CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning-Based Synthetic Face Detection

Aidan Boyd, Patrick Tinsley, Kevin Bowyer, Adam Czajka
University of Notre Dame, Notre Dame IN 46556, USA
{aboyd3,ptinsley,kwb,aczajka}@nd.edu

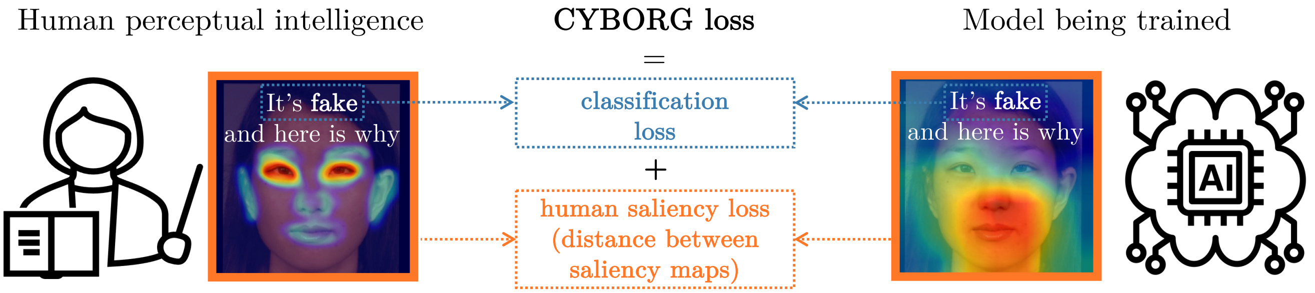


Figure 1: Our proposed training strategy **C**onve**Y**s **B**rain **O**versight to **R**aise **G**eneralization. CYBORG continually encourages the training process to look at image regions judged as salient for human visual perception. This results in a model that is more likely to learn features from regions that are salient to humans, and less likely to learn features that are accidentally correlated with class labels. A boost in generalization performance is demonstrated.

Abstract

Can deep learning models achieve greater generalization if their training is guided by reference to human perceptual abilities? And how can we implement this in a practical manner? This paper proposes a training strategy to ConveY Brain Oversight to Raise Generalization (CYBORG). This new approach incorporates human-annotated saliency maps into a loss function that guides the model's learning to focus on image regions that humans deem salient for the task. The Class Activation Mapping (CAM) mechanism is used to probe the model's current saliency in each training batch, juxtapose this model saliency with human saliency, and penalize large differences. Results on the task of synthetic face detection, selected to illustrate the effectiveness of the approach, show that CYBORG leads to significant improvement in accuracy on unseen samples consisting of face images generated from six Generative Adversarial Networks across multiple classification network architectures. We also show that scaling to even seven times the training data, or using non-human-saliency auxiliary information, such as segmentation masks, and standard loss cannot beat the performance of CYBORG-trained models. As a side effect of this work, we observe that the addition of explicit region annotation to the task of synthetic face detection increased human classification accuracy. This work opens a new area of research on how to incorporate human visual saliency into loss functions in practice. All data, code and trained models used in this work are offered with this paper.

1. Introduction

How do you teach a child to ride a bicycle? The passive option is to set the child on the bike, give the bike a push and then stand back silently, watching what happens. The active option is to set the child on the bike, give them a push, and then run alongside, continually giving advice on what to do. We argue that current state of training of deep learning-based models is more passive than active. We introduce a new training process that – by incorporating human visual perception into a loss function – continually reminds the model being trained of the image regions judged as salient to humans, as illustrated in Fig. 1.

The main goal of the proposed CYBORG approach is to **c**onve**y** **b**rain **o**versight to **r**aise **g**eneralization by encouraging the deep learning model to focus on human-salient regions. This is achieved by adding a new component to the loss, based on the difference between human saliency heatmaps and the model's class activation mapping-based [55] heatmaps in each training batch. Thus our new loss function blends classical data-driven optimization with human-derived oversight or "coaching" about the parts of the image that are salient to the problem.

To demonstrate the advantages of CYBORG training, we apply it to the challenging task of distinguishing face images that are authentic versus generated by various modern Generative Adversarial Nets (GANs). To generate human-derived saliency maps, we presented 1,000 pairs of face images to 363 humans. Each image pair contained one authentic (real) and one synthetic image (generated by an example deep learning-based approach, StyleGAN2 [31], and non deep learning-based method SREFI [2]). Viewers were asked to (a) choose which face was authentic and which was synthetic, and (b) annotate regions that support their decision. For each image, annotations from the viewers were compiled into a *saliency map* summarizing human judgment about the salient image regions.

Our experiments show that CYBORG learning increases the accuracy of detecting synthetic data in an open-set classification regime, in which test samples are generated with six different GAN architectures withheld during the training process. We also demonstrate that although adding human saliency to an example model implementing an attention mechanism [13] increases performance, this improvement is small compared to when the CYBORG approach is employed. The **main contributions** of this work are:

- **Introduction of the CYBORG training strategy**, which benefits from human judgment about salient regions by incorporating perceptual intelligence into loss function.
- **Open-set evaluation of CYBORG training** that shows a significant improvement for multiple state-of-the-art deep learning models (ResNet, DenseNet, Inception and Xception), as well as for the existing synthetic face detector.
- **Experiments assessing the "value" of human annotations in two ways:** (a) demonstrating that at least 7 times more training data is needed to train a model in classical fashion to achieve performance competitive with CYBORG, and (b) replacing human saliency maps with non-human-sourced cues offered by face segmentation masks, which did not achieve the level of generalization achieved by the CYBORG-trained models.
- **Evaluation of state-of-the-art "deep fake" detector on GAN-generated face images**, which illustrates that the solutions to "deep fake" detection and synthetic face detection are not cross-applicable.
- Results demonstrating that **human classification accuracy increases when participants are asked to annotate image regions that support their decisions**, compared to the same experiment without annotations.
- **Data and source codes to reproduce all experiments:** a test set containing 600,000 synthetic faces generated by six GAN architectures (ProGAN, StarGANv2, StyleGAN, StyleGAN2, StyleGAN2-ADA and StyleGAN3), human annotation data, and all neural network models at `https://github.com/BoydAidan/CYBORG-Loss`.

2. Related Work

Synthetic Image Generation and Detection. Since Goodfellow *et al.* [17], many open-source, and (often) pre-trained GANs for image synthesis have become available [26, 30, 31, 28, 29, 9, 6, 56, 39]. The authors of [42, 16] maintain that frequency domain analysis can unveil artifacts or manipulations in GAN-generated images across different model architectures, datasets, and resolutions. However, as documented by Marra *et al.* [35], conventional, non-deep-learning methods (such as steganalysis [11]) fail in the presence of compression. With a virtually infinite number of fake samples in their training processes, deep networks such as ResNet [20], DenseNet [23], InceptionNet [48], and Xception-Net [10] have achieved over 99% accuracy in fake image recall [49]. Even before public release of StyleGAN3 [29], there were several proactive efforts towards detecting StyleGAN3 images [47, 25, 53, 18, 34, 51]. These models can be complemented with the proposed CYBORG loss, and such an attempt, with a model proposed by Wang *et al.* [51], is described in the supp. materials.

Although the generation of never-before-seen images lends itself naturally to the creative process, the ability to manipulate existing images poses a significant security problem [4, 8]. A commonly commercialized scapegoat is deepfakes [14], which splices real identities onto realistic-looking videos. We demonstrate in this paper that state of the art deep fake detectors may not be effective in detecting fully-synthetic samples, which this paper focuses on.

Using Human Perception to Understand / Improve Computer Vision. O'Toole *et al.* [38] demonstrated that machines were never less accurate than humans on face images of various quality. RichardWebster *et al.* [43] showed that observing human face recognition behavior in certain contexts can be used to explain why face matchers succeed or fail, leading to better model explainability. In biometrics, human saliency was found complementary to algorithm saliency and thus beneficial to combine them [50, 36]. Czajka *et al.* measured human visual saliency via eye tracking and used it to build human-driven filtering kernels for iris recognition [12], achieving better performance than non-human-driven approaches. Human-guided training data augmentation was proposed by Boyd *et al.* [5] to build deep learning-based iris presentation attack detection methods generalizing exceptionally well to unknown attack types.

In broader machine learning, incorporation of results from psychophysics has aided in deep learning tasks such as image captioning for scene understanding [21, 24], handwriting analysis [19], and natural language processing [54].

Linsley *et al*. [32] proposed to incorporate human-sourced saliency into a self-attention mechanism, combining global and local attention in the "GALA" module. We demonstrate in the supp. materials how our human saliency maps can be incorporated into the attention mechanism, and show that CYBORG allows for a better gain in accuracy than using human saliency in the attention mechanism. Bruckert *et al*. [7] considered eye tracking-based human saliency to improve the model's saliency.

Differences between the proposed CYBORG method and previous works: (a) human spatial saliency and model spatial saliency have never before been directly compared and blended into overall loss; (b) CYBORG does not require architectural changes to the model *e.g.*, a specialized attention module.

3. Experimental Datasets

Two types of face image datasets are used: authentic datasets consisting of real images from three sources (CelebA-HQ [26], Flickr-Faces-HQ [30] and FRGC-Subset [40]), and synthetic datasets consisting of fake images from seven different generators (ProGAN, StyleGAN, StyleGAN2, StyleGAN2-ADA, StyleGAN3, StarGANv2 and SREFI [26, 30, 31, 28, 29, 9, 2]). Along with Fig. 2, the following sections briefly characterize data sources.

CelebA-HQ [26] is a high-quality version of the original CelebA dataset [33], containing 30,000 images of celebrities at a resolution of 1024×1024.

Flickr-Faces-HQ (FFHQ) includes 70,000 1024×1024 images of faces varying in age, ethnicity, and facial accessories (glasses, hats, etc.) [30].

FRGC-Subset dataset contains 16,433 face images, randomly sampled from a set of publicly available datasets collected by Phillips *et al*. [40]. Images show frontal faces varying in expression, ethnicity, gender, and age.

SREFI was generated by the "synthesis of realistic face images" (SREFI) [2] method, which works by first matching similar face images based on VGG-Face features, splitting them into region-specific triangles, and implanting from donor faces onto a base face to create a blended identity. To ensure consistency, important facial features, such as the mouth and eyes, on the generated image are required to come from the same donor.

ProGAN contains 100,000 images downloaded from [27]. Unlike its successors (StyleGAN), Karras *et al*.'s ProGAN generator network was trained on CelebA-HQ images [26].

The StyleGAN Family. The next four synthetic datasets were generated with StyleGAN architectures [30, 31, 28, 29]. The original StyleGAN was trained in a similar manner to its predecessor ProGAN [26], but with the added feature of mixable disentangled layers for style transfer. The next version, StyleGAN2 [31], removed artifacts found in original StyleGAN images and improved image reconstruction via path length regularization. The third iteration of StyleGAN, StyleGAN2 with adaptive discriminator augmentation) [28], solves for training GANs in data-limited scenarios. Finally, StyleGAN3 [29] mitigates aliasing in rotation- and translation-invariant generator networks.

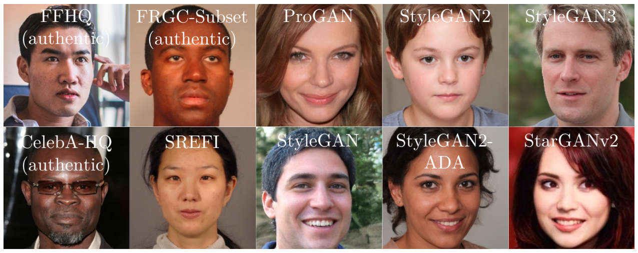


Figure 2: Examples from each data source.

For original StyleGAN and StyleGAN2, sets of 100,000 fake face images were downloaded from their GitHub repositories. For StyleGAN2-ADA and StyleGAN3, sets of 100,000 images were generated using default generator settings, including a truncation of (ψ) of 0.5 (as recommended by StyleGAN authors).

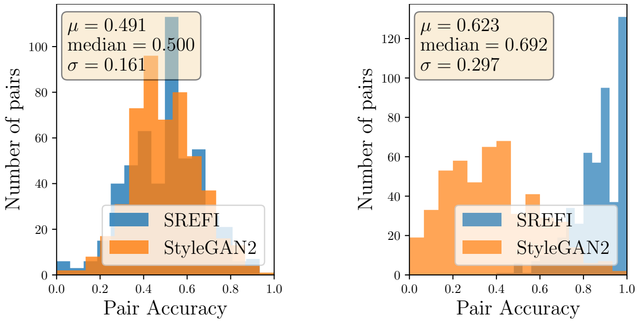
StarGANv2 produces images with the main focus of style transfer [9], unlike StyleGAN. Generated images show source identities "dressed" in the style of the supplied reference images. In order to ensure high facial quality of StarGANv2 generated images, 250,000 images were initially synthesized using a supplied network (pre-trained on CelebA-HQ). These synthetic samples were then scored and sorted according to facial quality using FaceQNet [22], a CNN designed to assess input images' suitability for face recognition tasks. The final dataset consisted of the top-ranked 100,000 images.

4. Human Saliency

4.1. Acquisition of Human Salient Regions

We replicate an experiment of Shen *et al*. [46], in which subjects judge pairs of non-masked face images as fake or real, but we require subjects to annotate regions supporting their decisions. Specifically, participants are presented with a pair of face images (one a synthetically-generated identity, and the other an authentic facial image), and asked to decide which image is either the synthetic image or the real image in a two-alternative forced choice (2AFC) manner. The prompt question alternated between asking *which is real* versus *which is fake*[1]. Next, users were asked to highlight regions (not size- nor location-constrained) of the image supporting their classification decision.

[1] The online annotation tool developed for this work is presented in supp. materials.



(a) Only classification

(b) With annotations

Figure 3: Human classification of face images as real/fake is more accurate when subjects annotate images: (a) Shen *et al*. [46] found that humans did not accurately classify face images as real/fake; (b) the same experiment in which we asked subjects to annotate the image regions that supported their classification. Average human accuracy is substantially higher in study (b). These histograms detail the accuracy of humans on the 1,000 pairs.

Data (decisions and annotations) was collected from 363 subjects (recruited via Amazon Mechanical Turk), with an average of 29.6 image pairs processed by each subject. The synthetic images consisted of 500 images generated using SREFI from the FRGC-Subset dataset, and 500 images synthesized by StyleGAN2 and downloaded from `thispersondoesnotexist.com`. 10,750 annotations were obtained, matching exactly the number of question/pair samples in Shen's work for fair comparison. For evaluating the CYBORG approach, only annotations for **correctly** classified pairs were used in the training process.

4.2. Do Annotations Improve Human Accuracy?

As the only difference between our protocol and that used in [46] was the annotation requirement, we can properly diagnose how annotating images impacts decision accuracy. Fig. 3(a) outlines the original results[2], where the blue and orange histograms represent results attained for 1,000 image pairs with synthetic data generated by SREFI and StyleGAN2 approaches, respectively. As can be seen, human accuracy is at random chance level when not asked to annotate regions to support their decision. However, Fig. 3(b) shows that the accuracy increased from 50% (random chance) to 69.2% simply by requesting users to annotate the images (and so spend more time on each pair). This experiment suggests that human accuracy in detecting synthetic faces can be improved simply by forcing annotations that support classification decisions. A side observation is that StyleGAN2 images may appear more realistic than SREFI images in such new setup (see the shift between distributions in Fig. 3(b)).

4.3. Building Human Saliency Maps

All correct annotations, as shown in eight individual images in Fig. 4(b), are combined together with equal weight to create image representations called **human saliency maps** shown in Fig. 4(c). A Gaussian blur of $\sigma = 5$ is applied to the combined array to smooth edges between regions of varying annotation density, and the map is scaled to the range of $\langle 0, 1\rangle$. White pixels in the saliency map correspond to regions that more subjects had annotated as important. Black pixels correspond to areas not annotated by any subject.

After data collection, there were 1,821 correctly classified images with annotations, consisting of 919 authentic images and 902 synthetic images. These 1,821 images represent the training set for the CYBORG loss experiments.

5. CYBORG Loss

In the same way a cyborg is a human-machine hybrid, the proposed CYBORG training strategy combines the human saliency information attained through annotations (*human saliency loss component*) with a requirement for high classification accuracy (*classification loss component*). The former component steers activations in the feature maps in the last convolutional layer to be aligned with human-defined regions of importance, while the model may still benefit from a data-driven learning approach owing to the latter component.

More specifically, the human saliency loss directly compares the difference in salient regions between machine and human during training. To accomplish this, we implemented a fully-differentiable Class Activation Mapping (CAM) approach [55] that, given the current weights, can generate CAMs for all samples in each training batch. Resultant CAMs are scaled to the range of $\langle 0, 1\rangle$, human saliency maps are downsized to the same size as CAMs, and then both heatmaps are compared via ℓ_2 norm. Formally, we define CYBORG loss $\mathcal{L}$ as:

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

where $\|\cdot\|$ is the ℓ_2 norm, y_k is a class label for the k-th sample, $\mathbf{1}$ is a class indicator function equal to 1 when $y_k \in \mathcal{C}_c$ (0 otherwise), C is the total number of classes, K is the number of samples in a batch, $\alpha = 0.5$ is a trade-off parameter weighting human- and model-based saliencies, $\mathbf{s}_k^{(\text{human})}$ is the human saliency for the k-th sample, and

$$\mathbf{s}_k^{(\text{model})} = \mathbf{f}_1 w_1^{(c)} + \mathbf{f}_2 w_2^{(c)} + \cdots + \mathbf{f}_N w_N^{(c)}$$

[2] We thank the authors of [46] for sharing the raw results with us.

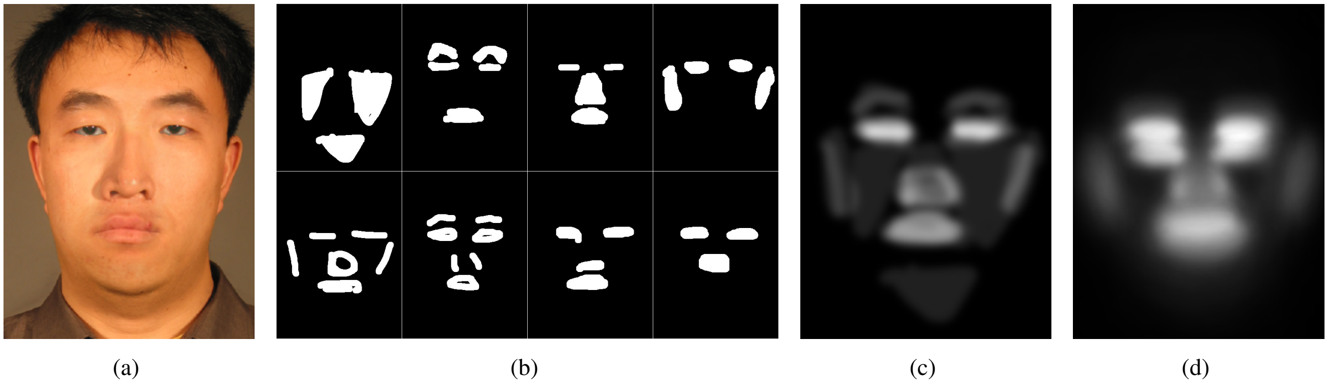


Figure 4: Creation of human saliency map: (a) an image (in this example generated by SREFI) presented to human annotators; (b) eight annotations from viewers who correctly classified the image; (c) averaged annotations defining the human salient features for that sample, as used by CYBORG; (d) average of all human annotations for all images in the training set.

is a class activation map-based model's saliency for the k-th sample, where N is the number of feature maps $\mathbf{f}$ in the last convolutional layer, and $w^{(c)}$ are the weights in the last classification layer belonging to predicted class $\mathcal{C}_c$. Both $\mathbf{s}_k^{(\text{model})}$ and $\mathbf{s}_k^{(\text{human})}$ are normalized to the range $\langle 0, 1 \rangle$.

The reason that the CAM method was implemented rather than a more modern approach (GradCAM [45] or EigenCAM [37]) is that the latter approaches require backpropagation to calculate gradients with respect to the input to determine salient regions (in addition to gradients with respect to the weights). This is expensive to do during training while maintaining differentiability, and these methods are typically only used on fully trained models where backward calls can be completed in a post-hoc fashion. For CAM, only a forward pass is necessary, meaning it can be bootstrapped into the training strategy directly.

6. Experimental Setup for CYBORG

Face images are aligned using *img2pose* [1], cropped, and resized to 224×224. Face bounding boxes are expanded 20% in all directions before cropping, with an additional 30% on the forehead to ensure the head is fully in view. Human saliency maps are resized and cropped the same, to keep spatial correspondence.

6.1. Training Scenarios

Scenario 1: Classical Training. The basic scenario consists of training the studied architectures in a task of synthetic face image detection, on image data for which human saliency information was collected, but using only the classification component in the loss function (*i.e.* no human annotations are used). The training set in this scenario consists of 919 authentic and 902 synthetic images. The validation set consists of 20,000 images: 10,000 authentic images, 5,000 images generated using SREFI, and 5,000 images downloaded from *thispersondoesnotexist.com*. The training and validation set used in this scenario will be further referred to as the *original data*.

Scenario 2: Classical Training with Large Data. To evaluate how much additional data is required to achieve CYBORG-level performance from learning with only classification loss (as in Scenario 1), we curate a larger dataset than that used in Scenario 1. Starting from the original data, we add six times more samples resulting in a training set 7× the initial size. Scarcity of authentic images in the source datasets prevented going beyond 7×, as adding data from different source could add bias to the comparison.

Scenario 3: CYBORG Training. Using the *original data* as in Scenario 1, we apply the same training strategy but include the human saliency component in the loss function to create CYBORG loss. The difference between Scenarios 1 and 3 is the loss function, so observations can be directly correlated with the effectiveness of CYBORG training.

Experimental Parameters. To ensure that observations are not architecture-specific, the base experiments are completed on four out-of-the-box architectures: DenseNet-121 [23], ResNet50 [20], Inception v3 [48] and XceptionNet [10]. For all methods, Stochastic Gradient Descent (SGD) is used, with learning rate of 0.005, modified by a factor of 0.1 every 12 epochs. Training ran for 50 epochs, and weights giving the highest validation accuracy were selected as the final model. The validation set is constant, as described in Scenario 1. Networks are instantiated from the pre-trained ImageNet weights [41]. For all experiments using CYBORG loss, the human saliency and the classification components are given equal weight ($\alpha = 0.5$). Each architecture/scenario pair is independently trained 10 times, to generate error statistics on the test set.

6.2. Testing Protocol

To evaluate accuracy of the models trained under the three scenarios, we composed a comprehensive test set of

100,000 synthetically generated images from each of six different GAN architectures, ending up with 600,000 total test samples. The authentic face datasets used for testing are the FFHQ dataset (70,000 images) and the CelebA-HQ dataset (30,000 images). For ProGAN and StarGANv2, the training data is CelebA-HQ; for the remaining four StyleGAN sets, the training data is FFHQ. This setup aims at demonstrating whether models can differentiate between authentic samples and synthetic samples, where the latter are generated by a GAN trained on the former.

6.3. Evaluating State-Of-The-Art DeepFake Detector on Test Data

In order to properly compare our CYBORG models against existing deepfake detectors, we evaluated the state-of-the-art ensemble method from Bonettini *et al*. [3] on *our* test set of synthetic images. Of the ten available models, five were trained on the DeepFake Detection Challenge (DFDC) dataset [15], and five were trained on the FaceForensics++ (FF++) dataset [44].

For each dataset, ensemble methods were composed, using models trained on DFDC or models trained on FF++. Before evaluating on *our* test data of synthetic images, we verified model performance by evaluating the reported top two ensemble methods (one for DFDC, one for FF++) on Bonettini *et al*. test deepfake data. We then ran the same two top-performing ensemble methods on our test data to compare results with CYBORG-trained models.

6.4. Assessing The Value Of Human Annotations

To determine the usefulness of human annotations in the CYBORG loss function, a comparison to a non-human-saliency-guided baseline is needed. A face parsing tool, BiSeNet [57], is applied to the training images to attain a mask detailing all facial regions[3] and CYBORG training is applied with BiSeNet segmentation masks instead of human saliency maps. The goal of this experiment is to determine whether human saliency maps provide better cues than automatically-determined face masks. An affirmative answer could limit the costs of human saliency acquisition.

7. Evaluation Results

Figure 5 summarizes the performance observed for each of the four studied architectures by presenting ROC curves obtained for the comprehensive set of all $100,000$ authentic and $600,000$ synthetically-generated test samples[4]. For all experiments, training and validation is repeated 10 times in order to assess statistical significance of the observed differences in the results. Area Under the Curve (AUC) is given along with ± one standard deviation across the 10 runs.

Scenario 1 vs Scenario 3 (*i.e*. Classical vs CYBORG Training). As shown in Fig. 5, the models trained just using the *original data* do not generalize well to the test sets. In contrast, when CYBORG training is applied on the same data, accuracy on the test sets increases significantly. Fig. 6 outlines the training and validation accuracies for both Scenario 1 (only classification loss) and Scenario 3 (with CYBORG loss) for training of the ResNet50 models. As it can be seen, training accuracy quickly reaches 100%, meaning both sets learn representative features of the training samples. However, the **CYBORG-trained model shows better validation accuracy across all epochs.** The decrease in validation accuracy for Scenario 1 models suggests overfitting, and the subsequent plateau (and even slight decline) can be explained by the training accuracy reaching 100% resulting in minimal optimization. The supplementary materials include plots for DenseNet, Inception-v3 and Xception models, showing very similar trends.

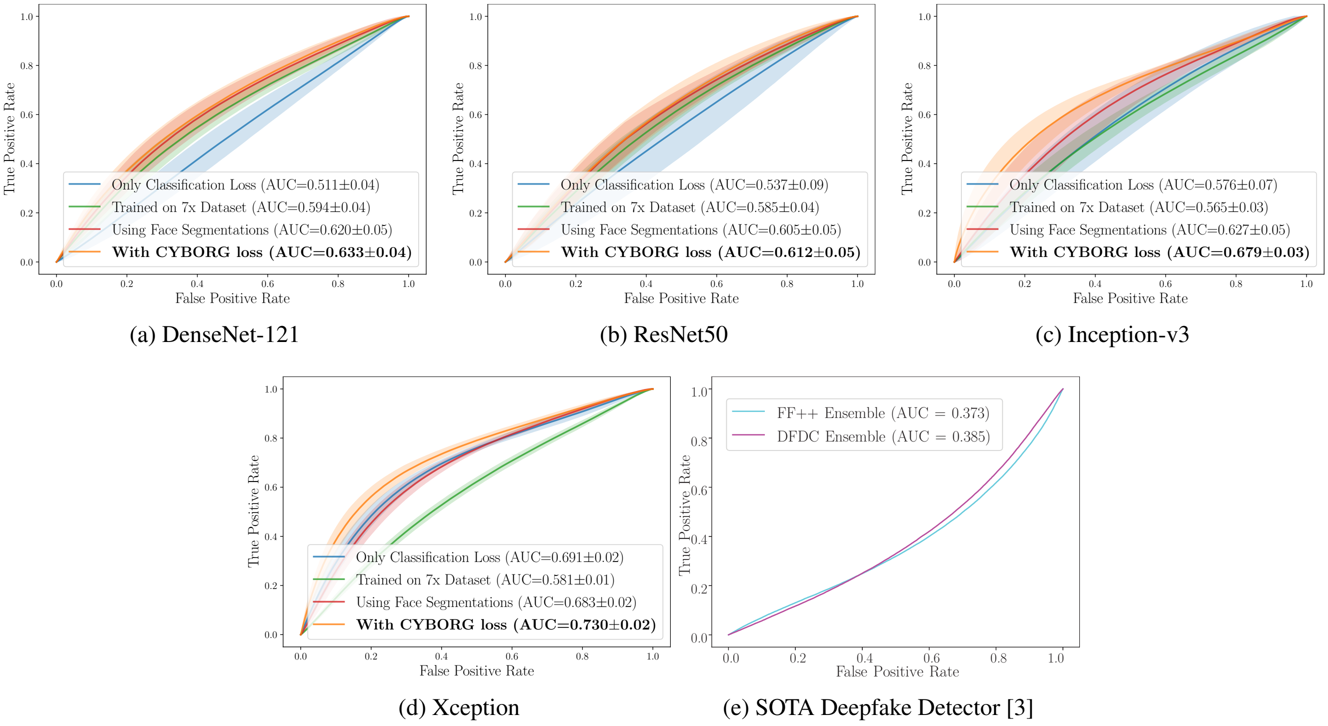
Scenario 2 (*i.e*. Classical Training with Large Data). Experiments were conducted to determine whether simply adding more data from the same sources as the original data to the Scenario 1 approach would bridge the performance gap. Giving the classical training process additional data, up to 7× the original amount, does not enable it to achieve CYBORG-level accuracy. In some cases, the performance of models trained on larger sets is even inferior to models trained with less data and CYBORG. The classical training simply overfits to the training data and so cannot generalize to samples generated by unknown GAN architecture. The CYBORG-trained models generalize better.

Evaluating An Off-The-Shelf Deepfake Detector on Test Data. The ensemble-based "deep fake" detection methods [3] demonstrated very high performance on the DFDC and FF++ test data with AUCs of 0.957 and 0.920, respectively. That means we were able to replicate the original results without any issues. However, when applied to the task of synthetic image detection, these top-performing "deep fake" ensemble methods are incapable of distinguishing between authentic and synthetically-generated images, as demonstrated by AUCs of less than 0.5 (0.385 and 0.373) for these methods in Fig. 5(e).

What The CYBORG-trained Models "Look" At? The results presented so far suggest that the CYBORG approach does guide deep learning towards models generalizing better on samples generated by never-seen-before GANs. However, are these CYBORG-trained models visually exhibiting behaviour akin to human annotators? To answer this question, visualizations of model saliency are generated on the test set, and illustrated in Fig. 7. For experimental Scenarios 1-3, a plot is created for each of the 10 independently trained models. To create each of these individual plots, the CAM is generated using the same mecha-

[3] example masks can be found in the supp. materials

[4] ROC curves for individual GANs can be found in the supp. material.



(a) DenseNet-121

(b) ResNet50

(c) Inception-v3

(d) Xception

(e) SOTA Deepfake Detector [3]

Figure 5: ROC curves presenting results on the test set consisting of six GAN types for four architectures and one off-the-shelf deepfake detector. Shaded regions in (a)-(d) correspond to ±1 standard deviation of the False Positive Rate (FPR) for a given True Positive Rate (TPR). Results outline that in all cases that CYBORG loss was employed (a-d), an increase in performance compared to classification loss alone can be observed. Additionally, in all (a-d) results CYBORG outperforms the models trained even on seven times the training set with just classification loss, and models trained with face segmentation masks instead of human saliency maps.

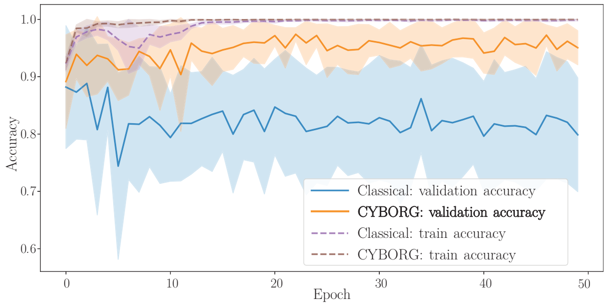


Figure 6: Comparison of training and validation accuracy for ResNet50 with only classification accuracy loss versus with CYBORG loss. Training accuracy quickly approaches 100% for both. But CYBORG-trained models achieve significantly higher validation accuracy throughout, indicating more effective learning. Shaded area represents ±1 standard deviation of the accuracy by epoch.

nism as the model saliency probing during training [55], but for each sample in the test set. The average of all 700,000 CAMs (100k authentic, 600k synthetic) is calculated. This details where the model deems important for classification on average over the entire test set for both classes combined. Because all images are aligned, facial features present in similar locations across test samples.

For both DenseNet and ResNet, the difference between Scenario 1 (“Classical”) and Scenario 3 (“CYBORG”) is immediately evident. Models trained with CYBORG exhibit CAMs that resemble facial features such as the mouth, nose and eyes. The models trained with classification loss alone show less compact CAMs, meaning there is no consensus of importance across the test images.

While the dominating features are comparable for Scenario 1 and Scenario 3 in the Inception-v3 experiment, the CYBORG models are more precisely focused on the facial region. For Xception, both the Scenario 1 and Scenario 3 models present similar CAMs, which is also indicated by the performance. However, CYBORG models exhibit more certainty as indicated by higher compactness of the corresponding CAMs. Tuning of the α value may be required for this model to attain average CAMs similar to ResNet.

For the Scenario 2 (“Classical – Large Data”), roughly similar CAMs are observed across all four architectures. For DenseNet and ResNet, this results in greater performance than classification alone. In these two cases, the Scenario 2 models are more concise than the Scenario 1 mod-

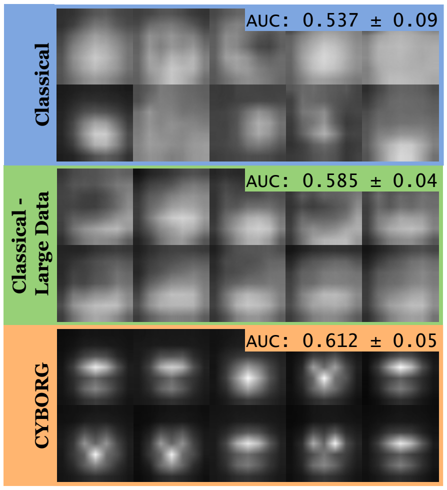


Figure 7: Average CAMs across the entire test set for 10 independently trained ResNet50 models in three experimental settings. Each individual plot is the average CAM obtained for all test images for a given model. (Similar results can be observed for other CNN architectures, included into the supp. materials). Compared to the average human annotation, shown in Fig. 4(d), it is clear CYBORG models are guided effectively by human annotation.

els. For Inception-v3 and Xception, Scenario 2 results in poorer performance than Scenario 1. These models loosely focused on similar features as in Scenarios 1 and 3. However, there is less consensus and more uncertainty indicating increased overfitting to the training data. This added uncertainty explains the degradation in performance.

By direct comparison to the average correct human annotation in Fig. 4(d), the models trained with CYBORG focus on features that are more similar to human detailed salient regions than models trained with classical cross-entropy loss in all cases. Thus, it can be concluded these models are effectively guided by the human annotations supplied during the training process.

Assessing Value Of Human Saliency. As demonstrated in Fig. 5, using human saliency maps results in a larger increase in performance than automatically determined segmentation masks. Thus, to answer the two questions posed in Sec. 6.4, deep learning-based segmentation masks can effectively guide the models using CYBORG loss, however, human saliency maps guide the network to **more generalized features**, thus achieving a **greater performance in open-set classification problem**.

Incorporation Of CYBORG Into An Existing Synthetic Face Detector. To determine whether the incorporation of CYBORG loss would improve upon existing methods, the CYBORG loss was added to Wang *et al.*'s [51] publicly available, re-trainable synthetic face detection model [52]. This resulted in a performance increase from AUC=0.554 ± 0.03 in the classical scenario to AUC=0.591 ± 0.03.

Incorporation Of Human Saliency Into An Attention Mechanism. A popular approach to force networks to focus on specified regions is self-attention. As an additional experiment, we investigated whether replacement of the attention masks proposed in [13] with human saliency results in higher accuracy. We train two models: (1) using the original approach with no human saliency, and (2) using our human saliency maps as the attention masks for authentic and synthetic images. In both cases, the parameters proposed by the authors are used. We found that replacement of the ground truth masks with human saliency increased performance from AUC=0.428 ± 0.04 to AUC=0.498 ± 0.06, suggesting that implanting human perception into the self-attention module narrows the model's search for areas of importance (even in the absence of ground truth) and boosts performance.

8. Conclusions

We proposed the CYBORG approach to CNN training, in which the learning is guided by information distilled from human visual abilities. CYBORG uses a human perception-based loss term for the disagreement between the CNN's class activation map and a human-derived saliency map. To emphasize that CYBORG is independent of CNN backbone, results are shown for four different models: ResNet, DenseNet, Inception and Xception. Applying CYBORG improved performance in detecting synthetic face images from six different GANs unseen in training (Fig. 5). CYBORG-trained models produced CAMs closer to human-annotated regions of saliency than classically trained models (Fig. 7). Comparing the training and validation accuracy of classical versus CYBORG training (Fig. 6) makes it clear that CYBORG results in a model that generalizes better to samples generated by never-seen-before GANs. Evaluation of a state-of-the-art "deep fake" detection model on our test set shows that this task and synthetic image detection are different domains.

Application of the CYBORG approach is possible for tasks in which human accuracy is not at the "expert level". The human saliency maps used in this work came from a task in which median human accuracy was 69.2% (Fig. 3), and only saliency maps associated with correct human decisions were applied.

Finally, we have shown that models trained classicaly with 7 times more training data does not achieve the performance of CYBORG-trained models, and that replacing human perception-driven maps with automatic face segmentation masks ends up with performance inferior to CYBORG. This shows value of the proposed mechanism to apply human perception to use limited data effectively in training.

References

[1] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7613–7623, 2021. 5

[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 Int. Joint Conf.on Biometrics (IJCB)*, pages 37–45, 2017. 2, 3

[3] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019. IEEE, 2021. 6, 7

[4] Johnny Botha and Heloise Pieterse. Fake news and deepfakes: A dangerous threat for 21st century information security. In *Int. Conf. on Cyber Warfare and Security (ICCWS)*, page 57. Academic Conferences and Publishing Limited, 2020. 2

[5] Aidan Boyd, Kevin Bowyer, and Adam Czajka. Human-Aided Saliency Maps Improve Generalization of Deep Learning. *arXiv preprint arXiv:2105.03492*, 2021. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[7] Alexandre Bruckert, Hamed R Tavakoli, Zhi Liu, Marc Christie, and Olivier Le Meur. Deep saliency models: The quest for the loss function. *Neurocomputing*, 453:693–704, 2021. 3

[8] Robert Chesney and Danielle Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, 98:147, 2019. 2

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8185–8194, 2020. 2, 3

[10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1251–1258, 2017. 2, 5

[11] Davide Cozzolino, Diego Gragnaniello, and Luisa Verdoliva. Image forgery detection through residual-based local descriptors and block-matching. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 5297–5301. IEEE, 2014. 2

[12] Adam Czajka, Daniel Moreira, Kevin Bowyer, and Patrick Flynn. Domain-specific human-inspired binarized statistical image features for iris recognition. In *IEEE/CVF Winter Conf. on App. of Comp. Vis. (WACV)*, pages 959–967. IEEE, 2019. 2

[13] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5781–5790, 2020. 2, 8

[14] DeepFakes. Faceswap. `https://github.com/deepfakes/faceswap`, 2021. 2

[15] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 6

[16] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Int. Conf. on Machine Learning (ICML)*, pages 3247–3258. PMLR, 2020. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[18] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. GANimageDetection. `https://github.com/grip-unina/GANimageDetection`, 2021. 2

[19] Samuel Grieggs, Bingyu Shen, Greta Rauch, Pei Li, Jiaqi Ma, David Chiang, Brian Price, and Walter Scheirer. Measuring human perception to improve handwritten document transcription. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 770–778, 2016. 2, 5

[21] Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 8529–8538, 2019. 2

[22] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, and Laurent Beslay. Biometric Quality: Review and Application to Face Recognition with FaceQnet. *arXiv preprint arXiv:2006.03298*, 2020. 3

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4700–4708, 2017. 2, 5

[24] Yupan Huang, Zhaoyang Zeng, and Yutong Lu. Be Specific, Be Clear: Bridging Machine and Human Captions by Scene-Guided Transformer. In *Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, pages 4–13, 2021. 2

[25] Yan Ju. GAN-generated-image-detector. `https://gitlab.com/littlejuyan/GAN-generated-image-detector`, 2021. 2

[26] 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. 2, 3

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation: Official TensorFlow Implementation. `https://github.com/tkarras/progressive_growing_of_gans`, 2021. 3

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training Generative Adversarial Networks with Limited Data. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2020. 2, 3

[29] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2021. 2, 3

[30] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4401–4410, 2019. 2, 3

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8110–8119, 2020. 2, 3

[32] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *arXiv preprint arXiv:1805.08819*, 2018. 3

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, December 2015. 3

[34] Sara Mandelli, Nicoló Bonettini, Paolo Bestagini, and Stefano Tubaro. Training cnns in presence of jpeg compression: Multimedia forensics vs computer vision. In *IEEE Int. Workshop on Inf. Forensics and Sec. (WIFS)*, pages 1–6, 2020. 2

[35] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-generated fake images over social networks. In *IEEE Conf. on Multimedia Inf. Proc. and Retrieval (MIPR)*, pages 384–389. IEEE, 2018. 2

[36] Daniel Moreira, Mateusz Trokielewicz, Adam Czajka, Kevin Bowyer, and Patrick Flynn. Performance of Humans in Iris Recognition: The Impact of Iris Condition and Annotation-driven Verification. In *IEEE/CVF Winter Conf. on App. of Comp. Vis. (WACV)*, pages 941–949. IEEE, 2019. 2

[37] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class Activation Map using Principal Components. In *IEEE Int. Joint Conf. on Neural Networks (IJCNN)*. IEEE, Jul 2020. 5

[38] Alice J O'Toole, Xaiobo An, Joseph Dunlop, Vaidehi Natu, and P Jonathon Phillips. Comparing face recognition algorithms to humans on challenging tasks. *ACM Trans. on Applied Perception*, 9(4):1–13, 2012. 2

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. GauGAN: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH Real-Time Live!*, pages 1–1. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR), 2019. 2

[40] 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. 3

[41] PyTorch. Pytorch Model Zoo. `https://pytorch.org/serve/model_zoo.html`, 2021. 5

[42] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *IEEE Eur. Conf. Comput. Vis. (ECCV)*, pages 86–103. Springer, 2020. 2

[43] Brandon RichardWebster, So Yon Kwon, Christopher Clarizio, Samuel E Anthony, and Walter J Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *IEEE Eur. Conf. Comput. Vis. (ECCV)*, pages 252–270, 2018. 2

[44] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 6

[45] 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 *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 618–626, 2017. 5

[46] Bingyu Shen, Brandon RichardWebster, Alice O'Toole, Kevin Bowyer, and Walter J. Scheirer. A study of the human perception of synthetic faces. *arXiv preprint arXiv:2111.04230*, 2021. 3, 4

[47] Harry Sun. Kitware Generated Image Detector. `https://github.com/Kitware/generated-image-detection`, 2021. 2

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2818–2826, 2016. 2, 5

[49] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Gan is a friend or foe? a framework to detect various fake face images. In *ACM/SIGAPP Symp. on Applied Comp.*, pages 1296–1303, 2019. 2

[50] Mateusz Trokielewicz, Adam Czajka, and Piotr Maciejewicz. Perception of image features in post-mortem iris recognition: Humans vs machines. In *IEEE Int. Conf. on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019. 2

[51] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot ... for now. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 2, 8

[52] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNNDetection. `https://github.com/peterwang512/CNNDetection`, 2020. 8

[53] Yaser Yacoob. GAN-Scanner. `https://github.com/yaseryacoob/GAN-Scanner`, 2021. 2

[54] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: a review. In *Int. Joint Conf. on Art. Intell. (IJCAI)*, volume 2020, page 4951. NIH Public Access, 2020. 2

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2921–2929, 2016. 2, 4, 7

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017. 2

[57] zll. BisSeNet – Face Parsing Tool. `https://github.com/zllrunning/face-parsing.PyTorch`, 2019. 6