard deviations of Area Under the Curve (AUC) are reported in brackets.



(a) ResNet (b) DenseNet (c) Inception

Figure 3. Same as in Fig. 2, except for **synthetic face detection**.

[11] P. Das, J. McFiratht, Z. Fang, A. Boyd, G. Jang, A. Mohammadi, S. Purnapatra, D. Yambay, S. Marcel, M. Trokielewicz, et al. Iris liveness detection competition (livdet-iris)-the 2020 edition. In *2020 IEEE international joint conference on biometrics (IJCB)*, pages 1–9. IEEE, 2020.

[12] P. Das, J. Mcfiratht, Z. Fang, A. Boyd, G. Jang, A. Mohammadi, S. Purnapatra, D. Yambay, S. Marcel, M. Trokielewicz, P. Maciejewicz, K. Bowyer, A. Czajka, S. Schuckers, J. Tapia, S. Gonzalez, M. Fang, N. Damer, F. Boutros, A. Kuijper, R. Sharma, C. Chen, and A. Ross. Iris Liveness Detection Competition (LivDet-Iris) - The 2020 Edition. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2020.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[14] J. Galbally, J. Ortiz-Lopez, J. Fierrez, and J. Ortega-Garcia. Iris liveness detection based on quality related features. In *2012 5th IAPR Int. Conf. on Biometrics (ICB)*, pages 271–276, New Delhi, India, March 2012. IEEE.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[17] P. Iakubovskii. Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`, 2019.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

[20] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *Proc. NeurIPS*, 2021.

[21] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[22] N. Kohli, D. Yadav, M. Vatsa, and R. Singh. Revisiting iris recognition with color cosmetic contact lenses. In *IEEE Int. Conf. on Biometrics (ICB)*, pages 1–7, Madrid, Spain, June 2013. IEEE.

[23] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore. Detecting medley of iris spoofing attacks using desist. In *IEEE Int. Conf. on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–6, Niagara Falls, NY, USA, Sept 2016. IEEE.

[24] S. J. Lee, K. R. Park, Y. J. Lee, K. Bae, and J. H. Kim. Multifeature-based fake iris detection method. *Optical Engineering*, 46(12):1 – 10, 2007.

[25] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[26] D. Linsley, D. Shiebler, S. Eberhardt, and T. Serre. Learning what and where to attend. *arXiv preprint arXiv:1805.08819*, 2018.

[27] S. Mishra, D. Z. Chen, and X. S. Hu. Image complexity guided network compression for biomedical image segmentation. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 18(2):1–23, 2021.

[28] G. Orrù, R. Casula, P. Tuveri, C. Bazzoni, G. Dessalvi, M. Micheletto, L. Ghiani, and G. L. Marcialis. Livdet in action-fingerprint liveness detection competition 2019. In *2019 international conference on biometrics (ICB)*, pages 1–6. IEEE, 2019.

[29] P. J. Phillips, P. J. Flynn, and K. W. Bowyer. Lessons from collecting a million biometric samples. *Image and Vision Computing*, 58:96–107, 2017.

[30] J. Piland, A. Czajka, and C. Sweet. Model focus improves performance of deep learning-based synthetic face detectors. *IEEE Access*, 2023.

[31] S. Purnapatra, N. Smalt, K. Bahmani, P. Das, D. Yambay, A. Mohammadi, A. George, T. Bourlai, S. Marcel, S. Schuckers, et al. Face liveness detection competition (livdet-face)-2021. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2021.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[35] P. Tinsley, S. Purnapatra, M. Mitcheff, A. Boyd, C. Crum, K. Bowyer, P. Flynn, S. Schuckers, A. Czajka, M. Fang, et al. Iris liveness detection competition (livdet-iris)–the 2023 edition. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[36] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Assessment of iris recognition reliability for eyes affected by ocular pathologies. In *IEEE Int. Conf. on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–6, 2015.

[37] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Post-mortem iris recognition with deep-learning-based image segmentation. *Image and Vision Computing*, 94:103866, 2020.

[38] T. van Sonsbeek, X. Zhen, D. Mahapatra, and M. Worring. Probabilistic integration of object level annotations in chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3630–3640, 2023.

[39] B. Wang, H. Pan, A. Aboah, Z. Zhang, E. Keles, D. Torigian, B. Turkbey, E. Krupinski, J. Udupa, and U. Bagci. Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2194–2203, 2024.

[40] Z. Wei, T. Tan, and Z. Sun. Synthesis of large realistic iris databases using patch-based sampling. In *Int. Conf. on Pattern Recognition (ICPR)*, pages 1–4, Tampa, FL, USA, Dec 2008. IEEE.

[41] D. Yambay, B. Becker, N. Kohli, D. Yadav, A. Czajka, K. W. Bowyer, S. Schuckers, R. Singh, M. Vatsa, A. Noore, D. Gragnaniello, C. Sansone, L. Verdoliva, L. He, Y. Ru, H. Li, N. Liu, Z. Sun, and T. Tan. LivDet-Iris 2017 – Iris Liveness Detection Competition 2017. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 733–741, 2017.

[42] D. Yambay, B. Becker, N. Kohli, D. Yadav, A. Czajka, K. W. Bowyer, S. Schuckers, R. Singh, M. Vatsa, A. Noore, D. Gragnaniello, C. Sansone, L. Verdoliva, L. He, Y. Ru, H. Li, N. Liu, Z. Sun, and T. Tan. LivDet Iris 2017 – Iris Liveness Detection Competition 2017. In *IEEE Int. Joint Conf. on Biometrics (IJCB)*, pages 1–6, Denver, CO, USA, 2017. IEEE.

[43] D. Yambay, S. Schuckers, S. Denning, C. Sandmann, A. Bachurinski, and J. Hogan. Livdet 2017-fingerprint systems liveness detection competition. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–9. IEEE, 2018.

[44] D. Yambay, B. Walczak, S. Schuckers, and A. Czajka. Livdet-iris 2015 - iris liveness detection competition 2015. In *IEEE Int. Conf. on Identity, Security and Behavior Analysis (ISBA)*, pages 1–6, New Delhi, India, Feb 2017. IEEE.

[45] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[46] zllrunning. face-parsing.pytorch. `https://github.com/zllrunning/face-parsing.PyTorch`, 2019.