Teaching AI to Teach: Leveraging Limited Human Salience Data Into Unlimited Saliency-Based Training

Colton R. Crum
ccrum@nd.edu
Aidan Boyd
aboyd3@nd.edu
Kevin W. Bowyer
kwb@nd.edu
Adam Czajka
aczajka@nd.edu

University of Notre Dame
Notre Dame, IN 46556, USA

Machine learning models have shown increased accuracy in classification tasks when the training process incorporates human perceptual information. However, a challenge in training human-guided models is the cost associated with collecting image annotations for human salience. Collecting annotation data for all images in a large training set can be prohibitively expensive. In this work, we utilize "teacher" models (trained on a small amount of human-annotated data) to annotate additional data by means of teacher models' saliency maps. Then, "student" models are trained using the larger amount of annotated training data. This approach makes it possible to supplement a limited number of *human-supplied* annotations with an arbitrarily large number of *model-generated* image annotations. We compare the accuracy achieved by our teacher-student training paradigm with (1) training using all available human salience annotations, and (2) using all available training data without human salience annotations. We use synthetic face detection and fake iris detection as example challenging problems, and report results across four model architectures (DenseNet, ResNet, Xception, and Inception), and two saliency estimation methods (CAM and RISE). Results show that our teacher-student training paradigm results in models that significantly exceed the performance of both baselines, demonstrating that our approach can usefully leverage a small amount of human annotations to generate salience maps for an arbitrary amount of additional training data.

1 Introduction

Computer vision architectures often take inspiration from brain physiology, mental models, and attention mechanisms, which can be incorporated into the training of models in many different ways. A common way to train human-guided models is through saliency-based training, which has shown to (a) generalize better to new data, which is vital in open-set recognition where not all classes are known, (b) speed up training time by using less samples that contain more meaningful information (data + associated saliency), (c) increase the model's focus on class features, limiting sensitivity to features accidentally correlated with class labels, and (d) produce more human explainable outputs. One limitation of human-guided models can be the high cost to acquire human perception-related information, such as image annotations. One potential solution to address this limitation is to build models that are capable of generating human-like saliency maps to annotate new data used to train subsequent models.

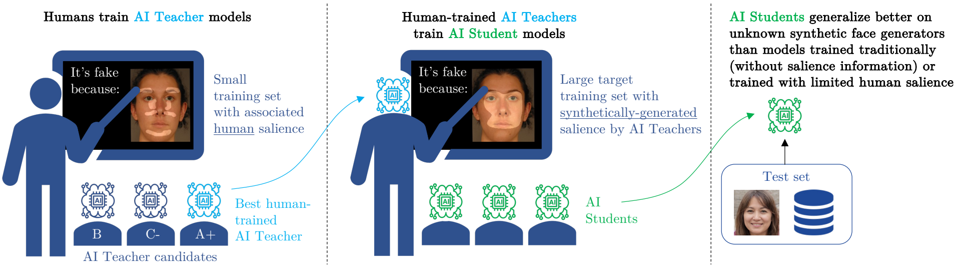

Humans train AI Teacher models
It's fake because:
Small training set with associated human salience
B
C-
A+
AI Teacher candidates
Best human-trained AI Teacher
Human-trained AI Teachers train AI Student models
It's fake because:
Large target training set with synthetically-generated salience by AI Teachers
AI Students
AI Students generalize better on unknown synthetic face generators than models trained traditionally (without salience information) or trained with limited human salience
Test set

Figure 1: This work explores training *AI Teachers* – models first trained in a human-guided way – to train future *AI Student* models. AI Teachers are selected from models that have the highest Area Under the ROC Curve (AUC) on the validation set and provide saliency for larger, unannotated training data, eventually used to train AI Students. With this approach, always-limited human annotations are efficiently leveraged to provide salience data to an unlimited number of training samples.

We explore a training framework which first uses the available human-salience data to train an *AI Teacher* model, which is then used to generate saliency maps, similar to human saliency, for large amounts of additional training data (see Fig. 1). The *AI Student* model is then trained using the training data annotated by the AI Teacher. The AI Teacher's saliency maps can be generated using white-box approaches (e.g., CAM [33]), or black-box approaches that rely on perturbing the input image and observing the effect on the output (*e.g.*, RISE [26]).

We experiment with our teacher-student framework for synthetic face detection using the CYBORG human-guided training paradigm [5]. However, our approach is applicable to any task for which humans can provide salience, and we present its viability also for fake iris detection.

We show that the performance of AI Student models, trained by the human-taught AI Teacher, surpasses the performance of both (a) models trained with limited human salience (Baseline 1), and (b) models trained without human salience but on large training data (Baseline 2). Thus, the proposed approach provides a means to efficiently convert a small amount of human-provided salience data into a large amount of effective human-like saliency. Our framework allows for increased data diversity and new information for each training sample, which exceeds performance rather than simply adding more data. Results in this paper are organized around the following **research questions**:

- **RQ1:** Which type of training produces better AI Teacher models: human-guided or purely data-driven? We consider four CNN architectures with the same architecture for teacher and student models. (*Sec. 4 and Tab. 1*)

- **RQ2:** Can the top-performing AI Teacher model improve the performance of AI Student models across different CNN architectures? (*Sec. 4 and Fig. 3*)
- **RQ3:** What are the potential performance benefits of the teacher-student training paradigm over the baselines? (*Sec. 4 and Tab. 2*)
- **RQ4:** Can this training approach be applied to domains beyond synthetic face detection? (*Sec. 4, Tab. 3.*)

2 Related Work

Estimating Model Salience Access to models' internal data (feature maps, gradients, weights) simplifies building saliency estimation methods. Class Activation Mapping (CAM) is the most popular approach to estimate salience of white-box models [33]. CAM works by making a forward pass through the model to get the activations of the last convolutional layer, which are weighted into a heat map. A potential downside of CAM is low resolution of the resulting visualization; *e.g.*, 7 × 7 for DenseNet. Recent advances such as Grad-CAM [30], Grad-CAM++ [8], HiResCAM [13], Score-CAM [32], Ablation-CAM [28], or Eigen-CAM [25] aim to provide more detailed saliency estimations, but require more computational resources.

In case of black-box models, dominant methods rely on a simple idea of randomly perturbing input regions and observing the impact on the output. Random Input Sampling Explanation (RISE) [26] is one such method, in which a weighted average (where the weights serves as the "confidence" scores) is used to generate a full-sized salience map. Black-box approaches, such as RISE, have two main benefits; (1) they require no information from inside the model, and (2) the resolution of generated salience may be as high as that of the input image, whereas CAMs are limited to the spatial dimensions of the last convolutional layer. Recent work on black-box explainers include methods of evaluating their usefulness for humans [6] and increasing their robustness against adversarial attacks [7].

The above mentioned techniques are part of the broader and dynamic "eXplainable AI" (XAI) area [29]. In this work, we use CAM and RISE to compare their usefulness in generating salience of teacher models, since saliency methods can be architecture-specific and may impact the performance of the Teacher-Student training paradigm.

Human Salience-Guided Model Training Incorporating human perceptual capabilities into the model training is non-trivial, and may involve human-sourced information in various forms: image/video annotations [5], eye-tracking [4, 11], reaction times [17], or even games [24]. Successful ways of incorporating human-collected information into training include adding specialized components to the loss functions [5, 17], augmenting training data [3], and introduction of human perception-based regularization [14, 17]. Specifically, ConveYing Brain Oversight to Raise Generalization (CYBORG) training strategy [5] combines both human and model's salience into the loss function by penalizing the divergence of the model's CAM from the human salience provided as image annotations. Application of the CYBORG loss function increased the performance in synthetic face detection across four, out-of-the-box CNN architectures using only 1,821 training samples with associated human annotations, compared to models trained with cross-entropy loss.

Although human salience-based training has been successfully implemented (*e.g.* CYBORG), to our knowledge the approaches that would enable more effective use of human annotations, and thus scale the human-guided training, have not yet been explored. We find this an important and essential research direction in the future of training human-guided models for a number of applications. This paper shows how to create "proxy" models, which by producing human-like salience for any input, allow human-inspired annotations to train models on without additional cost.

3 Methodology

In this section, we first describe the two baselines that we compare our training paradigm against. Secondly, we describe our dataset splits to train AI Teachers (TAIT), train AI Students (TAIS) and evaluate AI Students (EAIS). We then describe our method for creating representative human-guided AI Teachers, including experimental design parameters and teacher model selection. Finally, we describe the performance metrics used to evaluate research questions RQ1-RQ4.

3.1 Baseline Models

We benchmark our Teacher-Student training paradigm against two baselines. Baseline 1 is to simply train human-guided models using the available human annotations, which was previously proposed in [5]. Within the naming conventions of this paper, Baseline 1 can be thought of as simply using the teacher models on the test set, without training and using any student models. Baseline 2 is to train traditional (no saliency) models on all available training data. These two baselines represent contrasting viewpoints in achieving optimal model performance: (a) giving the model human-guided information on "where to look" in order to solve the task (Baseline 1), or (b) giving the model a large amount of data to train on (Baseline 2). For some tasks and domains, Baseline 1 or 2 may achieve the desired performance. However, for Baseline 1, the vast majority of training data remains un-annotated and is not used. And for Baseline 2, the hope is that "enough" training data has been used to train an optimal model. Our approach is a strategic blend of using human-generated salience for whatever fraction of training data it is available, and using model-generated salience for all remaining training data.

3.2 Datasets

This section presents three dataset splits that play essential roles in the proposed pipeline (cf. Fig. 2). First, the AI Teachers are trained on a (potentially small) training dataset with saliency maps sourced from human annotations. Next, the best AI Teacher is used to generate synthetic (yet human-like) saliency maps for samples in target (potentially large) training dataset. This creates AI Students, which generalize better to unknown samples in the test set, compared to (a) student models trained without any saliency, and (b) student models trained with synthetic saliency, but generated by teacher models trained without human perceptual inputs.

To evaluate our training framework, we use the task of synthetic face detection. Selected results are repeated also for iris presentation attack detection (PAD) to address the generalization capability of the proposed approach across domains.

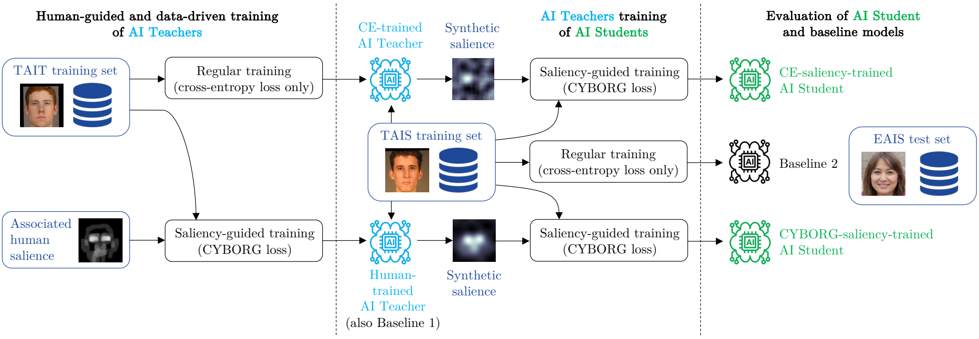

Human-guided and data-driven training of AI Teachers
TAIT training set
Regular training (cross-entropy loss only)
Associated human salience
Saliency-guided training (CYBORG loss)
CE-trained AI Teacher
Synthetic salience
TAIS training set
Human-trained AI Teacher (also Baseline 1)
Synthetic salience
AI Teachers training of AI Students
Saliency-guided training (CYBORG loss)
Regular training (cross-entropy loss only)
Saliency-guided training (CYBORG loss)
Evaluation of AI Student and baseline models
CE-saliency-trained AI Student
Baseline 2
EAIS test set
CYBORG-saliency-trained AI Student

Figure 2: Detailed diagram of data and saliency usage.

Dataset for Training AI Teachers (TAIT) These are the subsets of the overall training sets that have human-provided annotations of salience, and are used to **T**rain the **AI** **T**eachers. For the human-guided training of AI Teachers in the **synthetic face detection** task, we use the same dataset and human annotations as introduced in [5][1]. The training set consists of 919 authentic and 902 synthetic face images with annotations of regions selected by 363 humans (recruited via Amazon Mechanical Turk) as important to them judging the authenticity of a given face image. Only annotations for correctly classified image pairs are used for the training set. The validation set is composed of 10,000 authentic faces (sampled from FRGC-Subset [27]) and 10,000 synthetic faces (sampled from SREFI [2] and from StyleGAN2-generated images at *thispersondoesnotexist.com*). All images were preprocessed using img2pose [1], resized to 224 × 224, and cropped to ensure the face is in full view. Subjects were presented with a pair of face images, one authentic and one synthetic, and asked an alternating prompt of which face is real (fake). After answering the question, subjects used their cursor to highlight regions of the selected face that support their decision. The human salience maps were cropped and resized to 224 × 224 in the same way as the image data to match the corresponding input images.

For training **iris PAD** AI Teachers, we used 765 samples annotated by humans, offered with [3], including bona fide irises, and seven spoof attack types (artificial, textured contact lens, post mortem, paper print outs, synthetic, diseased, textured contact lens & printed). Only correctly classified samples were used in model training.

The TAIT image sets, without the human-salience heatmaps, are also used to train models with cross-entropy loss in order to answer Research Question 1.

Datasets for Training AI Students (TAIS) Larger training sets, for which no human annotations of salience are available, were used to **T**rain the **AI** **S**tudents. For **synthetic face detection** task, the dataset was collected from the same sources as the TAIT (FRGC, SREFI and StyleGAN2; see example images added to the supplementary materials). This resulted in a TAIS dataset six times larger than, and image-disjoint from, the smaller TAIT dataset for synthetic face detection task. The TAIS dataset for **iris PAD** task was collected from the same sources as TAIT for iris PAD, and is certainly image-disjoint. However, due to a more challenging scenario of collecting physical iris spoofs, we kept the size of TAIS similar

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

as the size of TAIT except so that classes (live / spoof) could be completely balanced (764 samples).

Both TAIS datasets have no associated human saliency for the images. Instead, salient regions for each image are given by a salience map (either either CAM- or RISE-based) generated from the AI Teacher model that was trained using TAIT.

Datasets for Evaluating AI Students (EAIS) In conventional machine learning terms, these are the test sets. However, we are careful to make this distinction as the teacher models are completely withheld from the test set, and only the student models are evaluated on this set. Instead, teacher models are assessed by their performance on the validation set (Sec. 3.4 discusses the teacher model selection process in depth). For **synthetic face detection** task, the EAIS set contains (a) 600,000 synthetic face images, evenly sampled from six GAN architectures (ProGAN [19], StarGANv2 [9], StyleGAN [20], StyleGAN2 [22], StyleGAN2-ADA [21], and StyleGAN3[23]; samples are presented in supplementary materials), and (b) 100,000 authentic face images: 70,000 from FFHQ and 30,000 from CelebA-HQ [18]. ProGAN and StarGANv2 were trained using CelebA-HQ, whereas the rest of the GAN generators (StyleGAN, StyleGAN2, StyleGAN2-ADA, and StyleGAN3) were trained using FFHQ. For the **iris PAD** task, the test set is comprised of 12,432 samples across six categories (live, artificial, texted contact lenses, display, post mortem, and paper print outs), which is identical to the test set used in the LivDet-Iris-2020 competition benchmark [12].

3.3 Performance Metrics

In an effort to benchmark our results against the most recent human saliency-based training, we first conducted experiments using the exact same dataset sources, model backbones, and experimental environment as in [5]. To assess the uncertainty related to random training seeds, we trained 10 models for each discussed dataset-model configuration. Area Under the ROC Curve (AUC) is used to compare the performance of the models.

3.4 Generation of Human-Guided AI Teachers

Saliency-based Model Training Our framework for teaching AI Teachers begins by first training 10 models on the TAIT dataset using human annotations and the CYBORG loss, which simultaneously maximizes the classification performance, and closeness of the model and human saliency maps [5]:

$$\mathcal{L} = \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\mathbf{1}_{y_k\in\mathcal{C}_c}\Bigg[\underbrace{(1-\alpha)\|\mathbf{s}_k^{(\text{teacher})}-\mathbf{s}_k^{(\text{model})}\|^2}_{\text{teacher saliency loss component}}-\underbrace{\alpha\log p_{\text{model}}\big(y_k\in\mathcal{C}_c\big)}_{\text{classification loss component}}\Bigg] \tag{1}$$

where $\|\cdot\|$ is the ℓ_2 norm, y_k is a class label for the k-th sample, $\mathbf{1}$ equals to 1 when $y_k \in \mathcal{C}_c$ (and equals to 0 otherwise), C is the number of classes, K is a batch size, $\alpha = 0.5$ is a trade-off parameter weighting teacher- and model-based saliency maps. The $\mathbf{s}_k^{(\text{teacher})}$ is the salience generated by the teacher (*i.e.* by a human in case of teaching the AI Teacher models, or by the AI Teacher in case of teaching the AI Students) for the k-th sample. The $\mathbf{s}_k^{(\text{model})}$ is the model saliency estimated by weighting all features maps in the last convolutional layer using weights in the last classification layer belonging to the predicted class $\mathcal{C}_c$. We follow [5] and normalize both $\mathbf{s}_k^{(\text{model})}$ and $\mathbf{s}_k^{(\text{teacher})}$ to the range of $\langle 0, 1\rangle$.

Table 1: Mean Area Under the Curve (AUC) for synthetic face detection task solved by all variants of AI Students. Ten models were trained for each variant and standard deviations are given. Only one architecture (Inception) did not benefit from the salience generated by AI Teachers taught initially by humans. Best results for each model architecture are **bolded**, and better type of saliency is color coded: RISE, and CAM.

How the AI Teachers	Mean AUC on the EAIS data			
were trained on TAIT data	**DenseNet**	**ResNet**	**Xception**	**Inception**
Without human salience	0.591 ±0.036	0.601±0.019	0.694±0.011	**0.645±0.020**
With human salience	**0.696±0.016**	**0.634±0.021**	**0.722±0.011**	0.617±0.040

Next, the model with the highest AUC on the validation part of the TAIT dataset is selected as the AI Teacher. The selected teacher model then generates saliency (using either RISE or CAM approach) on the larger unannotated TAIS training set. Finally, 10 subsequent AI Students are trained on the TAIS dataset with the associated AI Teacher-generated salience maps using CYBORG loss.

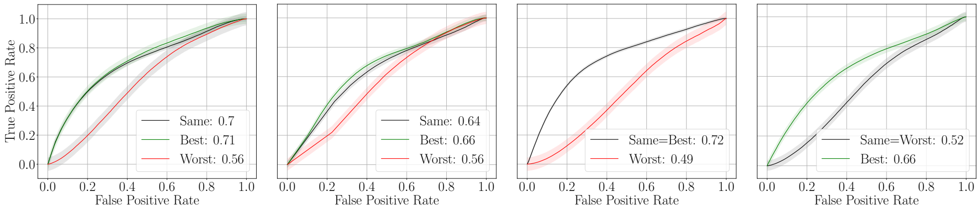
Model Architectures We used four out-of-the-box architectures across all experiments: DenseNet121 [16], ResNet50 [15], Xception [10], and Inception v3 [31]. All model weights were instantiated from the pre-trained ImageNet weights. All models were trained using Stochastic Gradient Descent (SGD) for maximum 50 epochs, with learning rate of 0.005, modified by a factor of 0.1 every 12 epochs. The initial teacher saliency and model saliency components in the human-guided (CYBORG) loss were given equal weighting, *i.e.* $\alpha = 0.5$ in Eq. (1), as in [5] and [3]. Optimal student model configurations were achieved by lowering $\alpha = 0.01$ (the exploration of the weighting parameter α is detailed in Section 4).

4 Results

Answering RQ1: Which type of training makes better AI Teacher models: human guided or purely data driven? To fairly assess the value of using human salience in training AI Teachers, we first taught 10 of such models using TAIT data with human saliency. In order to answer research question RQ1, we additionally trained another 10 AI Teachers with only cross-entropy loss ("CE-trained AI Teacher" in Fig. 2). This is to investigate if human annotations are at all needed at any step of the entire framework. For both tasks, three out of four AI Student model architectures benefited from AI Teachers being trained with human annotations as opposed to being trained without human salience, as seen in Tab. 1 for synthetic face detection task. For the one AI Student model architecture that did not benefit from AI Teachers being trained with human salience (Inception), we believe the standard deviations indicate these differences weren't statistically significant. More specifically, we believe this result is due to the selection of a poor Teacher model. The Teacher-Student training paradigm selects the highest-performing model on the validation set as the Teacher. However, this may not always generate the best salience due to overfitting, or latching onto spurious features despite presence of human salience. This is illustrated in Fig. S4 in the supplementary materials, which shows that the selected Inception-based Teacher model failed to focus on important regions of the input image (see specifically col (e) row (ii) in that figure).

With Inception Teacher's saliency maps unfocused on the wrong features, the AI Student's performance will inevitably suffer.

Thus, **the answer to RQ1 is affirmative: effective student models benefit from being trained with AI Teachers trained with human-salience, compared to AI Students taught by teachers not exposed to human salience**.



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

Figure 3: Mean ROC curves, along with bands representing standard deviations across 10 independent train-test runs, illustrating how the performance of the AI Teacher (and resulting salience generated by that teacher model) impacts the AI Student's performance. "Same" indicates that both AI Teacher and AI Student shared the same architecture. For example, "Same" for DenseNet indicates that the student model was trained using saliency generated by the DenseNet-based teacher model. "Best" means that the student model was trained with saliency generated by the best AI Teacher, possibly with a different architecture. For comparison, the "Worst" means that the student model was trained with saliency generated by the worst AI Teacher. For ROCs denoted as "Same = Best" or "Same = Worst", there was no difference in performance for these options, so we keep one ROC curve.

Answering RQ2: Can the top performing AI Teacher improve the performance of AI Students across different CNN architectures? One of the most profound insights is that AI Teacher's saliency is transferable across different CNN architectures. To explore this, we applied the AI Teacher which taught the best-performing student model (Xception-based teacher with CAM-based saliency, AUC=0.722 in Tab. 1), to train another student models, but based on different architectures (DenseNet, ResNet, and Inception) than the AI teacher. Every model achieved better performance using Xceptions's CAM-based saliency instead of the saliency generated by a teacher sharing the same architecture as a student. As a sanity check, we then performed the opposite experiment: applying the AI Teacher that resulted in the worst-performing student models (in this case Inception-based teacher model with RISE-based saliency, AUC=0.508) to train the student model (again, based on a different architecture than the AI Teacher: DenseNet, ResNet and Xception). Every model that used Inception's saliency decreased its performance significantly. Fig. 3 illustrates the results via ROC curves. From these experiments, we can conclude that the best teacher models are learning more salient features from the data, and passing that information along effectively to student models. Thus, the **answer to RQ2 is affirmative: the top performing AI Teacher can improve the performance of AI Students and does not need to have the same experimental parameters or even architecture to convey saliency-related information efficiently to future student models**.

Table 2: Area Under the Curve (AUC, mean ± std) achieved by the baselines and the optimal student model in synthetic face detection task. Optimal AI Student configurations were achieved by training the model using the optimal AI Teacher's configuration (Xception + CAM) with $\alpha = 0.01$ (*i.e.* encouraging the model to focus on salience instead of class label).

Model	**Baseline 1** (small set, entire human salience available)	**Baseline 2** (larger set, no human salience)	**Optimal AI Student** (this paper: large set, optimal use of human salience)
DenseNet	0.633 ± 0.04	0.629 ± 0.039	**0.767** ± **0.020**
ResNet	0.612 ± 0.05	0.555 ± 0.061	**0.718** ± **0.012**
Xception	0.730 ± 0.02	0.586 ± 0.074	**0.743** ± **0.005**
Inception	0.679 ± 0.03	0.610 ± 0.035	**0.746** ± **0.019**

Table 3: Same as in Table 2, except that results for iris PAD are shown. Optimal AI Students may have different CNN architectures than their teacher models, and may have "aggressive" ($\alpha = 0.01$) or modest ($\alpha = 0.50$) weighting towards using human saliency.

Model	**Baseline 1** (small set with the entire human salience available)	**Baseline 2** (large set, no human salience)	**Optimal AI Student** (this paper: large set, optimal use of human salience)
DenseNet	0.920 ± 0.017	0.917 ± 0.017	**0.950** ± **0.013**
ResNet	0.854 ± 0.031	0.905 ± 0.013	**0.920** ± **0.022**
Xception	0.852 ± 0.018	0.948 ± 0.008	**0.952** ± **0.003**
Inception	0.888 ± 0.018	0.905 ± 0.029	**0.947** ± **0.010**

Answering RQ3: What are the potential performance benefits of the teacher-student training paradigm over the baselines? In answering this research question, we build upon the insights found from the previous two research questions in order to maximize the full capabilities of the proposed training paradigm. We explore increasing the performance of teacher-student training by: (1) using the "best" teacher model's saliency (conclusion from answering RQ1: teachers trained using human-salience teach better student models & conclusion from answering RQ2: Xception's teacher using CAM saliency is best to teach students detecting synthetic faces), and (2) training the student models to "look" more aggressively at the teacher's saliency maps by lowering α in Eq. (1) during training to near zero ($\alpha = 0.01$). In addition to using the optimal teacher's saliency maps, once the AI Students have a more accurate map of "where to look," the classification (cross entropy-based) component of the loss becomes less important to the student models. The results from these experiments are reflected in Tab. 2. As illustrated, **this approach significantly boosted the performance across all CNN architectures, and the accuracy of optimal student models trained with the proposed approach surpassed the accuracy of the baseline models.**.

Answering RQ4: Can this training approach be applied to domains beyond synthetic face detection? In order to validate our findings, we repeated our experiments for iris PAD task. For RQ1, we saw similar findings as for synthetic face detection, as three out of the four student model architectures benefited from AI Teachers taught using human saliency. For RQ2, the top-performing AI Teachers improved the performance of AI Students across different CNN architectures. Finally, we were able to increase the performance of AI Student models over human-guided teacher models by selecting optimal teacher saliency and alpha values (RQ3), as shown in Tab. 3. We included non-essential, yet potentially informative results and graphs related to the iris PAD results in the supplementary materials.

5 Conclusions

We have proposed, designed and evaluated a learning framework that makes an efficient use of limited human saliency data, allowing to significantly scale human-guided training strategies. To accomplish this goal, we first use a small amount of human annotations to train AI Teachers, that is, models that generate saliency for subsequent AI Students. These student models are trained using existing saliency-guided training paradigms, but utilizing synthetically-generated salience rather than human-supplied salience.

We extensively tested our framework in a task of synthetic face detection, and explored selected variants in a task of iris presentation attach detection (to check the domain generalization hypothesis). We observed a boosted performance of the resulting student models trained by AI Teachers built using human salience, when compared to student models trained without any salience information. Even more importantly, we also saw a better performance when models trained with human salience were used as AI Teachers, compared to teacher models not exposed to human salience before. That confirms the usefulness of incorporating human salience into CNN training, and this paper – to our knowledge – for the first time demonstrates how to leverage small availability of human annotations and scale the human perception-augmented training. The proposed way of learning can thus serve as one of the ideas to match the growing size of datasets in any domain in which humans can provide initial limited salience information sufficient to train teacher models.

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7617–7627, 2021.

[2] Sandipan Banerjee, John S. Bernhard, Walter J. Scheirer, Kevin W. Bowyer, and Patrick J. Flynn. SREFI: Synthesis of realistic example face images. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 37–45, 2017. doi: 10.1109/BTAS.2017.8272680.

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

[4] Aidan Boyd, Daniel Moreira, Andrey Kuehlkamp, Kevin Bowyer, and Adam Czajka. Human saliency-driven patch-based matching for interpretable post-mortem iris recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 701–710, 2023.

[5] Aidan Boyd, Patrick Tinsley, Kevin Bowyer, and Adam 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. doi: 10.1109/WACV56688.2023.00605.

[6] Zachariah Carmichael and Walter J. Scheirer. On the objective evaluation of post hoc explainers. *CoRR*, abs/2106.08376, 2021. URL https://arxiv.org/abs/2106.08376.

[7] Zachariah Carmichael and Walter J Scheirer. Unfooling perturbation-based post hoc explainers. In *AAAI Conference on Artificial Intelligence, Washington D.C.*, pages 1–9, 2023.

[8] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo 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.

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[11] Adam Czajka, Daniel Moreira, Kevin Bowyer, and Patrick 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. doi: 10.1109/WACV.2019.00107.

[12] Priyanka Das, Joseph McFiratht, Zhaoyuan Fang, Aidan Boyd, Ganghee Jang, Amir Mohammadi, Sandip Purnapatra, David Yambay, Sébastien Marcel, Mateusz 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.

[13] Rachel Lea Draelos and Lawrence Carin. Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification. *arXiv preprint arXiv:2011.08891*, 2020.

[14] Justin Dulay and Walter J Scheirer. Using human perception to regularize transfer learning. *arXiv preprint arXiv:2211.07885*, 2022.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[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*, pages 4700–4708, 2017.

[17] Jin Huang, Derek Prijatelj, Justin Dulay, and Walter Scheirer. Measuring human perception to improve open set recognition. *arXiv preprint arXiv:2209.03519*, 2022.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[21] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo 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.

[23] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[24] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *arXiv preprint arXiv:1805.08819*, 2018.

[25] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[26] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[27] P. Jonathon Phillips, Patrick J. Flynn, and Kevin W. Bowyer. Lessons from collecting a million biometric samples. *Image and Vision Computing*, 58:96–107, 2017. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2016.08.004. URL https://www.sciencedirect.com/science/article/pii/S0262885616301287.

[28] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[29] Rabia Saleem, Bo Yuan, Fatih Kurugollu, Ashiq Anjum, and Lu Liu. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513:165–180, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.09.129. URL https://www.sciencedirect.com/science/article/pii/S0925231222012218.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[32] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[33] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.