Human-Aided Saliency Maps Improve Generalization of Deep Learning

Aidan Boyd, Kevin Bowyer, Adam Czajka

University of Notre Dame, Notre Dame, IN 46556

{aboyd3,kwb,aczajka}@nd.edu

Abstract

Deep learning has driven remarkable accuracy increases in many computer vision problems. One ongoing challenge is how to achieve the greatest accuracy in cases where training data is limited. A second ongoing challenge is that trained models oftentimes do not generalize well even to new data that is subjectively similar to the training set. We address these challenges in a novel way, with the first-ever (to our knowledge) exploration of encoding human judgement about salient regions of images into the training data. We compare the accuracy and generalization of a state-of-the-art deep learning algorithm for a difficult problem in biometric presentation attack detection when trained on (a) original images with typical data augmentations, and (b) the same original images transformed to encode human judgement about salient image regions. The latter approach results in models that achieve higher accuracy and better generalization, decreasing the error of the LivDet-Iris 2020 winner from 29.78% to 16.37%, and achieving impressive generalization in a leave-one-attack-type-out evaluation scenario. This work opens a new area of study for how to embed human intelligence into training strategies for deep learning to achieve high accuracy and generalization in cases of limited training data.

1. Introduction

Deep learning methods have had huge impact in many areas of computer vision, including generic object detection [27], face recognition [29, 13], medical image analysis [26] and biometrics [37]. They are known for their need for large amounts of training data and for solutions that are often fragile, in the sense of producing state-of-the-art results on a well-defined problem but not generalizing well to what might seem to be related problems or datasets. The generalization problem can be addressed to some degree by design of a large and varied training dataset, or regularization techniques. But the strengths and weaknesses here are like two sides of the deep learning coin. Deep learning-based methods can learn whatever exists in the training data that can
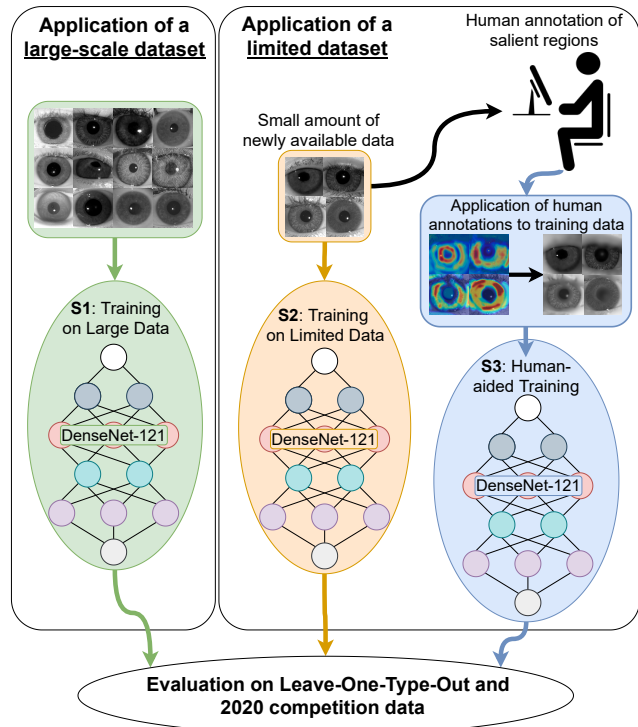

Figure 1: **Do Human-Aided Saliency Maps Improve Generalization of Deep Learning?** We use the same training and architecture to compare models trained on a newly-acquired limited dataset, with typical augmentation techniques (S2), and the same images encoding human judgement of region saliency using multiple levels of blur (S3). For reference, we also present results for training with a much larger original dataset (S1). The use of training images encoding human saliency results in models demonstrating **improved accuracy and generalization over the conventional approach**.

be used to solve the problem. Learning what is incidental to the training data causes the solution to not generalize well.

In this work, we make the novel proposition to transform training data for deep learning in a way that incorporates **human judgement** about salient parts of images. We asked humans to annotate image regions that are salient to their

decision about that image. Then we produced transformed versions of the original images, in which the degree of blur applied to a region is inversely related to its human-rated salience. The transformed images have full original clarity in regions salient to humans, and increasing levels of blur in regions rated increasingly less salient for humans. Such transformed images are then used as training data. In this way, the deep learning process is encouraged to learn whatever it can from the image regions rated salient for humans, but discouraged from learning from high-frequency details of regions rated less salient for humans, Fig. 1.

Various methods of experimental psychology have been already used to gauge limits of human perception [25, 33], understand how humans process visual information [5], build more explainable computer vision models [34], or strengthen feature extractors in iris recognition with human-driven way of processing images [10]. However, to our knowledge, this is the first-ever exploration of transforming the training data for deep learning based on human perception. This may initially seem a radical proposition. But we present two important advantages of this approach when applied to a quite difficult problem of presentation attack detection in biometrics, which always suffers from small and unrepresentative training corpora. First, comparing the accuracy achieved by training with the original image dataset versus the saliency-transformed version of that dataset, the saliency-transformed version achieves significantly higher accuracy. Second, evaluating the open-set scenario of training using all-except-one attack types and testing on the held-out attack type, the saliency-transformed training generates models with significantly higher accuracy on generalizing to the held-out attack type. It's noteworthy that the idea proposed in this paper is different from and independent of recent research on "attention" mechanisms [2, 3], and actually may complement them with human guidance.

This work opens up a new area of study for how to best incorporate human judgement about visual saliency into training data for deep learning, and makes the following **novel contributions**:

(a) a framework that **employs human intelligence to use effectively a low amount of training data** to increase the performance of deep learning models,

(b) demonstrated **greater generalization** of the method when applied to the iris presentation attack detection problem being solved by a deep learning classifier,

(c) all elements to reproduce this work made publicly available, including a **database of human-annotated iris images and sources codes**.

2. Related Work

Training data augmentation. Deep learning's voracious appetite for training datasets has been dealt with in multiple ways, and data augmentation techniques (such as flipping orientation of images, adding noise, using multiple crop offsets, and combination of those) are now a staple element of training [36, 9, 28]. These techniques can be applied to raw training data, subsets of training data within a batch [15], or in feature space instead of image space [19]. Another approach, benefiting from a renaissance of generative models, is to generate large amounts of synthetic images for use in training [41]. Recent trends are to mix synthetic and actual data [23], *e.g.* pre-train a model on a larger synthetically-generated corpus, and fine-tune on a smaller actual sample set [42], or guide the generation of synthetic data to make it more domain-specific [18].

To our knowledge, no previous work has applied human saliency encoding to augmentation techniques.

Combining human and algorithm capabilities when solving vision tasks has been studied in the context of biometrics. O'Toole *et al.* [31] demonstrated that fusing humans and algorithms increased face recognition accuracy to near perfect values for the Face Recognition Grand Challenge dataset. A few research groups concluded that human's and algorithm's visual saliencies differ and their integration increases the quality of image captioning [14] and of post-mortem iris recognition [30, 39]. Peterson *et al.* have even shown that humans' perceptual uncertainty may positively impact the generalization of deep learning-based models [32]. Human-machine pairing was also proposed to speed-up large-scale segmentation tasks, in which humans "correct" algorithm results, without a need to solve an entire segmentation task manually [4].

These past efforts demonstrate a boost in performance when machines and humans cooperate on the same task. However, again, we know of no previous work that has integrated human saliency judgements into training to increase generalization of the models.

Presentation Attack Detection (PAD) refers to determining the validity of an object presented to a biometric sensor, which – if not detected – could drive the system to an incorrect decision. While many iris PAD works show promising cross-dataset and cross-attack performance, generalization against true *unknown* attack types is an open research problem and a crucial aspect of deployable solutions [7]. Modern iris PAD approaches mainly employ deep-learning to achieve state-of-the-art accuracy, as evidenced in the recent LivDet-Iris 2020 competition [11]. In particular, Sharma and Ross [35] propose the application of DenseNet-121 [16] to iris PAD with a focus on interpretability. More recently, Chen and Ross proposed a novel method of attention-guided training using class activation mapping and attention modules [8]. They apply positional and channel attention modules to extract refined features from a DenseNet-121 backbone. Through experiments on public and private datasets, superior performance is demonstrated on both known and

Table 1: Number of samples in the **train**, **validation** and **test** partitions extracted from the superset of all available iris PAD datasets, broken by abnormality type. [P] denotes our proprietary data set.

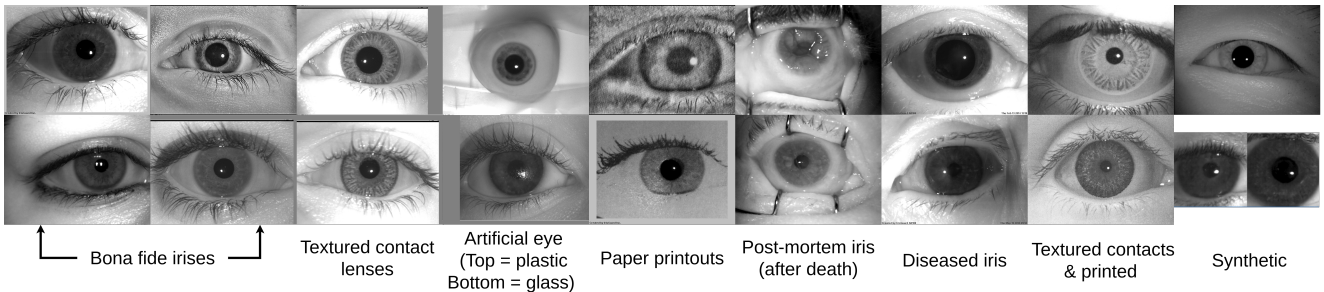| Bona fide	Artificial	Textured contact lenses	Display	Post mortem	Paper printouts	Synthetic	Diseased	Textured contact lenses & printed	Total
Train and validation partitions:									
399,053 [1] [24] [12] [38] [22] [20] [46] [43] [45] [P]	277 [24] [P]	27,372 [24] [20] [46] [45] [P]	×	2,259 [40]	16,393 [12] [24] [21]	10,000 [44]	1,537 [38]	1,899 [45]	**458,790**
Test partition (equivalent to LivDet-Iris 2020 benchmark) [11]:									
5,331	541	4,336	81	1,094	1,049	×	×	×	**12,432**

Figure 2: Examples of *bona fide* and *abnormal* samples from the acquired databases. The eight image types shown represent all attacks represented in the datasets. Each human annotator was presented with multiple examples of each type.

unknown presentation attack samples. The attention modules shift the focus of the networks to the iris regions rather than peripheral image features. Our work differs in that, instead of applying attention modules to learn network-defined salient features, the collected human-annotated features are used as predefined salient features and incorporated directly into the training process.

3. Experimental Datasets

Superset of Available Iris PAD Datasets. We made an effort to acquire all known (to us) public research iris PAD datasets, to have a full representation of known presentation attack instruments (PAI), ending up with more than 800,000 samples available. This collection was first cleaned by removing duplicated and not ISO-compliant [17] images. Second, it was split into **training and validation** comprising 458,790 samples in 8 categories (live + seven PAIs), and a disjoint **test set**, which is identical to the most recent LivDet-Iris competition benchmark [11] comprising 12,432 samples in six categories (live + 5 PAIs), as shown in Tab. 1. This test dataset was excluded from all training and validation processes, and was withheld solely for final testing. This allows for comparison with the results of the LivDet-Iris 2020 competition, as well as providing an independently collected dataset to assess the generalization capabilities of the proposed approach. Comprehensive descriptions of individual datasets can be found in the asso-

ciated papers listed in Table 1. Examples of PAIs available for this research, along with *bona fide* samples are illustrated in Figure 2, and individual dataset sample contribution statistics can be seen in the supplemental materials. In this work the term *abnormal* is assigned to the samples that differ from *bona fide (live)* samples including presentation attacks.

Human-Guided Region Saliency. To facilitate human data collection, an online annotation tool was developed. Subjects were presented 8 types of images: bona fide and 7 abnormal types, as presented in Figure 2. Participants were not specifically trained in iris PAD or iris recognition tasks, and were associated with the University of Notre Dame at the time of data collection.

On presentation of an iris image, users were asked to first select the type of image they believed it to be (one of eight types as above or *unsure*). Next, users were asked to highlight at least five regions of the image supporting their decision. The regions highlighted were not constrained on size or on location within the image. The objective was to collect data on what information present in an ISO-compliant iris image leads humans (non-experts) to a classification decision. There are two reasons for using non-experts: (1) there are no experts formally trained in iris image examination (such experts do exist in, *e.g.* fingerprint analysis); (2) to investigate whether a generalization boost can be obtained with help of non-experts in a given domain. The online an-

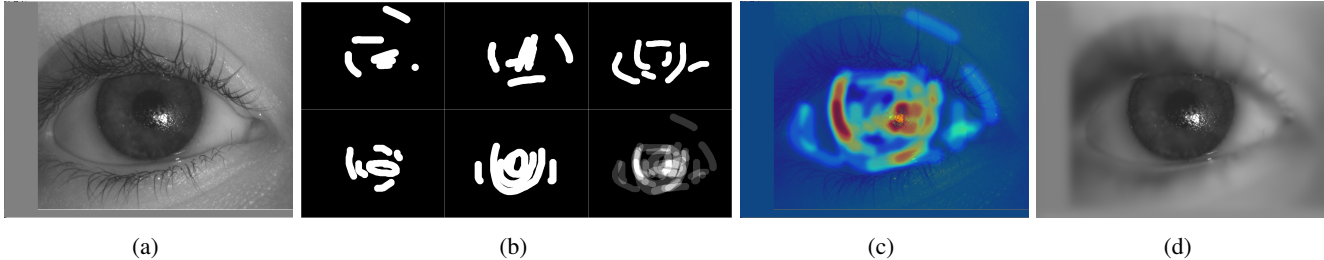| | (a) | (b) | (c) | (d) |

Figure 3: Creating human-guided training samples: the original image presented to human annotators (a); five individual correct annotations and a combination of those into the **saliency map** (b); a heatmap representation of the annotation density (c); the resulting training sample blurred locally with a magnitude (as defined in Sec. 4) inversely proportional to the annotation density (d).

notation tool, with green highlights corresponding to user annotated regions, is presented in supplementary materials.

Data from 150 subjects was collected in this experiment who annotated 30 unique image sets in total. Users were presented with an average of 27 images from one of the image sets. Image sets were assigned randomly to users and on average five subjects annotated each image. This simulates a scenario where some images have more annotation data available than others and hence the proposed approach must account for this imbalance. Only annotations from correctly classified samples were kept. As PAD is a binary classification problem, it was deemed a correct classification if the subject selected correctly either a bona fide samples, or marked any of the 7 abnormal types as abnormal.

Note that merely collecting more labeled samples (bonafide / abnormal) may (a) be impossible in the context of biometric attacks as these may be sparsely represented in datasets of ample size, and (b) not guide the network "where to look", opposite to the idea proposed in this paper. That is, the network, by simply observing more data, would still need to figure out relevant and irrelevant features without other guidance than the one through the loss function.

The annotations are next used to create image representations called **human saliency maps**, as shown in the bottom right image of Fig. 3(b). Each correct annotation, as shown in five individual plots in Figure 3(b), is weighted equally and combined. The closer the pixel is to white in the saliency map, the more subjects selected that area as supporting their decisions. Black regions correspond to areas in which no subject annotated interesting features. These saliency maps provide the basis to incorporate the data into the training process, as outlined in the next Section 4.

Classification accuracy from the human annotators is shown in Table 2. The most difficult type to classify correctly is textured contact lenses. This might be due to the fact that lens manufacturers design them to mimic genuine patterns. Conversely, the highest human classification accuracy (98%) was for post-mortem samples. The presence

Table 2: Human performance on the limited data used to construct saliency-encoded training images, measured in two scenarios: bona fide vs abnormal (independently of the abnormality type) and multi-class classification (indication of exact abnormality class was important).

Image Type	Accuracy (%)	
	Bona fide / abnormal	Exact type
Bona fide	56.3	56.3
Textured contact lenses	65.13	32.89
Paper printouts	94.27	64.53
Post-mortem	98.0	79.43
Synthetic	83.81	49.0
Artificial	70.19	34.82
Textured contact lenses printed	88.17	27.46
Diseased	77.32	47.73

of metal retractors to hold the eyelids open potentially contributes to this. Human annotators also performed well on paper printouts, possible due to well-visible pattern spread through the images. Interestingly, humans not trained in iris recognition, and classifying near-infrared iris samples, tend to have lower accuracy for bona fide samples.

4. Saliency-Encoded Training Images

Given a set of human-annotated iris samples, various levels of Gaussian filtering are applied to de-emphasize regions not marked as salient by humans. The intuition is that human annotators are able to restrict attention to regions relevant to the decision, as opposed to features of the training samples that may have incidental correlation with class labels. The magnitude of the blur (Gaussian kernel width σ) has a simple relation to how frequently a given region was annotated, and in particular regions selected by the largest number of annotators remain unchanged. We use blurring rather than binary masks due to (a) the need of gradual information suppression reflecting the human saliency, and (b) sharp edges around a binary mask could constitute "fake" image features that impact training. This approach mitigates these sharp edges between edges by ap-
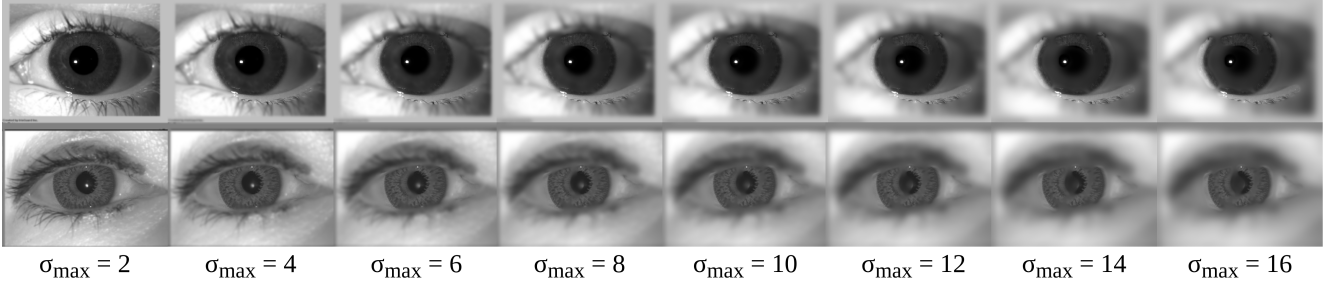
| $\sigma_{\max} = 2$ | $\sigma_{\max} = 4$ | $\sigma_{\max} = 6$ | $\sigma_{\max} = 8$ | $\sigma_{\max} = 10$ | $\sigma_{\max} = 12$ | $\sigma_{\max} = 14$ | $\sigma_{\max} = 16$ |

Figure 4: Illustration of annotations translating into saliency-encoded images for training. $\sigma_{\max}$ denotes the maximum blur is applied to unannotated regions. The top row is a bona fide iris and the bottom is a sample wearing a textured contact lens.

plying Gaussian blur of $\sigma = 5$ to the saliency map.

An important degree of freedom is the maximum strength of blurring $\sigma_{\max}$ applied to non-annotated regions, serving as baseline when calculating all remaining annotation map-dependent blur levels σ for a given sample. Instead of making arbitrary choices about $\sigma_{\max}$, we use a combination of $\sigma_{\max} \in \{2, 4, 6, 8, 10, 12, 14, 16\}$ as particular levels of blur "aggressivness" may make more sense for some abnormal types, or some values are better than others across all abnormal types. Thus, all annotated regions are blurred with a blur level σ based on a function of $\sigma_{\max}$ and the number of subjects that highlighted a specific region:

$$\sigma = (\sigma_{\max}(1 - \rho))^4 / \sigma_{\max}^3$$

where ρ is the fraction between 0 and 1 of annotators that selected a given image area. Thus, if zero subjects highlight a region ($\rho = 0$) then $\sigma = \sigma_{\max}$. And if all annotators working on a given image marked that region as important to them, then $\sigma = 0$ and these regions will be passed to the network unchanged. Figure 4 illustrates how the human annotation maps translate to the human-aided training data for various $\sigma_{\max}$. Increasing $\sigma_{\max}$ subjects unannotated regions to stronger blurring.

This straightforward mechanism of guiding the model to learn more human-like decisions has never been explored before. The next section will demonstrate its effectiveness in the case of limited training data, which is particularly relevant for biometric presentation attack detection.

5. Experiments

Setup. We selected D-NetPAD [35] as the most recent, open-source deep-learning-based iris PAD algorithm, demonstrating good results on the LivDet-Iris 2020 benchmark [11]. To ensure that the results presented in the evaluation section are the result of the application of human data, and not due to parameter optimization or modification to the method, no changes were made to the model parameters from the publicly available code. That is, the learning rate was set at 0.005, batch size was 20 and the number of

epochs was 50 for all experiments in this section. SGD with a momentum of 0.9 was used as the optimization algorithm and cross-entropy loss function was applied. No additional train time image augmentations were applied. All data is segmented (a SegNet-based method [40]), cropped, resized to 224×224 and used as input.

Leave-one-type-out experiments. Increased generalization means that the model can better classify new types of inputs unseen in training. This perfectly fits into biometric presentation attack detection, where – realistically – we cannot assume that only abnormal types present in the training data will be observed during testing. To evaluate effectiveness of the proposed method, seven "leave-one-abnormal-type-out" experiments are run. For each experiment, one abnormal type is omitted from both training and validation, and present only in testing. As this is a binary classification problem, bona fide samples (yet from disjoint sets) are used in all training, validation and test sets. In each of the 7 experiments, the left-out class represents an unknown type of abnormality for the model, and the performance on these left-out sets is indicative of the model's generalization capability. Because the *textured contacts & printout* type contains information about both contacts and printouts, these samples were excluded from training and validation for experiments with the *textured contacts* and *paper printouts*, thus maintaining complete model ignorance of the nature of the test data. Similarly, *textured contacts* and *paper printouts* were excluded from the experiment where *textured contacts & printouts* was the held-out test set. The models trained in the leave-one-abnormal-type-out scenario are then tested on unseen bona fide samples taken from the LivDet-Iris 2020 dataset, and unseen abnormal samples originating from the large dataset. Taking bona fide samples from LivDet-Iris 2020 corpus maximizes the "open-setness" of these experiments, as these bona fide samples were acquired independently of all the training and validation data shown in Table 1.

Evaluation scenario S1: Training on the Large Dataset. In this scenario, models are trained on the unmodified train-
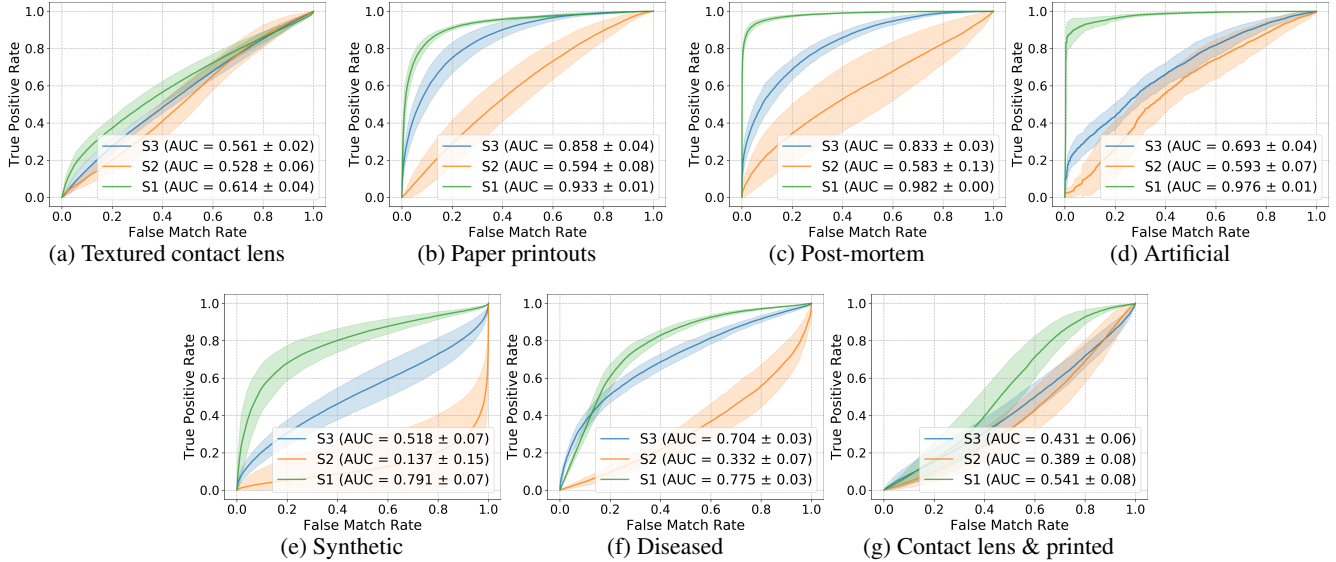
Figure 5: ROC curves in leave-one-abnormal-type-out experiments obtained in three training scenarios. The left-out type is named in the plot caption. Shaded bands represent ±1 standard deviation along the TPR axis obtained in repeated experiments. **S1** refers to scenario 1 using the large dataset in training. **S2** refers to limited training dataset with **no** human salience info incorporated. **S3** refers to the training with human-salience-encoded versions of the same original images as **S2**. In all cases, there is a significant performance increase when human salience info is used in training (compare **S3** with **S2**).

ing data described in Table 1, starting from pre-trained ImageNet weights, as common practice suggests [6]. Before model training, all samples that have corresponding correct human annotations (765 in total) are removed from the training and validation sets. This means the human-annotated data used in other scenarios was *unseen*. There are many more bona fide samples available than abnormal samples, as shown in Tab. 1. To prevent the over-representation of one class in training, all abnormal samples are taken along with a randomly sampled set of bona fide samples equal to the number of abnormal samples. Models trained in this scenario are referred to as the **large data models**, achievable in the situation of having a "large-enough" training dataset, which usually is not the case in biometric presentation attack detection.

Evaluation scenario S2: Training on Regular Limited Data. In this scenario, the same ImageNet-initiated models are tuned with 765 live and abnormal samples minus the samples corresponding to the left-out type, for which correct human annotations were collected (but **not** used here). Additionally, these images are augmented by applying a Gaussian blur $\sigma \in \{2, 4, 6, 8, 10, 12, 14, 16\}$ to the entire image and combined with the unblurred versions of the samples, increasing the size of the dataset nine-fold. This is to ensure that the results obtained for human-annotated training data (scenario S3 below) are due to the model learning human-aided features, not the addition of the global blur as image augmentations. This scenario simulates a situation

where limited dataset representing a given domain is available, and human-aided augmentations are not used. Models trained according to this scenario will be referred to as the **limited data models**.

Evaluation scenario S3: Training on Human-Aided Limited Data. The same training as scenario S2 is applied, but using human saliency encoded versions of all 765 training samples, again minus samples corresponding to the left-out type. Networks are trained on a combined set comprised of images transformed with all maximum blur levels, as described in Sec. 4. Samples with no blur are not included in this scenario as they were in S2. Models trained according to this scenario are referred to as the **human-aided models**.

Common settings across all scenarios. The validation set used for best epoch selection was the same in all scenarios, and the validation images were not blurred. Since all models accept cropped and resized images as input, both original and the user-annotated images were cropped in the same way. As with the large training dataset, the common validation set is also balanced such that the number of bona fide and abnormal samples is equal.

6. Evaluation Results

Improvement through human-saliency-aided training. Since the evaluation is focused to assess generalization capabilities, all samples of one abnormal type are left out in each experiment. Each plot in Fig. 5 represents one leave-

Table 3: LivDet-Iris 2020 competition results (in %) compared to the proposed approaches.

Method category	Algorithm	APCER					Overall Performance		ACER
		PP	CL	DP	AR	PM	APCER$_{average}$	BPCER	
Livet Iris 2020 Submissions	Team: USACH/TOC	23.64	66.01	9.87	25.69	86.10	59.10	0.46	29.78
	Team: FraunhoferIGD	14.87	72.80	53.08	19.04	0	48.68	11.59	30.14
	Competitor-3	72.64	43.68	83.95	73.19	89.85	57.8	40.31	49.06
This Work	Large Training Data (S1)	9.34	32.89	3.70	2.03	0.55	21.74	0.47	11.1
	Limited Training Data (S2)	1.91	5.05	1.23	3.512	0.09	3.66	93.08	48.37
	Human-Aided (S3)	9.06	36.65	0.0	2.77	1.37	24.14	8.61	**16.37**

PP: Paper printouts; **CL:** Textured contacts ; **DP:** Display; **AR:** Artificial; **PM:** Post-mortem; **APCER:** Attack Presentation Classification Error Rate (abnormal called bona fide); **APCER$_{average}$:** APCER averaged across all attack types; **BPCER:** Bonafide Presentation Classification Error Rate (bona fide called abnormal); **ACER:** average of BPCER and APCER$_{average}$.
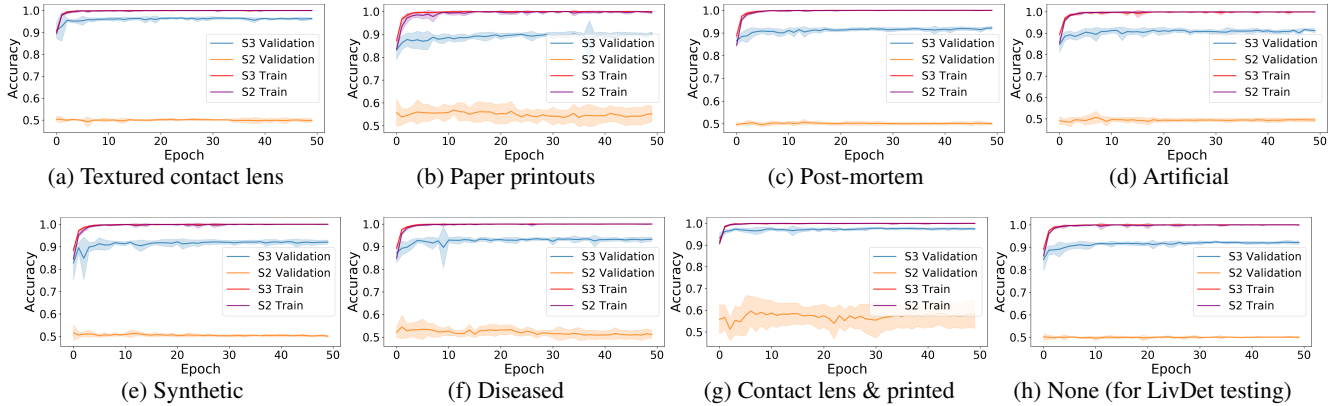

(a) Textured contact lens   (b) Paper printouts   (c) Post-mortem   (d) Artificial

(e) Synthetic   (f) Diseased   (g) Contact lens & printed   (h) None (for LivDet testing)

Figure 6: Accuracy on the training and validation subsets while training the models without (scenario S2) and with (S3) human saliency maps encoded into the training data. We can observe a severe overfit, and a close-to-random-chance performance of models trained without employing human intelligence (S2). In turn, human-aided saliency maps help in obtaining high validation accuracy almost instantly during training (S3).

one-out experiment, with the class that was held out indicated in the graph's title. For all experiments, each network initiated from pre-trained ImageNet weights was trained ten times independently. The ROC curves represent the means of ten runs in a given scenario. The shaded region surrounding each curve represents ±1 standard deviation of the True Positive Rate (TPR) obtained on the test set.

In all graphs, a similar observation can be made. The *large data* models, trained with the largest-possible dataset (scenario S1), results in the highest accuracy. That is, domain knowledge about abnormality delivered by dataset of ample size allows for more accurate class representations to be learned. Models trained with a limited number of samples, and with no human-aided augmentation (scenario S2) have lowest accuracy for each left out abnormal type. This clearly demonstrates that such limited training set, although non-trivial to collect if we think about collecting more than 700 exemplars of real attacks on biometric systems, is not sufficient to build an effective PAD system.

Can we do better with this limited data? The ROC plots show that the use of human-guided features (scenario S3) provides a significant increase in accuracy, when compared

to scenario S2, across all experiments. This shows that using human judgement about which parts of the image contain information that is salient to the decision guides networks to learn solutions that generalize better. While only trained and tested in the iris PAD context, this approach does not consider any biometrics-specific training or annotations, hence we hypothesise that a similar approach can be applied to a wide family of visual tasks, in which humans present better-than-random-chance classification accuracy. It is important to note that direct comparison of the large data (scenario S1) models with human-aided (scenario S3) models would not be fair due to significant discrepancies in the training set size.

Varying performance can be seen across the left-out abnormal types. Abnormal types such as textured contacts and synthetic are most difficult due to their intentional resemblance to bona fide samples. When these types are removed from training and validation, the models struggle to make effective classifications on these samples. This is a common observation in iris PAD, as shown in results of the LivDet-Iris 2020 competition. However, in all cases the performance is significantly increased when human annota-

tions are incorporated into training.

Testing on LivDet-Iris 2020 Competition data. The final evaluation demonstrates how the proposed human-aided training strategy performs on the most recent iris PAD competition. This competition included abnormal types unknown from previously published works, and hence is inherently focused on assessing generalization capabilities of iris PAD algorithms. In this evaluation, we again test models trained according to three scenarios (S1, S2 and S3) and test on the previously unseen LivDet-Iris 2020 dataset using the classification threshold defined by the competition organizers of 0.5, where 0 is bona fide and 1 is abnormal. For this experiment, no abnormal type from the training data was left out. The competition protocol is applied such that we assume no knowledge of the testing set during training.

As shown in Table 1 there are five abnormal image types present in the LivDet dataset, and also bona fide data. One of the abnormal image types (display attack) was not present in the training data. The other four abnormal image types were represented in the training with disjoint samples, but were collected by different teams, with different subjects, and were excluded from any training. To ensure fairness, no modifications were made to the algorithm to improve results after attaining performance metrics on the LivDet-Iris 2020 benchmark. All baselines were disqualified from the competition by LivDet organizers since these institutions had access to this test data which originated from the same source as their train data, unlike the competitors. Hence, it was decided to strictly follow the LivDet competition protocol and compare only to the competitors.

As shown in Table 3, the highest accuracy is again the "large data" model (scenario S1). This model improves upon the best results in the competition by decreasing the average error rate from the winning level of 29.78% to a much lower 11.1%. It is certainly expected, as the large and balanced training set, composed of 93,190 (46,595 bona fide/46,595 abnormal) samples contains the richest information about the task domain. However, the most interesting comparison is to juxtapose the "limited data" and "human-aided" approaches. We can see that using only 765 training images encoded with human saliency maps (scenario S3), the average classification error rate (ACER) decreased from 29.78% (obtained by the competition winner) to 16.37%. In contrast, when using the same training images but without human-aided training (scenario S2) the obtained ACER = 48.37%, which is worse than the top two LivDet submissions and only marginally better than the last place submission. This suggests that **the human guided approach (S3) would have won the LivDet-Iris 2020 competition** by a large margin of 13% whereas an equivalent solution without human annotation incorporation (S2) would have placed third.

Validation accuracy increases significantly using models trained with human-aided saliency maps. Figure 6 outlines an interesting finding when analysing validation accuracies as the training progresses in both S2 and S3 scenarios. While the training accuracy in both scenarios stays practically identical, it is clear that regular training (S2) leads to a severe overfit to the training data, as performance on the validation data is close to random chance and plateaus quickly. Conversely, the validation accuracy on the same set when the models are trained with human saliency encoded into the training data (S3) is significantly better, and can be achieved rapidly during the training process. It's yet another demonstration that the addition of human saliency information to the training data reduces overfitting, while increasing performance over models trained with the same images without human saliency information encoded.

7. Summary and Conclusions

We propose a novel framework for incorporating human saliency judgements into the training images for deep learning, with the goal of increasing accuracy, especially from limited training data, and improving generalization. We validated our approach on the difficult problem of iris presentation attack detection. Iris PAD is an excellent example of a security-critical task that inherently has limited training data available, and for which generalization is of paramount importance. Adversaries create an ever-evolving landscape of attack samples, and algorithms must be able to generalize to novel attacks with only a handful of new images.

We collected a unique dataset of human annotations from 150 non-expert subjects who highlighted salient regions for this visual classification task. These annotations fed into localized blurring of image regions judged as less salient for humans. The result is a human-saliency-transformed version of the original training images. We experimentally compared the performance of the model trained with (a) human-saliency-transformed data, and (b) original images (with standard augmentation techniques). Training with human-saliency-transformed images achieves significantly greater accuracy on all leave-one-attack-type-out experiments. Further, evaluating our approach in the framework of the state-of-the-art LivDet-Iris 2020 competition, our human-saliency-transformed model achieves average classification error rate of 16.37% on the challenging LivDet-Iris 2020 benchmark, a substantial improvement over the competition winner's ACER = 29.78%.

No part of our approach is specific to iris PAD and it can readily be applied to any visual classification task recognizable to humans. This opens a whole new area of research related to effective incorporation of human saliency judgements into training strategies for deep learning. We release our human annotation dataset collected for this work along with source codes to prepare human-aided training data.

References

[1] Chinese academy of sciences institute of automation. http://www.cbsr.ia.ac.cn/china/Iris%20Databases%20CH.asp. Accessed: 03-12-2021. 3, 11

[2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. Jan. 2015. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. 2

[3] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. Attention augmented convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294, 2019. 2

[4] R. Benenson, S. Popov, and V. Ferrari. Large-scale interactive object segmentation with human annotators. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11692–11701, 2019. 2

[5] David Berga. Understanding eye movements: Psychophysics and a model of primary visual cortex. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 18(2):13–15, 2020. 2

[6] Aidan Boyd, Adam Czajka, and Kevin Bowyer. Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch? In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2019. 6

[7] Aidan Boyd, Zhaoyuan Fang, Adam Czajka, and Kevin W. Bowyer. Iris presentation attack detection: Where are we now? *Pattern Recognition Letters*, 138:483–489, 2020. 2

[8] Cunjian Chen and A. Ross. An explainable attention-guided iris presentation attack detector. In *Workshop on Explainable & Interpretable Artificial Intelligence for Biometrics (xAI4Biometrics) at the IEEE Winter Conference on Applications of Computer Vision (WACV)*, January 2021. 2

[9] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019. 2

[10] A. Czajka, D. Moreira, K. Bowyer, and P. Flynn. Domain-specific human-inspired binarized statistical image features for iris recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 959–967, 2019. 2

[11] 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. 2, 3, 5

[12] Javier Galbally, Jaime Ortiz-Lopez, Julian Fierrez, and Javier 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. 3, 11

[13] Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189, 2019. 1

[14] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault. Human attention in image captioning: Dataset and analysis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8528–8537, 2019. 2

[15] E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry. Augment your batch: Improving generalization through instance repetition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8126–8135, 2020. 2

[16] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[17] ISO/IEC 19794-6:2011. Information technology – Biometric data interchange formats – Part 6: Iris image data, 2011. 3

[18] A. Kar, A. Prakash, M. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler. Meta-sim: Learning to generate synthetic datasets. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4550–4559, 2019. 2

[19] B. Ko and G. Gu. Embedding expansion: Augmentation in embedding space for deep metric learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7253–7262, 2020. 2

[20] Naman Kohli, Daksha Yadav, Mayank Vatsa, and Richa Singh. Revisiting iris recognition with color cosmetic contact lenses. In *IEEE Int. Conf. on Biometrics (ICB)*, pages 1–7, Madrid, Spain, June 2013. IEEE. 3, 11

[21] Naman Kohli, Daksha Yadav, Mayank Vatsa, Richa Singh, and Afzel 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. 3, 11

[22] Oleg Komogortsev. Eye Tracker Print-Attack Database (ET-PAD) v2, 2014. 3, 11

[23] P. N. V. R. Koutilya, H. Zhou, and D. Jacobs. Sharingan: Combining synthetic and real data for unsupervised geometry estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13971–13980, 2020. 2

[24] Sung Joo Lee, Kang Ryoung Park, Youn Joo Lee, Kwanghyuk Bae, and Jai Hie Kim. Multifeature-based fake iris detection method. *Optical Engineering*, 46(12):1 – 10, 2007. 3, 11

[25] Fei-Fei Li, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(10), 2007. 2

[26] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra, Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60—88, 2017. 1

[27] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning

for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318, Feb 2020. 1

[28] C. Luo, Y. Zhu, L. Jin, and Y. Wang. Learn to augment: Joint data augmentation and network optimization for text recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13743–13752, 2020. 2

[29] I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478, 2018. 1

[30] D. Moreira, M. Trokielewicz, A. Czajka, K. Bowyer, and P. Flynn. Performance of humans in iris recognition: The impact of iris condition and annotation-driven verification. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 941–949, 2019. 2

[31] A. J. O'Toole, H. Abdi, F. Jiang, and P. J. Phillips. Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1149–1155, 2007. 2

[32] J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625, 2019. 2

[33] B. Richard Webster, S. E. Anthony, and W. J. Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, 2019. 2

[34] Brandon RichardWebster, So Yon Kwon, Christopher Clarizio, Samuel E. Anthony, and W. Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *European Conference on Computer Vision (ECCV)*, pages 263–281, 2018. 2

[35] R. Sharma and A. Ross. D-NetPAD: An Explainable and Interpretable Iris Presentation Attack Detector. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020. 2, 5

[36] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019. 2

[37] Kalaivani Sundararajan and Damon L. Woodard. Deep learning for biometrics: A survey. *ACM Comput. Surv.*, 51(3), May 2018. 1

[38] 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. 3, 11

[39] M. Trokielewicz, A. Czajka, and P. Maciejewicz. Perception of image features in post-mortem iris recognition: Humans vs machines. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2019. 2

[40] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Post-mortem iris recognition with deep-learning-based image segmentation. *Image and Vision Computing*, 94:103866, 2020. 3, 5, 11

[41] A. Tsirikoglou, G. Eilertsen, and J. Unger. A survey of image synthesis methods for visual machine learning. *Computer Graphics Forum*, 39(6):426–451, 2020. 2

[42] Q. Wang, J. Gao, W. Lin, and Y. Yuan. Learning from synthetic data for crowd counting in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8190–8199, 2019. 2

[43] Warsaw University of Technology. Warsaw Datasets Webpage. http://zbum.ia.pw.edu.pl/EN/node/46, 2013. 3

[44] Zhuoshi Wei, Tieniu Tan, and Zhenan 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. 3, 11

[45] David Yambay, Benedict Becker, Naman Kohli, Daksha Yadav, Adam Czajka, Kevin W. Bowyer, Stephanie Schuckers, Richa Singh, Mayank Vatsa, Afzel Noore, Diego Gragnaniello, C. Sansone, L. Verdoliva, Lingxiao He, Yiwei Ru, Haiqing Li, Nianfeng Liu, Zhenan Sun, and Tieniu 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. 3, 11

[46] David Yambay, Brian Walczak, Stephanie Schuckers, and Adam 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. 3, 11

A. Detailed Dataset Description

Table 4: Full dataset used for **training and validation** broken down by individual contributing dataset. The in-house currently unpublished data used in this work is denoted as *University of Notre Dame data*.

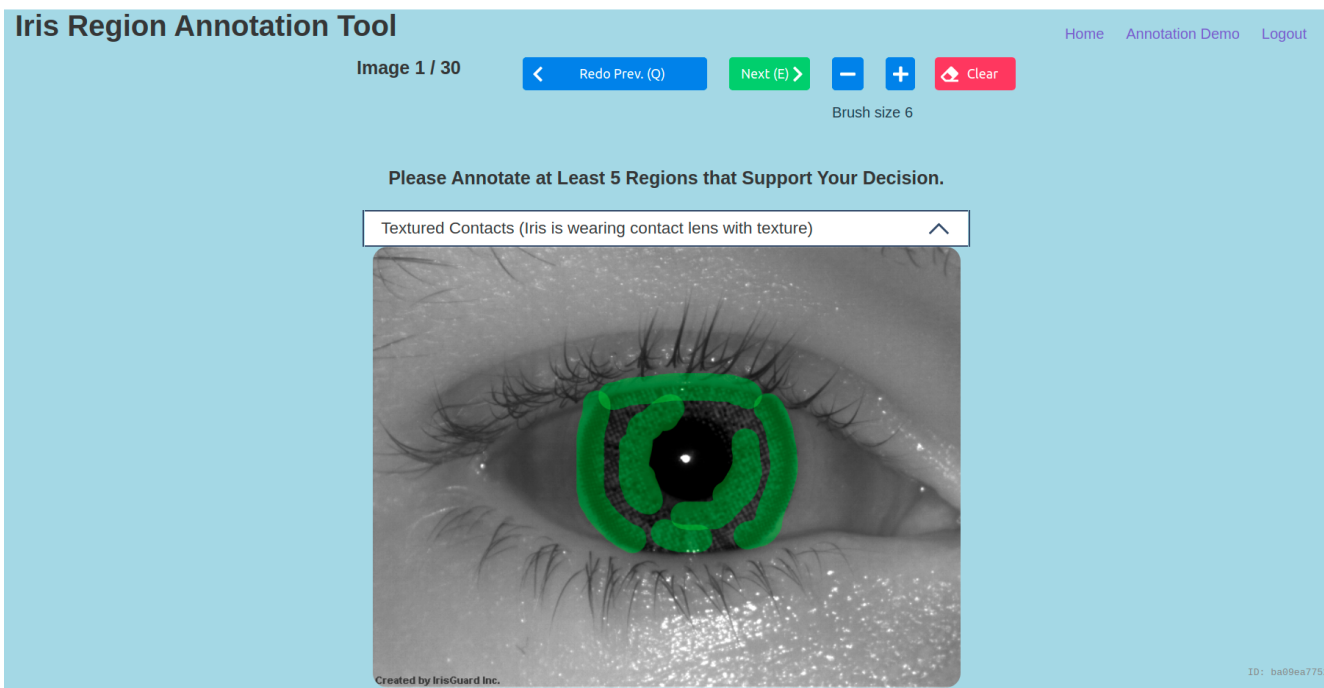| Image Type	Contributing Dataset	# of Samples	Total Samples
Bona fide	ATVS-FIr [12]	800	**399,053**
	BERC_IRIS_FAKE [24]	2,776	
	CASIA-Iris-Thousand [1]	19,952	
	CASIA-Iris-Twins [1]	3,181	
	Disease-Iris v2.1 [38]	255	
	ETPAD v2 [22]	400	
	IIITD Contact Lens Iris [20]	13	
	IIITD Combined Spoofing Database [21]	4,531	
	LivDet-Iris Clarkson 2015 [46]	813	
	LivDet-Iris Warsaw 2015 [46]	36	
	LivDet-Iris Clarkson 2017 [45]	3,949	
	LivDet-Iris IIITD-WVU 2017 [45]	2,944	
	LivDet-Iris Warsaw 2017 [45]	5,167	
	University of Notre Dame data	354,236	
Textured contact lens	BERC_IRIS_FAKE [24]	140	**27,372**
	IIITD Contact Lens Iris [20]	3,420	
	LivDet-Iris Clarkson 2015 [46]	1,107	
	LivDet-Iris Clarkson 2017 [45]	1,881	
	LivDet-Iris IIITD-WVU 2017 [45]	1,700	
	University of Notre Dame data	19,124	
Paper printouts	ATVS-FIr [12]	800	**16,393**
	BERC_IRIS_FAKE [24]	1,600	
	IIITD Combined Spoofing Database [21]	1,371	
	LivDet-Iris Clarkson 2015 [46]	1,745	
	LivDet-Iris Warsaw 2015 [46]	20	
	LivDet-Iris Clarkson 2017 [45]	2,250	
	LivDet-Iris IIITD-WVU 2017 [45]	1,766	
	LivDet-Iris Warsaw 2017 [45]	6,841	
Post-mortem Irises	Post-Mortem-Iris v3.0 [40]	2,259	**2,259**
Synthetic	CASIA-Iris-Syn V4 [44]	10,000	**10,000**
Artificial	BERC_IRIS_FAKE [24]	80	**277**
	University of Notre Dame data	197	
Diseased irises	Disease-Iris v2.1 [38]	1,537	**1,537**
Textured contacts & printed	LivDet-Iris IIITD-WVU 2017 [45]	1,899	**1,899**

B. Annotation Tool

Figure 7: Online annotation tool developed to collect annotation data with an example input of a human solving the iris presentation attack detection task for a textured contact lens sample.