Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking

İlke Demir  
Intel Corporation  
Santa Clara, CA, USA

Umur Aybars Çiftçi  
Binghamton University  
Binghamton, NY, USA

Figure 1: Overview. Our fake detector extracts eye and gaze features from real (top) and fake (bottom) videos. Frame-based features (blue) are converted to temporal signals and spectral densities to forge signatures (orange). Finally, a deep network (purple) predicts authenticity from signatures, and aggregates predictions.

ABSTRACT

Following the recent initiatives for the democratization of AI, deep fake generators have become increasingly popular and accessible, causing dystopian scenarios towards social erosion of trust. A particular domain, such as biological signals, attracted attention towards detection methods that are capable of exploiting authenticity signatures in real videos that are not yet faked by generative approaches. In this paper, we first propose several prominent eye and gaze features that deep fakes exhibit differently. Second, we compile those features into signatures and analyze and compare those of real and fake videos, formulating geometric, visual, metric, temporal, and spectral variations. Third, we generalize this formulation to the deep fake detection problem by a deep neural network, to classify any video in the wild as fake or real. We evaluate our approach on several deep fake datasets, achieving 92.48% accuracy on FaceForensics++, 80.0% on Deep Fakes (in the wild), 88.35% on CelebDF, and 99.27% on DeeperForensics datasets. Our approach outperforms most deep and biological fake detectors with complex network architectures without the proposed gaze signatures. We conduct ablation studies involving different features, architectures, sequence durations, and post-processing artifacts.

1 INTRODUCTION

With the growth of computational resources and the availability of deep learning models, AI applications thrive on finding new use cases. In particular, generative models invigorate by creating more photorealistic results with each technological advancement. One particular example is in face synthesis domain: The resolution and potent of generated faces have been increasing exponentially since the introduction of Generative Adversarial Networks (GANs) [Goodfellow et al. 2014]. Although realistic synthetic faces are promoted for mixed reality applications, avatars, digital humans, and games; detrimental uses also emerged, such as deep fakes.

Deep fakes are photorealistic synthetic portrait videos generated by deep neural networks, in which the main actor is replaced or reanimated with another person or with a synthetic face. Deep fakes emerged as a novel technique in facial reenactment [Fac
We evaluate our approach on four publicly available datasets, namely the metric domain to learn from the deviation of the coupled odors. Although forgery detection is an established domain with post-generation image manipulation. Consequences are already faced or forecasted, such as impersonation, celebrity porn, and political misinformation. Moreover, interdisciplinary groups and governments provided action plans and acts as a precaution for deep fakes.

To combat malicious use of deep fakes, several detection approaches are proposed, mainly categorized as pure deep learning methods and semantic methods. Deep methods classify authenticity only based on the input image/audio/video dataset, without an intermediate representation. Semantic approaches exploit some priors hidden in real or fake data, such as blinks, heart rates, phoneme-viseme mismatches, or eye reflections. First, we observe that the eye-related fake detectors only consider single artifacts of real eyes, as blinks or reflections. Second, by inspecting digital humans in production, we conclude that inconsistent gaze directions, vergence points, and eye symmetries create the "uncanny valley" for synthetic faces. We want to analyze and integrate these variants in a deep fake detector.

In this paper, we first examine the spatial, temporal, and spectral consistency of eyes and gazes in five domains: (i) The visual domain to cover color and shape deformations, (ii) the geometric domain to include characteristics of 3D vergence points and gaze directions, (iii) the temporal domain to assess the consistency of all signals, (iv) the spectral domain to investigate signal and noise correlation, and (v) the metric domain to learn from the deviation of the coupled signals as approximated by the generators. The rest of the paper discloses our classification network that collates described features into a deep fake detector. Our main contributions are listed below.

- We present the first holistic analysis on deep fake eyes and gazes; proposing geometric, visual, temporal, spectral, and metric features.
- Based on our analysis, we exploit inconsistencies of synthetic eyes to build a robust deep fake detector.
- We achieve significant accuracies in our experiments on different fake generators, with different visual artifacts, and on unseen deep fakes in the wild; using only the proposed eye and gaze features.

We evaluate our approach on four publicly available datasets, namely FaceForensics++, CelebDF, DeeperForensics, Deep Fakes Dataset, achieving 92.48%, 88.35%, 99.27% and 80.0% detection accuracies respectively. We conduct ablation studies with different generators, network architectures, sequence lengths, and feature sets. We compare our approach against three biological detectors and six learning-based detectors on three datasets. Finally, we perform experiments including (1) cross validation between dataset splits, (2) cross-dataset evaluations, (3) weighing the contribution of different feature domains, and (4) assessing the robustness against post-generation image manipulation.

2 RELATED WORK

We briefly introduce face synthesis, then delve into detection methods. Although forgery detection is an established domain with immense literature, we limit our compilation to deep fake detection, especially with biological priors.

2.1 Parametric Face Synthesis

The proliferation of deep generative models enabled realistic face creation and manipulation techniques such as reenactment, attribute or domain manipulation, expression molding, and inpainting. In particular, physically grounded models most of these approaches need some portion of the skin to be visible, decaying in accuracy under heavy make-up or insufficient illumination. Being invariant to skin reflectance, we aim for a more robust approach under these conditions to outperform and/or complement other fake detectors.

2.2 Fake Detection with Biological Priors

Deep fake detectors either employ complex networks to classify videos without an intermediate representation, or they make use of simpler networks with some semantic interpretation of authenticity signatures or generative residuals. The first family of detectors has limitations such as only processing images, detecting the fakes of specific generators, or using small datasets. We refer the reader to a recent survey for the details of each network. The second branch extracts intermediary physiological, biological, or physical signals, then learns the authenticity from their correlations. Such approaches enclose blinks, head pose, eye reflections, heart rates, facial attributes, depth of field, eye brows, phoneme-viseme mismatches, and affective cues. Most of these approaches need some portion of the skin to be visible, decaying in accuracy under heavy make-up or insufficient illumination. Being invariant to skin reflectance, we aim for a more robust approach under these conditions to outperform and/or complement other fake detectors.

2.3 Gaze-based Fake Detection

Recently, proposes a robust detector by monitoring irregular blinking patterns successfully. Further investigation of eye traits reveals that the corneal specular highlights for real eyes have strong similarities while those for GAN-synthesized eyes are different. Introduces a physiological/physical detection method based on that observation. However, this approach
exposes only one feature of eyes and it is still sensitive to both illumination and pose variations.

In addition to blinking and corneal reflection patterns, gaze is also an important trait to distinguish real and fake videos. Nevertheless, there is no existing work to utilize the eye gaze discrepancy for identifying deep fakes. This paper is to assemble such important but subtle signals to tackle the deep fake detection problem.

To address this issue, essential steps include eye landmark extraction and gaze estimation. In general, eye detection techniques cover (1) shape-based [Hansen and Pece 2005], (2) appearance-based [Hansen and Ji 2010], or (3) hybrid detection [Hansen et al. 2002; Santini et al. 2018]. Gaze estimation methods rely on 2D features [Hansen and Ji 2010] or 3D eye models [Dierkes et al. 2019; Hansen and Ji 2010; Reale et al. 2013]. Recent developments using CNN based gaze estimation has also improved gaze estimation accuracy [Akinleye and Blignaut 2020; Krafa et al. 2016; Wood et al. 2015; Zhang et al. 2019]. In this paper, we utilize [Baltrusaitis et al. 2018] for extracting eye landmarks, and [Wood et al. 2015] for extracting pupil and iris landmarks in addition to estimating the gaze direction. We also experiment with other approaches and implementations, however they do not provide iris tracking [Zhang et al. 2019], require special software/hardware [Kassner et al. 2014; Zieliski 2007], calibration [Dalmajer et al. 2014], or configured for mobile [Krafa et al. 2016] or online [Deja 2019] users.

3 3D EYE AND GAZE FEATURES

In order to learn the characteristics of fake gazes, we investigate several eye and gaze features and their contributions to differentiate authenticity. We explore these signals in five subsets, although some features belong to multiple domains. In addition to Figure 1 briefly depicting our proposed features, Figure 2 provides examples for all of them. We also categorically document all features in Table 7 including their normalizations and contributions to the signatures.

3.1 Visual Features

Visual features \( V \) include color \( C \), area \( A \) of eye \( X_E \), iris \( X_I \), and pupil \( X_P \), from left \( X^L \) and right \( X^R \) eyes:

\[
V = \{C^L, C^R, C^L_p, C^R_p, A^L, A^R, A^L_p, A^R_p, \rho^L, \rho^R, \rho^L_p, \rho^R_p, |A^L - A^R|, |C^L - C^R|, |C^L_p - C^R_p| \}
\]

(1)

As shown in Figure 2d, areas correspond to the pixels that are included in the larger region (i.e., iris) but not in the smaller region (i.e., pupil). The color is computed in the CIElab [Connolly and Fleiss 1997] space in order to create illumination independent features, then averaged within the corresponding area (Figure 2c). Assuming that the color difference of iris and pupils from left and right eyes would stay constant (in an illumination invariant color space), we also append color differences (Eqn. 1). Note that we excluded the eye color as it is brittle against illumination and alignment errors.

3.2 Geometric Features

Geometric gaze features \( L_G \) include left and right gaze vectors \( G^L, G^R \), 3D gaze point \( \rho \), 3D gaze point approximation \( \hat{\rho} \), and the value of the cost function at the approximation \( \delta_p \). Geometric eye features include the eye distance \( d \), pupil distance \( d_p \), and the difference of areas \( |A^L - A^R| \):

\[
L_G = \{G^L, G^R, \rho, \hat{\rho}, \delta_p \}
\]

(2)

\[
L_E = \{d, d_p, |A^L - A^R|, |A^L_p - A^R_p| \}
\]

(3)

We use [Wood et al. 2015] to extract gaze vectors and pupil centers. Then we shoot rays from the pupil centers along the gaze vectors to find their intersection. We aim to optimize for the midpoint of the closes points of the two rays, because it is possible that the rays

Figure 2: Eye and Gaze Features. Real (left) and fake (right) videos (a), exhibit different geometric (a,b), visual (c,d), and metric (e) features. We draw gaze vectors \( G^* \), 3D gaze points \( \rho \), and pupil distances \( d \) on (a).
are skew lines in 3D. In order to approximate the intersection of two 2D rays in 3D space, we use a least squares optimization with $1e-08$ threshold. In addition to the approximated 3D gaze point $\rho$ (green points in Figure 2a), we compute the distance between the projections of $\rho$ onto $G_l$ and $G_r$ (orange vectors in Figure 2a), indicated as $\hat{\rho}$ (the green line in fake Figure 2a). We also incorporate the cost at the solution as $\delta_{\rho}$.

### 3.3 Temporal Features

As of now, the features we present are computed per frame. However we are less interested in their exact values but more interested in their consistency, therefore we introduce the concept of sequences: a fixed $\omega$-length of consecutive frames from a video. We divide all videos into $\omega$ length sequences and create signals from the previously defined visual and geometric features, in order to perform our temporal consistency analysis.

$$ T = \{V^i, L^i_G, L^i_E\text{ such that } i \in [0, \omega) \} \quad (4) $$

Although $\omega$ can be set as a user parameter for detector sensitivity, we experiment with different $\omega$ values in Section 6.2.

### 3.4 Metric Features

In addition to temporal consistency, spatial coherence is an embedded property of authentic videos, especially for symmetries. We employ cross correlation $\phi(\ast, \ast)$ of left/right, eye/gaze features to utilize their spatial coherence as an authenticity cue (Figure 2e).

$$ M = \{\phi(C^i_l, C^i_r), \phi(C^i_p, C^i_p), \phi(A^i_l, A^i_l), \phi(A^i_p, A^i_p), \phi(A^i_E, A^i_E), \phi(G^i_l, G^i_r)\} \quad (5) $$

### 3.5 Spectral Features

Although we launch our analysis in temporal domain, the environment or the face detection algorithm can introduce noise which deteriorates our temporal features. Thus, we calculate the power spectral density $\Theta(\ast)$ of all the previously defined signals. Frequency bands play a crucial role in differentiating noise from actual gaze or eye movements.

$$ S = \{\Theta(V^i), \Theta(L^i_G), \Theta(L^i_E), \Theta(M)\} \quad (6) $$

In total, we conclude with $40 \times 3 \times \omega$ size signatures with the combination of $S$ and $T$, all of which are listed in Table 7.

### 4 GAZE AUTHENTICITY

Having defined our feature space, we explore and analyze the behavior of the predefined features on real and fake gazes. As different generative models create different levels of realism for synthetic faces, they also leave different traces behind [Ciftci et al. 2020b]. We experiment with several generator outputs [Dee [n.d.]; Fac [n.d.]; Thies et al. 2019, 2016] corresponding to a real sample. For the domains in which the artifacts of different generators are not significantly varying, we discuss them collectively.

### 4.1 Color and Size Deformations

The visual features $V$, namely color and size of the eye regions, tend to stay coherent for real videos, even if there is some movement involved. Figure 3 demonstrates close ups of real (first column) and realistic fake (others) eyes of a sample from FF++ dataset. Even for the best fake generator (last column), some of the resolution, striding, or alignment artifacts become visible in close ups. We also emphasize the color differences of those real-fake pairs (even rows) and area changes of regions (i.e., the second column). We delve into those deformations to detect fake videos.

---

**Figure 3: Eye Deformations.** Real (first column) and fake (others) eyes are demonstrated to emphasize color differences (even rows) and size differences (i.e., the areas of eye regions in the second column).
Figure 4: 3D Gaze Points of a real video (a) and its fake counterparts (b-e) are observed for ω frames of a video. Fake gazes exhibit noise (b), miss some saccades (c,d), and have irregular distribution (e). Sample problematic points are boxed in red.

4.2 Geometric Attestation
A visually invisible but computationally powerful way to verify natural gazes is to geometrically validate them. Intuitively, the gaze vectors in real videos converge at a point in 3D space. However, slight deformations of pupil centers or gaze directions may lead to two completely off-plane gaze vectors that never intersect (i.e., Figure 2a fake, or Figure 1 fake). We use vergence points to verify authentic gazes, such that (i) ρ exists and is in front of the face, (ii) $\hat{\rho} = 0$ or is relatively small, and (iii) $\delta \rho \times \rho$ is close to $\rho$ within the optimization $\epsilon$. Another nice property of $\rho_t$ is that it corresponds to where we naturally look, which has its own form and properties (i.e., smooth pursuits, saccades, etc.) as shown in Figure 4.

4.3 Temporal Consistency
To illustrate temporal consistency of gaze features, we compare the form of 3D gaze points from real (first) and fake (others) versions of a video in Figure 4, color coded by time. In red boxes, we observe that fake gazes exhibit noise (Figure 4b), miss sudden motions (4c,d), and have ragged distributions (4e) compared to real gazes.

Similarly, we show exemplary temporal signals $G^r_1, G^r_0, |A^r_1 - A^r_0|, A'^r_1, \hat{\rho}_t, \rho_t$ in Figure 5, of real (green) and fake (blue and orange) videos. Fake gazes may break temporal consistency because of over smoothing (i.e., $G^r_1[0], G^r_0[0]$), or noise (i.e., left-right iris size changes by $30mm^2$), or random distributions (i.e., gaze error).

4.4 Spectral Artifacts
Depending on video resolution, post-processing operations, and illumination, real videos may produce imperfect signals. Those imperfections differ in nature from the generative noise, and the distinctions are best observed in spectral domain. Therefore, we expert power spectral densities of our temporal signals in our analysis, similar to the detection approaches in biological domain [Ciftci et al. 2020a]. Synthetic features produce a richer spectra than reals, as we expect most of our real features to remain constant or coherent. Figure 6 shows the signature blocks, as well as zoom ins to several features, including PSDs of real and fake pupil areas $\Theta(A^r_p)$ at top right. The colorful bands indicate different spectral densities, so the fake PSD gleams whereas the real PSD contains only a few bands.

Figure 5: Temporal Signals. Real (green) and fake (orange and blue) signals of (a) $G^r_1[0], (b) G^r_0[0], (c) G^r_1[1], (d) G^r_2[1], (e) |A^r_1 - A^r_0|, (f) A'^r_1, (g) \hat{\rho}_t$ and (h) $\rho_t$. Fake gazes cover less range in (a, b) and they exhibit more noise in (c,d). Fake iris area differences has a $30mm^2$ range in (e) and fake pupil areas can differ by $14mm^2$ in (f).
4.5 Symmetry Flaws

Early signs of detecting GAN-generated images were claimed to be asymmetric eyes, ears, and earrings. As the convergence of GANs needs more resources than training other deep networks, results from sooner epochs still exhibit asymmetric eyes as there is no explicit control for coupling eyes. We exploit this shortcoming and experiment with the cross correlation of symmetric pairs of eye and gaze features, with the hypothesis of fake features breaking the symmetry. Our experiments yield that \( \Phi = \{\phi(C^l_I, C^r_I), \ldots, \phi(G^l_I, G^r_I)\} \) (Eqn. 5) increases the detection accuracy.

5 LEARNING AUTHENTICITY BY EYE & GAZE FEATURES

In this section, we build a deep fake detector using our features from Section 3 and our analysis from Section 4.

5.1 Datasets

As our hypothesis of utilizing eye and gaze features for deep fake detection needs to be tested on several deep fake generators, so we need datasets with multiple generators and in the wild samples. For this purpose, we use FaceForensics++ (FF++) [Rossler et al. 2019] dataset with known and Deep Fakes dataset (DF) [Ciftci et al. 2020a] with unknown sources. FF++ dataset includes 1,000 real videos and 4,000 fake videos, generated by four generative models which are FaceSwap [Fac [n.d.].a], Face2Face [Thies et al. 2016], Deepfakes [Dee [n.d.]], and Neural Textures [Thies et al. 2019]. DF dataset, on the other hand, contains 140 in-the-wild deep fakes collected from online sources, thus they are generated by various unknown generators, undergone various post-processing and compression operations. Such deep fakes do not meet any quality standards and their source generative model is unknown, in addition to experiencing occlusion, illumination, etc. artifacts.

We complement our experiments with CelebDF [Li et al. 2020b] (CDF) and DeeperForensics [Jiang et al. 2020] (DFor) datasets to further justify our claim of trusting eye and gaze features as a deep fake litmus. CDF includes 590 real and 5639 fake videos created by their own generative model. DFor contains 50,000 real videos of 100 people captured under controlled conditions, and 1000 fake videos, extrapolated to 10,000 videos by 10 image distortion algorithms.

5.2 Network Architecture

We treat authenticity detection as a binary classification problem. Our network is made up of three dense layers with 256, 128, and 64 nodes, followed by a last dense layer with 2 nodes. As discussed in the previous section, our features are representative and interpretable, nonetheless our dense architecture may tend to overfit without regularization. Therefore we include batch normalization and dropout layers in between the dense layers to prohibit the network from overfitting. We also add leaky ReLU activations to add non-linearity, and end with a sigmoid activation to obtain the outputs. The detailed model architecture can be observed in Table 8.

5.3 Training Setup

After we extract feature sets \( S \) (Eqn. 6) and \( T \) (Eqn. 4), we normalize all features to be in \([0, 1]\), by shifting and scaling with different values (Table 7). Then we concatenate the processed signals to create our input tensors, or signatures (Figure 1, orange boxes), of size \((40, 6, 3)\). The decomposition of our input tensors are depicted on the left side of Figure 6 and in Table 7. We use the following hyper-parameters determined empirically: Adam optimizer with a learning rate of \(1e - 04\), 0.3 dropout probability, batch size of \(32\), and \(0.2\) threshold for leaky ReLU layers. We train the models for 100 epochs and validate every 10 epochs.

For training our networks, we gather equal number of real and fake videos from the aforementioned datasets, and create train and test subsets with a 70%-30% split (unless otherwise is noted in the results section). This split is completely random for all datasets and it does not affect the results as demonstrated in Section 6.1.3. Only for CDF, authors provide a predetermined test set of 178 real and 340 fake videos for benchmarking results, thus we use the provided subsets. For DFor, we use 1000 fake videos with no post-processing and 616 real videos with uniform lighting, frontal view, and all expressions. Each train/test pair belongs to one dataset, except cross dataset evaluation in 6.1.3.

5.4 Video Predictions

As our signatures are per sequence (Section 3.3), we need to aggregate sequence predictions from our network into video predictions. Although deep fakes mostly focus on stable faces, they can contain a moving actor, illumination changes, occlusions, and frames without faces. To accommodate for these irregularities, we can trust the
average of the sequence predictions, e.g., if the mean of sequence predictions is closer to fake, we accept the whole video as fake. Another approach is majority voting, where each sequence of a video contributes to a fake or real vote. Instead of counting each sequence as fake or real, we can use the confidence of sequence predictions as a voting score. Or more suitably, we can level with the negative likelihood of sequence predictions contributing to the video prediction. In order to accomplish that, we use the average of log of odds of prediction confidences as the voting scheme:

\[
\frac{1}{N_p} \sum_{i < N_p} \log \left( \frac{p_i}{1 - p_i} \right)
\]  

(7)

6 RESULTS

We use OpenFace [Baltrusaitis et al. 2018] for eye and face tracking, OpenCV for image processing, SciPy and NumPy for scientific computing, Keras for deep learning, and Matplotlib for visualizations. We train our system on a single NVIDIA GTX 1060 GPU, and a full training loop of 100 epochs takes less than an hour. We continue with our evaluations on different datasets, from different generators, comparisons, validations, experiments, and ablation studies.

6.1 Evaluations and Comparisons

We evaluate our approach on several single and multi-source datasets discussed in Section 5.1 and list our sequence and video accuracy percentages (abbreviated as S. Acc. and V. Acc.) in Table 1. CDF and DFor represent single generator datasets, FF++ is a known multi-generator dataset, and DF is a completely in-the-wild dataset. Our approach detects fakes with 88.35%, 99.27%, 92.48%, and 80% accuracies, respectively. We also evaluate our detector on four known generators of FF++ separately. We observe that for GANs that generate masks without eyes, such as [Thies et al. 2019, 2016], our accuracy is low as expected – they simply do not modify the eyes.

Table 1: Classification accuracies on FF++ generators and single / multi / unknown source datasets.

| Source          | S. Acc. | V. Acc. |
|-----------------|---------|---------|
| Real            | 80.36   | 91.30   |
| DeepFakes       | 81.02   | 93.28   |
| Face2Face       | 58.18   | 59.69   |
| FaceSwap        | 80.63   | 91.62   |
| Neural Tex.     | 57.46   | 57.02   |
| CDF [Li et al. 2020b] | 85.76   | 88.35   |
| DFor [Jiang et al. 2020] | 95.57   | 97.27   |
| FF++ [Rossler et al. 2019] | 83.12   | 92.48   |
| DF [Ciftci et al. 2020a] | 79.84   | 80.00   |

6.1.1 Comparison with Biological Detectors. In Table 2, we compare our approach against three other biological deep fake detectors built on blinks [Li et al. 2018], which uses only the eye region similar to our approach; head pose, which uses similar information from [Baltrusaitis et al. 2018]; and heart rates [Ciftci et al. 2020a], which uses the information from exposed skin. For a fair comparison, we use the same train and test subsets from FF++ and DF for all approaches as discussed in Section 5.3, where FF/FS and FF/DF corresponds to FaceSwap and DeepFakes subsets of FF++, respectively. Observing first two rows, blink patterns and head poses are no longer a strong indicator of fake content (50-67%), as generators learn to fake these features. In contrast, heart rates (third row) contain significant authenticity clues similar to our approach (on the average 2.8% better on FF++).

Table 2: Comparison with state-of-the-art biological (1-3) and deep (4-5) detectors on FF and DF. Same data and procedure are used, except results (*) from [Rossler et al. 2019].

| Approach          | FF/DF | FF/FS | DF   |
|-------------------|-------|-------|------|
| Blink [Li et al. 2018] | 67.14 | 54.15 | 57.69|
| Head pose         | 55.69 | 50.08 | 67.85|
| PPG [Ciftci et al. 2020a] | 94.87 | 95.75 | 91.07|
| Inception [Szegedy et al. 2016] | 65.4* | 54.4* | 68.88|
| Xception [Chollet 2017] | 74.5* | 70.9* | 75.55|
| Ours              | 93.28 | 91.62 | 80.00|

Based on the slightly better results, we investigate and compare [Ciftci et al. 2020a] and our approach. We observe that the failure cases are mutually exclusive, as demonstrated in Figure 7. For false negatives (left half), PPG signals extracted from the skin may be smooth enough to trick [Ciftci et al. 2020a] to classify a fake video as real, however, our approach correctly classifies the same fake video with 84.93% confidence from the eye artifacts (bottom row). For false positives (right half), heavy make up on the skin may deteriorate PPG signals and lead [Ciftci et al. 2020a] to detect a real as fake, however, our approach correctly classifies it due to coherent eye signatures. This motivates us to propose ensemble networks, merging biological detectors working in separate domains to cover for each others’ limitations.

Figure 7: Comparison to FakeCatcher [Ciftci et al. 2020a]. A fake video classified as real by [Ciftci et al. 2020a] and as fake by our approach. Artifacts are more visible when zoomed in (bottom). Fake frames from [Rossler et al. 2019].
6.1.2 Comparison with Other Networks. We also compare our approach against six standard deep network architectures, without integrating our eye and gaze features. In Table 3, we listed [Szegedy et al. 2016] and [Chollet 2017] results on three datasets, and in Table 2, we listed [Afchar et al. 2018] and [Tariq et al. 2018] variation results in addition to those of two basic architectures, a three-layer CNN and a ConvLSTM. Our motivation in these comparisons is to justify that the eye and gaze features we formulate contain sufficient information to enable our simple architecture to perform better than deeper and more complex models with more capacity and full frame information.

Table 3: Comparison with other deep fake detectors on DF and FF datasets. Same train/test data and procedure are used for all, except marked results (*) from [Rossler et al. 2019].

| Approach           | Dataset | Acc. |
|--------------------|---------|------|
| 3-layer CNN        | DF      | 48.88|
| ConvLSTM           | DF      | 48.83|
| [Tariq et al. 2018] V3 | DF    | 73.33|
| [Tariq et al. 2018] ensemble | DF | 80.00|
| MesoNet [Afchar et al. 2018] | FF/DF | 87.3*|
| MesoNet [Afchar et al. 2018] | FF/FS | 61.2*|

We obtained the results on FF++ from their benchmark, concluding with our approach having 5.98% better accuracy than the second best [Afchar et al. 2018], verifying the contribution of our authenticity formulation. For the results on DF, we use the same setup for training all of the models, namely, the same train/test splits with the same hyper parameters listed in Section 5.3. The only difference is that others are trained with source frames (real or fake) without preprocessing, and ours is trained with our signatures. Despite the in-the-wild setting of DF dataset, our approach outperforms all listed detectors – except having the same accuracy with the complex ensemble network of [Tariq et al. 2018] utilizing full frame information with two order of magnitude more parameters than our model. We emphasize that we only input eyes in our fake detector, and still obtain better results, proving the power of the proposed eye and gaze features.

6.1.3 Cross Validation. Giving the benefit of doubt to our random train/test split of datasets, we perform two 5-fold validations with different splits. In the first one, we create five different train-test subset pairs by randomly splitting the aforementioned FF++ dataset as 70%-30%. Then we train and test our approach independently five times on these subsets, with the same procedure discussed in Section 5.3. Five test accuracies have a mean of 90.6% with 1.5% standard deviation.

In the second validation test, we use 80%-20% split to create non-overlapping subsets. In other words, 2000 training videos (1000 fake, 1000 real) are randomly split to mutually exclusive subsets of 400 videos (200 fake, 200 real) as test sets. For each test set, remaining 1600 videos are used for training. Five train and test sessions yield 90.3% mean accuracy with 1.1% standard deviation, validating that our results are independent of the random dataset splits.

Lastly, we conduct a cross dataset evaluation between CDF and DFor datasets. Using the same train and test setup, we train a model on CDF and test on DFor, then we train a model on DFor and test on CDF. We obtain 80.72% and 73.68% accuracies respectively, which are close to our in-the-wild results.

6.2 Experiments

We summarize the effects of different sequence durations, using several feature sets, under two image-space artifacts.

6.2.1 Sequence Duration. As our method benefits from both temporal and spectral artifacts of continuous signals, such as 3D gaze points in time or frequency bands of iris colors, selection of $\omega$ is crucial. We experimentally validate our selection of $\omega$ with an ablation study (Table 4), testing different sequence lengths of 16, 32, 64, and 128 on FaceSwap subset of FF++. We monitor that short sequences may contain only noise or missing faces, while long sequences saturate authenticity cues. We can recover from the former during video prediction, but not from the latter one. In short, we conclude on the optimum value of $\omega = 32$ from Table 4.

Table 4: Sequence Durations. We validate the optimum choice of $\omega = 32$ by sequence and video accuracies.

| $\omega$ | S. Acc. | V. Acc. |
|---------|---------|---------|
| 16      | 77.72   | 79.79   |
| 32      | 80.63   | 86.93   |
| 64      | 81.62   | 85.04   |
| 128     | 79.45   | 88.96   |

6.2.2 Feature Sets. Next, we train and test our system on FF++ dataset with different feature collections, namely (i) as is, only with (ii) spectral, (iii) temporal, (iv) geometric, (v) visual, and (vi) metric features, (vii-ix) only with raw gaze variants, (x) without metric features, and (xi) without geometric features. All other hyper parameters are held the same. According to Table 5, all five domains of features are contributing to the detection accuracy, however metric features have relatively less importance. Although metric-only results are high (row 6), results without metric features are higher (row 10). We speculate that the size of metric features are relatively smaller in the signature, so they influence the classification less than others, and they can be compensated by other features.

Table 5: Ablation study with and without some of the feature domains from Section 3.

| Condition            | S. Acc. | V. Acc. |
|----------------------|---------|---------|
| All                  | 83.12   | 92.48   |
| Spec-only            | 75.42   | 84.46   |
| Temp-only            | 76.39   | 85.28   |
| Geo-only             | 79.84   | 91.82   |
| Visual-only          | 78.56   | 88.46   |
| Metric-only          | 78.52   | 87.79   |
| Raw gaze             | 72.88   | 81.04   |
| Gaze vectors         | 72.28   | 80.80   |
| Rows 6 and 7         | 72.66   | 81.39   |
| No metric            | 79.45   | 88.96   |
| No geometric         | 79.52   | 87.98   |
6.2.3 Postprocessing Artifacts. Generative noise is not the only artifact that deep fakes exhibit, especially in the wild. Post-processing operations, such as compression and resampling, may deteriorate focused signals. In order to test the robustness of our approach, we approximate the post-processing artifacts with two commonly used image processing operations: Gaussian blur and median filtering. All other parameters are kept the same. As enumerated in Table 6, our system is robust against Gaussian and median noise, because effects of such artifacts are minimal for the eye region. As long as the eyes are still visible in videos, features from different domains enforces the detection accuracy to remain the same – i.e., high resolution eyes enable the contributions of geometric features, and low resolution eyes enable the contributions of visual features. This experiment supports the claim that eye/gaze based fake detection is more persistent when other biological fake detectors fail under post-processing operations, compared to Table 14 of [Citcet al. 2020a]. Their accuracy drops by 20.66% under 9x9 Gaussian blur, whereas our accuracy only drops by 0.37%.

Table 6: Robustness. We manually add Gaussian blur and median noise to input videos and evaluate our robustness against such post-processing artifacts.

| Kernel | Initial | Gaussian Blur | Median Filter |
|--------|---------|---------------|---------------|
|        | 92.48 | 89.13 | 85.88 | 92.11 | 91.27 | 90.77 | 89.76 |

7 CONCLUSION AND FUTURE WORK
In this paper, we exhaustively analyze real and fake gazes to answer “Where Do Deep Fakes Look?”, and morph our findings into a deep fake detector. To the best of our knowledge, our paper conducts the first extensive analysis of deep fake gazes, and it is the first approach to build a detector solely based on holistic eye and gaze features (instead of cherry-picking a few). We evaluate our approach on four datasets, compare against biological and deep detectors, and conduct several ablation studies. As eyes in real videos exhibit natural signatures, which fake ones yet mimic consistently, we put forward that our visual, geometric, temporal, metric, and spectral features to be integrated to existing fake detectors.

One advantage of our approach is that our signatures are based on authenticity signals instead of generative noise. Blind classifiers without ‘real’ priors are prone to adversarial attacks, as shown in [Hussain et al. 2021] for [Chollet 2017] and [Afchar et al. 2018]. In contrast, fake eye/gaze physics should be spatio-temporally accurate in order to fool our classifier.

On the other hand, as the arms race continues, we also believe that our findings can be used to create better deep fakes with consistent fake gazes. Such synthetic gazes with increased photorealism can be useful for virtual/augmented reality applications, avatars, data augmentation, transfer learning, and controlled learning and testing environments for eye tracking research.

Another future direction is to include gaze classification for our 3D gaze points in order to boost the accuracy of our fake detector. We observe that time-varying 3D gaze points contain much richer biological information that can be utilized if we can interpret them further – starting from gaze movements such as saccades, fixations, etc. We anticipate that our first step in deciphering fake gazes will lead the way for other deep learning approaches to be inherited for eye or gaze based deep fake detection techniques.

REFERENCES
[1]. Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy? https://www.lawfareblog.com/deep-fakes-loomimg-crisis-national-security-democracy-and-privacy. Accessed: 2018-11-15.
[2]. DeepFakes. https://github.com/deepfakes/face swap. Accessed: 2020-03-16.
[3]. FaceSwap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2020-03-16.
[4]. FaceSwap-GAN. https://github.com/shaolanlu/face swap-GAN. Accessed: 2020-03-16.
[5]. Deepfake porn is blwing up on Reddit, thanks to artificial intelligence. https://www.theverge.com/2018/1/24/16929418/deepfake-porn-artificial-intelligence-redi. Accessed: 2018-11-15.
[6]. Deep Fakes Prohibition Act of 2018. https://www.congress.gov/bill/115th-congress/senate-bill/3805/text. Accessed: 2020-11-25.
[7]. When seeing is no longer believing: Inside the Pentagon’s race against deep fake videos. https://www.cnn.com/2019/01/21/tech/deepfakes/. Accessed: 2020-11-15.
[8]. GazePointer. https://sourceforge.net/projects/gazepointer/.
[9]. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments. Behavior research methods 46, 4 (2014), 913–921.
[10]. Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking ETRA ’21 Full Papers, May 25–27, 2021, Virtual Event, Germany
Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. 2017. Two-Stream Neural Networks for Tampered Face Detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 1831–1839. https://doi.org/10.1109/CVPRW.2017.229

Piotr Zieliński. 2007. Opengazer: open-source gaze tracker for ordinary webcams. http://www.inference.phy.cam.ac.uk/opengazer/

A SIGNATURE COMPOSITION

In Table 7, we list elements of our input tensor (i.e., the signature) that collates the spectral and temporal versions of the aforementioned visual, geometric, and metric features. SS means “shift by min+scale by max” and \( d^+ \) corresponds to maximum IPD. Multiple features occupying one row corresponds to features being merged into a three channel value. Channels with stars (*) indicate that a single channel value is duplicated into multiple channels per its importance. Cross correlation features intake double signals and output single signals. The signatures include 24 visual, 24* geometric, and 12 metric features. Finally, all \( |T| = 20 \times 3 \times \omega \) values are transformed and resized to the spectral domain, creating \( |S| = 20 \times 3 \times \omega \) values, constituting the \( 40 \times \omega \times 3 \) size signatures.

Table 7: Signatures. We enlist the elements, their parts, sides, channels, and normalizations that make up our fake and real signatures.

| Set   | Feature | Part | Side | Chan. | Norm. |
|-------|---------|------|------|-------|-------|
| Visual| Color   | Iris | Left | 3     | /256  |
| Visual| Color   | Iris | Right| 3     | /256  |
| Visual| Color   | Pupil| Left | 3     | /256  |
| Visual| Color   | Pupil| Right| 3     | /256  |
| Visual| Area    | Eye  | Left | 1     | SS    |
| Visual| Area    | Eye  | Right| 1     | SS    |
| Geometric| Area  | Eye  | Left-Right | 1 | /\(d^+\) |
| Visual| Area    | Iris | Left  | 1     | SS    |
| Visual| Area    | Iris | Right | 1     | SS    |
| Geometric| Area  | Iris | Left-Right | 1 | SS    |
| Visual| Area    | Pupil| Left  | 1     | SS    |
| Visual| Area    | Pupil| Right | 1     | SS    |
| Geometric| Area  | Pupil| Left-Right | 1 | /\(d^+ − 1\) |
| Visual| Color   | Iris | Left-Right | 3     | /256  |
| Visual| Color   | Pupil| Left-Right | 3     | /256  |
| Geometric| Vector | Gaze | Left  | 3     | SS2   |
| Geometric| Vector | Gaze | Right | 3     | SS2   |
| Geometric| Point  | Gaze | -     | 3     | SS    |
| Geometric| Error  | Gaze | -     | 3     | /\(d^+\) |
| Geometric| Cost   | Gaze | -     | 3*    | SS    |
| Geometric| Distance| Eye  | -     | 3*    | SS    |
| Geometric| Distance| Pupil| -     | 3*    | SS    |
| Cross  | Color   | Iris | Both | 2x3 > 3 | SS  |
| Cross  | Color   | Pupil| Both | 2x3 > 3 | SS  |
| Cross  | Area    | Iris | Both | 2 > 1   | SS  |
| Cross  | Area    | Pupil| Both | 2 > 1   | SS  |
| Cross  | Area    | Eye  | Both | 2 > 1   | SS  |
| Cross  | Vector  | Gaze | Both | 2x3 > 3 | SS  |

B NETWORK LAYERS

We log our network layers, their arguments, and their output shapes in Table 8. Note that the order of layers and the network architecture are finalized empirically and theoretically [Han et al. 2017; Li et al. 2019]. We would also like to note that the main motivation of our paper is to analyze eye and gaze behavior of various deep fakes, and this simple but strong network is a part of the verification process for our findings. For production-level accuracies, we encourage researchers to invest in neural architecture search for building better alternatives utilizing our eye and gaze features.

Table 8: Network Layers. We document sizes, arguments, and outputs of each layer in our network.

| Layer     | Arguments | Output |
|-----------|-----------|--------|
| Signature | -         | 40x64x3 |
| Flatten   | -         | 7680x1x1 |
| Batch Norm| -         | 7680   |
| Dense     | 256       | 256    |
| Batch Norm| -         | 256    |
| Leaky ReLu| 0.2       | 256    |
| Dropout   | 0.3       | 256    |
| Dense     | 128       | 128    |
| Batch Norm| -         | 128    |
| Leaky ReLu| 0.2       | 128    |
| Dropout   | 0.3       | 128    |
| Dense     | 64        | 64     |
| Dense     | 2         | 2      |
| Sigmoid   | -         | 2      |