Is It Certainly a Deepfake?
Reliability Analysis in Detection & Generation Ecosystem

Neslihan Kose
Intel Labs

Anthony Rhodes
Intel Labs

Umur Aybars Çiftçi
Binghamton University

İlke Demir
Cauth AI

{neslihan.kose.cihangir, anthony.rhodes}@intel.com, uciftci@binghamton.edu, ilke@cauth.ai

Abstract

As generative models are advancing in quality and quantity for creating synthetic content, deepfakes begin to cause online mistrust. Deepfake detectors are proposed to counter this effect, however, misuse of detectors claiming fake content as real or vice versa further fuels this misinformation problem. We present the first comprehensive uncertainty analysis of deepfake detectors, systematically investigating how generative artifacts influence prediction confidence. As reflected in detectors' responses, deepfake generators also contribute to this uncertainty as their generative residues vary, so we cross the uncertainty analysis of deepfake detectors and generators. Based on our observations, the uncertainty manifold holds enough consistent information to leverage uncertainty for deepfake source detection. Our approach leverages Bayesian Neural Networks and Monte Carlo dropout to quantify both aleatoric and epistemic uncertainties across diverse detector architectures. We evaluate uncertainty on two datasets with nine generators, with four blind and two biological detectors, compare different uncertainty methods, explore region- and pixel-based uncertainty, and conduct ablation studies. We conduct and analyze binary real/fake, multi-class real/fake, source detection, and leave-one-out experiments between the generator/detector combinations to share their generalization capability, model calibration, uncertainty, and robustness against adversarial attacks. We further introduce uncertainty maps that localize prediction confidence at the pixel level, revealing distinct patterns correlated with generator-specific artifacts. Our analysis provides critical insights for deploying reliable deepfake detection systems and establishes uncertainty quantification as a fundamental requirement for trustworthy synthetic media detection.

1. Introduction

Recently, synthetic content has become a part of our daily lives with the proliferation of generative models. Specifically, human faces have always been the focus of computer vision algorithms with a growing interest, pursuing the same paradigm with generative models since the introduction of Generative Adversarial Networks [28] (GAN) in 2014. In the intersection was born deepfakes: images, audio clips, or videos, where the actor or the action of the actor in the content is fabricated using deep generative models.

Although synthetic content creation brought up many positive use cases, deepfakes are usually exploited in politics, entertainment, and security [1, 2]; causing the need for a line of defense [20]. Deepfake detectors are proposed to satisfy this need, however, their generalization and robustness vary depending on the intermediate signals they use, their model architecture, and the datasets they are trained on. Contemporary deepfake detection research has predominantly pursued accuracy maximization through increasingly sophisticated neural architectures and training methodologies. However, this accuracy-centric evaluation paradigm fundamentally overlooks a critical dimension: **prediction uncertainty**. As detection systems transition from controlled laboratory environments to real-world deployment scenarios, understanding when and why detectors exhibit uncertainty becomes as crucial as their peak performance capabilities. As opposed to evaluating them based on traditional accuracy and AUC metrics on many datasets, we analyze their core performance and understanding of data, by uncertainty analysis, which has not garnered much attention in the research community. Understanding model response of these deepfake detectors help compare their generalization in-the-wild, overfitting to the artifacts, performance beyond the training distribution, robustness against adversarial attacks, and effectiveness in source detection. This analysis is the key to focus on robust and reliable deepfake detectors to establish a trusted online future.

Current detection methodologies broadly fall into two paradigmatic approaches, using the existence or non-existence of priors to classify content as real or fake. These priors are either irreproducible authentic signals (e.g., corneal reflections [33], blood flow [16], phoeneme-viseme mismatches [9]) in real data or small generative artifacts in fake data [8, 19, 21], often achieving impressive performance on in-distribution benchmarks while struggling with cross-domain generalization. As per generative models, face generation [34, 54], face swapping [5, 44], and face reenactment [53, 66] methods manipulate using different techniques operating on different facial regions. As a result, different generative models leave different residual traces behind [72]. Those traces are directly correlated to the detector response, tying deepfake detectors and generators. This connection resumes and is observable in the uncertainty estimation of detectors, aggregated by model-specific uncertainty contributors such as architecture, signal manifold, and training data. Despite their widespread adoption across industry and academic applications, the uncertainty characteristics of these complementary approaches remain poorly understood. This knowledge gap proves particularly concerning given the high-stakes nature of deepfake detection deployments, limiting our understanding of when these systems can be trusted in production environments, where false positive classifications can irreparably damage reputations while false negatives enable malicious manipulation campaigns.

In this paper, we analyze uncertainty of various deepfake detectors in the presence of data generated by various deepfake generators. We use this analysis comprehensively to compare the robustness and reliability of detectors, to explain the detector response towards different generative sources. As the uncertainty can stem from multiple sources, multiple uncertainty measures are needed for an in-depth analysis. Our contributions include,

- comprehensive uncertainty quantification across diverse detector and generator architectures
- generator-specific uncertainty analysis revealing how different synthesis paradigms influence detector confidence patterns and calibration quality
- image-, region-, and pixel-wise uncertainty comparisons of authenticity- and fakery-based deepfake detectors to provide an in-depth understanding of detection performance,
- source detection capabilities showing how uncertainty patterns encode generator-specific signatures enabling forensic analysis beyond binary classification tasks
- novel uncertainty visualization techniques through pixel-level confidence maps that complement existing explainability approaches and provide interpretable insights

We conduct our experiments with two uncertainty methods, on two datasets, nine generators, and six detectors (including four blind and two biological detectors). We relate generator properties to detector predictions through predictive and model uncertainty using Bayesian Neural Networks (BNNs) and through model uncertainty based on model variance using Monte-Carlo (MC) dropout approach. We compare and contrast uncertainties on both traditional deepfake detection and the more elaborate deepfake source detection tasks. We measure robustness in terms of parameter sensitivity and adversarial attacks. Finally, we formulate different uncertainty maps to intersect our uncertainty analysis with explainability methods to form the big picture of detector-generator relations in deepfakes.

2. Related Work

2.1. Deepfake Generation

Deepfakes have been increasing in quality and quantity [50], mainly (1) creating entirely artificial faces from learned latent distributions [18, 23, 34], (2) replacing facial identity while attempting to preserve original conditions [3, 4, 44], or (3) modifying facial expressions and mouth movements while maintaining identity consistency [53, 65, 67]. Historically, autoregressive models [71] (AR), Variational Autoencoders [12] (VAE), Generative Adversarial Networks [28] (GAN), or diffusion models [54] are used to create such manipulated content; all of which leave behind different generative residues based on the architecture, the noise, and the operations [72, 76].

2.2. Deepfake Detection

The arms race between generation and detection intensifies as it becomes impossible to distinguish deepfakes from real faces [68]. Deepfake detectors first focused on *artifacts of fakery*, learning directly from data with "blind" detectors [8, 10, 13, 29, 30, 36, 45, 51, 64, 78, 79, 79]. Although they provide high accuracy on small datasets; they tend to overfit, they are easily manipulated by adversarial samples, and their generalization is limited across different domains, image transformations, and compression levels [15, 59].

Another branch of deepfake detection explores authenticity signals, mostly hidden in biometric data. These detectors explore low to high level signals such as blinks [46], blood flow [16], head-pose [74], emotions [32], gaze [22], and breathing [38]. These signals tend to be much inconsistent in fake videos, so the preservation of spatial, temporal, and spectral features in real videos provide an advantage for generalization over blind detectors. However, some of these inconsistencies are easily "fixed" in newer models [57].

The third and newest branch of deepfake detection aims to trace back the source generative model behind a given synthetic sample [17, 24, 25, 49, 75], following the hidden generative residue of the deep models. Some approaches even try to infer model parameters from these artifacts [11].

2.3. Uncertainty Estimation

Uncertainty estimation in machine learning involves quantifying the quality of predictions with respect to the confidence or to the model parameters. There are various approaches for uncertainty estimation including Bayesian [14, 26, 27, 73] and non-Bayesian [43, 47, 70] methods. This important step towards evaluating prediction reliability can be designed with (1) probabilistic models to cover full probability distributions over predictions (e.g., using Bayesian Neural Networks [73] (BNN)), (2) bootstrap methods to evaluate variability on controlled subsets of data or controlled subsets of the model weights (e.g., Monte-Carlo Dropout [27]), or (3) ensemble methods to combine multiple model predictions (e.g., Deep Ensembles [43]). Tangentially, uncertainty calibration also gains attention to tune these techniques for capturing the prediction distributions as close to the sample distributions. Information theoretic approaches to use entropy and mutual information for estimating uncertainty by information gain (e.g., [41]) or calibration methods to align prediction probabilities to sample frequencies (e.g., [39, 40]) are widely used for this purpose.

3. Methodology and Experimental Design

3.1. Detector Architecture Selection

Our comprehensive analysis encompasses six detector architectures representing major paradigmatic families and spanning the accuracy-efficiency trade-off space. We select these as representatives from their family of detectors to keep the number of detectors tractable (i.e., Inception [63] is in the family of Xception [19], ShuffleNet [77] is in the family of MobileNet [58], etc.).

- ResNet18 [31]: a generic lightweight blind detector
- Xception [19]: most used generic blind detector [6]
- EfficientNet [21]: one of the top scoring detectors [6]
- MobileNet [58]: a compact blind detector
- FakeCatcher [16]: an industry-adopted bio-detector [7]
- Motion-based detector [24]: a new bio-detector [24]

Deepfake detection studies only fake and real classes, where fake class equals to one source subset if it is a per-generator experiment, else covers samples of all generators. Source detection studies number of generators plus one class (for real), which is formulated as classification.

3.2. Generator Evaluation Landscape

Although there are several deepfake datasets in the literature, there exists only two multi-source datasets with known generators, namely FaceForensics++ [55] (FF) and FakeAVCeleb [35] (FAVC). FF contains 1000 real and 5000 deepfake videos, each 1000 created by FaceSwap [5], Face2Face [66], Deepfakes [3], Neural Textures [67], and FaceShifter [44], presenting a representative dataset covering various aforementioned face manipulation methods. FAVC contains unbalanced number of real and fake videos created by FaceSwapGAN [4], FSGAN [52], and Wav-to-Lip [53]. As real class has the lowest number of videos (500), we balance our setup by randomly selecting 500 videos from each class. We utilize FF as our main dataset and use FAVC for generalization, using 70/30 train/test splits for all detectors. Lastly, for the adversarial robustness experiment, we use a simple adversarial generator as outlined in [59] on all subsets of FF where the black-box attack model is selected as the ResNet18 detector.

3.3. Uncertainty Estimation

For our analysis, we employ Bayesian Neural Networks [73] to extend deterministic deep neural network architectures to corresponding Bayesian variants in order to perform stochastic variational inference. This inference captures certainty measures that help us better understand the quality of predictions. Our implementation employs mean-field Gaussian variational families (Eq. 1).

$$q(\omega) = \mathcal{N}(\omega; \mu, diag(\sigma^2)) \tag{1}$$

where training optimizes the evidence lower bound, applying KL (Kullback-Leibler divergence) loss in addition to the cross entropy loss (Eq. 2), scaling of which can be controlled by β (kl_{factor}) parameter as shown in our ablations.

$$\mathcal{L} = \mathbb{E}_{q(\omega)}[\log p(y|x,\omega)] - \beta D_{KL}[q(\omega)||p(\omega)] \tag{2}$$

We quantify predictive uncertainty (predictive entropy) capturing both aleatoric and epistemic components (Eq. 3), and model uncertainty (mutual information) computing the difference between the entropy of the mean of the predictive distribution and the mean of the entropy (Eq. 4).

$$H(y|x,D) := -\sum_{i=0}^{K-1}(p_{i\mu} \cdot log(p_{i\mu})) \tag{3}$$

$$I(y,\omega|x,D) := H(y|x,D) - E_{p(\omega|D)}[H(y|x,D)] \tag{4}$$

where $p_{i\mu}$ is the predictive mean probability of i^{th} class from n MC samples and K is the number of output classes. BNN conversion of all models is achieved using Bayesian-torch repo [42]. In order to help training convergence of models, we use MOPED method [41], which enables initializing variational parameters from a pretrained deterministic model. During inference, multiple stochastic forward passes are performed over the network via sampling from posterior distribution of the weights (with n MC samples).

As a computationally efficient alternative, we employ MC Dropout [27] during inference, treating dropout masks as approximate posterior samples. Model uncertainty is quantified through prediction variance across multiple stochastic forward passes. Performance of both methods depend on multiple parameters, set optimally by our ablation studies. In MC dropout experiments, we report model uncertainty as the mean of the variance of sampling outputs. Finally, model calibration analyses are conducted using retention plots for deepfake detection tasks.

3.4. Pixel-wise Uncertainty

One of the most prominent techniques in Explainable AI has been saliency maps [60], tracing the gradients back to input pixels to understand which pixels contribute more to the model's decision. Traditional saliency maps identify discriminative pixels but provide no information about *how certain* this contribution is. We propose uncertainty maps to visualize this information to relate the model uncertainty back to generative artifacts on images. This duality can be thought analogous to having density plots in addition to retention plots for observing the model uncertainty with respect to its accuracy. We propose two types of maps: (1) conventional saliency maps derived from Bayesian variants of regular detectors, and (2) uncertainty maps tracing the uncertainty back to pixels of original images.

3.4.1. Bayesian Saliency Maps

Saliency is computed in the traditional way by calculating a weighted average of penultimate layer activation maps, however using the BNN-converted versions of the aforementioned detectors.

$$\alpha_k = \frac{1}{n}\sum_{n} y_{\max}\left(\frac{1}{Z}\sum_{i,j}\frac{\partial y_{\max}}{\partial A^k_{ij}}\right),\; S = ReLU(\sum_{k}\alpha_k A^k) \tag{5}$$

The α_k activation weights are calculated as the pooled gradient magnitude of the k^{th} activation map A^k, scaled by the predictive confidence $y_{\max}$ of the model, and averaged over the n MC samples provided to the model, and computing the final saliency map S by a linear combination of the A^k activations with respect to α_k activation weights.

3.4.2. Uncertainty Maps

Although the previous approach pulls regular saliency maps towards uncertainty-informed saliency maps, they still do not represent pure uncertainty distribution on the input images. Thus, we formulate uncertainty maps by calculating predictive uncertainty over MC samples, and then map the gradient information from the predictive uncertainty back to input pixels. We define per-pixel uncertainty-based saliency in Eq. 6, following our notation in Eq. 3.

$$s_{ij} = \frac{\partial H(y|x,D)}{\partial x_{ij}} \tag{6}$$

This gradient-based approach reveals spatial patterns of detector confidence, identifying regions where the model exhibits high uncertainty about its predictions.

4. Analysis

We conduct experiments on uncertainty of deepfake (source) detectors, region- and pixel-based uncertainty, uncertainty estimation techniques, with ablation studies.

4.1. Uncertainty of Deepfake Detectors

M	Eval	DF	F2F	FSh	FSw	NT	All
R	acc	96.96	93.96	99.15	92.35	94.51	94.08
R$_B$	acc	96.38	94.91	98.81	93.64	93.39	95.32
	PU	0.075	0.077	0.043	0.097	0.069	0.037
	MU	0.031	0.028	0.026	0.051	0.038	0.018
E	acc	99.96	99.28	99.08	99.42	99.67	99.38
E$_B$	acc	95.87	93.24	98.82	97.75	92.31	90.93
	PU	0.209	0.151	0.203	0.168	0.246	0.263
	MU	0.098	0.092	0.107	0.095	0.132	0.128
F	acc	96.73	95.12	95.65	96.04	93.31	96.14
F$_B$	acc	96.30	94.37	95.52	95.76	91.59	95.77
	PU	0.015	0.026	0.056	0.016	0.089	0.028
	MU	0.001	0.003	0.008	0.002	0.006	0.002
M	acc	97.54	93.50	97.63	97.71	87.83	88.19
M$_B$	acc	94.16	84.91	95.91	92.95	77.71	87.40
	PU	0.083	0.147	0.071	0.103	0.241	0.182
	MU	0.007	0.007	0.008	0.007	0.006	0.007

Table 1. Accuracy and predictive (PU) and model (MU) uncertainties for (R)esnet18 [31], (E)fficientNet [21], (F)akeCatcher [16], and (M)otion-based detector [24], with their Bayesian X_B variants; per generator in FaceForensics++ [55].

Tab.1 presents binary classification performance for each synthetic content generator in the FF dataset. For instance, the DF column displays results from models trained and evaluated exclusively on DF-generated samples versus authentic content using the aforementioned split. The rightmost column combines all synthetic categories against real samples in a unified binary classification task.

Models	Eval	Results
Resnet18	accuracy	94.23
Resnet18$_B$	accuracy	93.54
	predictive uncertainty	0.119
	model uncertainty	0.054
FakeCatcher	accuracy	97.99
FakeCatcher$_B$	accuracy	98.21
	predictive uncertainty	0.030
	model uncertainty	0.009

Table 2. Accuracy and uncertainty results of regular/Bayesian variants of blind/biological detectors on FAVC.

The findings reveal a notable performance gap between traditional neural networks and their Bayesian counterparts: while complex ensemble architectures achieve high accuracy rates (99.38% in row 4), their BNN implementations experience substantial accuracy degradation, exceeding 8% (90.93% in row 5). In contrast, biological detectors maintain remarkable stability, with accuracy drops below 1%

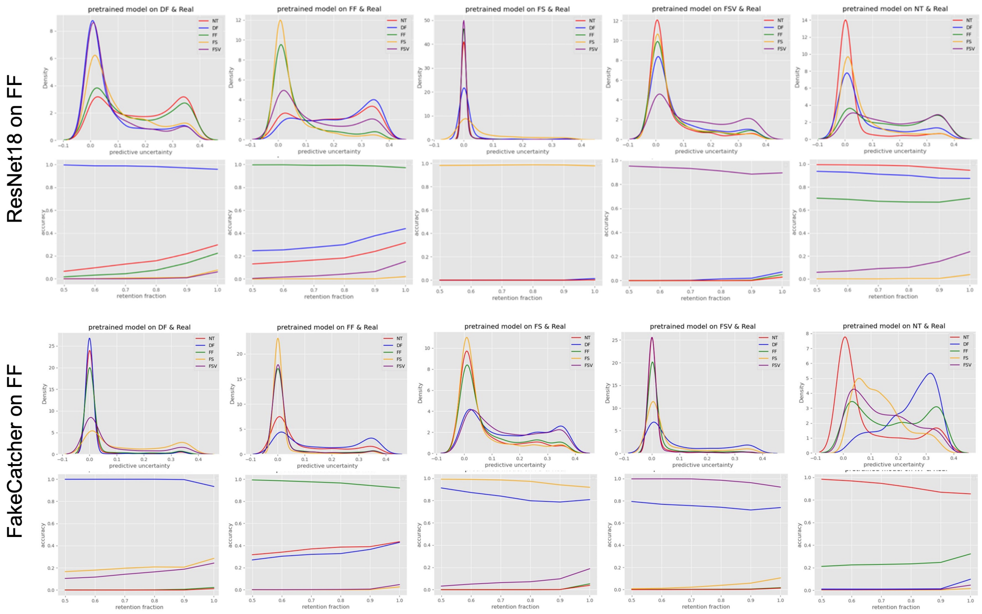


Figure 1. Each column shows density histograms and accuracy retention curves for 5 generators in FF, trained and tested per generator.

(comparing rows 7-8 and 10-11). This performance differential is further substantiated by the uncertainty metrics, where blind models exhibit considerably higher predictive and model uncertainties (0.263 PU and 0.128 MU versus 0.028 PU and 0.002 MU in rows 6 and 9, respectively). Moreover, more complex blind detectors show greater accuracy degradation and higher uncertainty with Bayesian variants. In Tab. 2, we repeat the evaluation on FAVC dataset with 3 generators and obtain similar insights, especially for validating the certainty of biological detectors.

Fig. 1 shows density histograms and corresponding retention curves, which are computed by testing ResNet18 and FakeCatcher for real/fake detection with five generators in FF. We observe that (1) biological detectors have a narrower variance of uncertainty in this binary setting, (2) similar face manipulations provide relatively better generalizability (closer curves for DF, FSh, and FSw vs. NT and F2F) and (3) for biological detectors, per-generator models can generalize to similar fakes (last row, cols. 3-4). For both detector types, we observe a unique behavior for NT, suggesting fundamental challenges in this synthesis paradigm.

To mimic a production environment, we conduct leave-one-out (LOO) experiments (Tab. 3). For LOO, we use the same split for train/validation on five classes and test the best model's accuracy on the left-out class's test set. Overall, generalizing to FSw's artifacts is harder, however FakeCatcher can achieve it. Another interpretation of Tab. 3 is that generalization capability increases as detector uses more modalities, from spatial (blind) to spatio-temporal (motion) to spectro-temporal (PPG) representations.

M	Eval	L$_{DF}$	L$_{F2F}$	L$_{FSh}$	L$_{FSw}$	L$_{NT}$
R$_B$	acc	97.75	91.25	46.75	18.25	70.92
	PU	0.074	0.179	0.246	0.193	0.252
	MU	0.036	0.139	0.142	0.090	0.151
F$_B$	acc	96.14	70.27	71.91	83.43	67.86
	PU	0.143	0.219	0.231	0.225	0.215
	MU	0.003	0.015	0.008	0.007	0.013
M$_B$	acc	93.91	64.50	82.17	52.00	69.75
	PU	0.236	0.271	0.271	0.256	0.253
	MU	0.008	0.009	0.011	0.009	0.010

Table 3. Accuracy and predictive (PU) and model (MU) uncertainties for Bayesian (R)esnet18 [31], (F)akeCatcher [16], and (M)otion-based detector [24], with leave-one-out trainings in FF.

Fig. 2 visualizes retention curves of the LOO experiment in the first column. Retention fraction represents percentage of retained data based on predictive uncertainty, with the expectation of uncertain samples decreasing accuracy for cal-

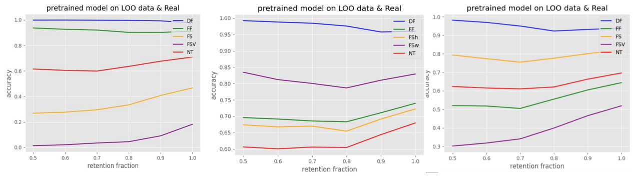


Figure 2. Retention plots of (a) ResNet18, (b) FakeCatcher, (c) motion-based detector on FF for DF, following Tab. 3.

ibrated models. Here, LOO_{DF} (blue curve) refers to the use of F2F, FSh, FSw, NT, and real data in training and validation, and DF data only for testing. Each column shows retention curves for the corresponding model in Tab. 3. Retention curves confirm our findings about generalization.

4.2. Uncertainty of Deepfake Source Detection

Tab. 4 shows uncertainty and accuracy on FF for deepfake source detection. Source detection observations are similar: Complex and large networks overfit and their Bayesian versions cannot reproduce the same accuracy, exhibit dramatic accuracy drops for source detection (EfficientNet: 99.46% → 89.77%) with extremely high uncertainty (middle blocks). Smaller networks (ResNet18, first block) and biological detectors (FakeCatcher, last block) maintain relatively stable performance and low uncertainty.

M	Eval	DF	F2F	FSh	FSw	NT	All
R	acc	98.58	96.66	98.16	95.5	92.66	96.32
R_B	acc	97.83	97.41	97.33	96.75	93.66	95.95
	PU	0.055	0.068	0.134	0.103	0.108	0.131
	MU	0.024	0.033	0.082	0.060	0.055	0.074
X	acc	99.83	99.41	98.92	99.00	99.00	99.17
X_B	acc	97.08	98.25	91.41	95.58	99.91	89.37
	PU	0.257	0.109	0.518	0.388	0.054	0.344
	MU	0.160	0.075	0.366	0.266	0.031	0.227
E	acc	99.91	99.08	98.92	99.75	99.33	99.46
E_B	acc	82.75	88.75	85.75	94.33	92.5	89.77
	PU	0.984	0.806	1.091	0.714	0.844	0.894
	MU	0.437	0.444	0.571	0.382	0.372	0.432
B	acc	99.91	99.33	99.00	99.58	99.08	99.38
B_B	acc	87.41	91.83	93.58	97.91	84.00	91.08
	PU	1.328	1.021	1.210	0.824	1.259	1.154
	MU	0.326	0.449	0.535	0.402	0.447	0.447
F	acc	92.08	90.31	92.41	90.97	85.27	91.26
F_B	acc	93.58	88.34	93.20	91.67	80.98	90.18
	PU	0.125	0.223	0.122	0.139	0.310	0.198
	MU	0.004	0.015	0.006	0.005	0.032	0.013

Table 4. Accuracy and uncertainty results of regular and Bayesian detectors (R)esnet18 [31], (X)ception [19], (E)fficientNet [21], Mo(B)ileNet [58], (F)akeCatcher [16] for source detection on real and five generators in FF.

4.3. Region-based Uncertainty Analysis

In order to couple generator types per face manipulations to detector uncertainty, we conduct region-based experiments for deepfake detection (left half) and source detection (right half) in Tab. 5. For example, when the training samples exclude bottom half of the faces, source detection for NT shows dramatic performance drop (95.81% → 91.68%), confirming mouth-centric manipulation focus. Similarly for F2F, removing symmetry elements from the training set (half mouth or one eye) reduces its source detection, as F2F is a mask-based technique creating symmetric priors. Region-based results also indicate that uncertainty measures are highly correlated with accuracy measures. It is also observed that each generator exhibits characteristic regional uncertainty patterns that can serve as forensic signatures.

5. Uncertainty Maps for Deepfake Detection

We visually compare saliency map of ResNet18 detector, its Bayesian saliency, and its uncertainty map in Fig. 3, for NT.

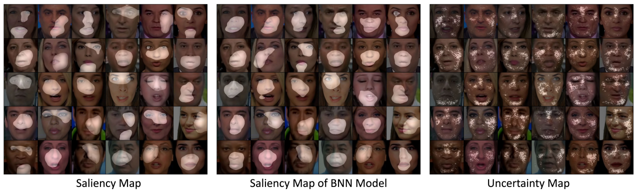


Figure 3. Saliency, Bayesian saliency, and uncertainty maps of ResNet18 detector on NT samples.

Our uncertainty maps reveal interpretable spatial patterns that complement traditional saliency analysis while providing novel insights. *From a generator perspective*, uncertainty concentrates around mouth regions and facial boundaries for NT, creating distinctive ”skull-like” patterns in cheek and forehead areas where fewer artifacts exist. For DF, uncertainty localizes primarily around facial perimeters and identity-critical regions, reflecting the encoder-decoder compression artifacts. F2F exhibits symmetric uncertainty patterns consistent with mask-based manipulation approaches. *From an interpretability angle,* uncertainty maps effectively identify regions where detection confidence decreases, often corresponding to subtle artifact boundaries. High uncertainty regions frequently correspond to areas where biological constraints are violated. While saliency maps highlight discriminative features, uncertainty maps reveal areas requiring additional evidence for confident classification. Lastly *for a comparative view*, traditional saliency focuses on obvious discriminative artifacts. Bayesian saliency produces more diffuse, averaged activations that create blob-like patterns in central facial regions. This is also expected considering that noise, mouth, and middle areas contain most of the artifacts. Uncertainty maps generate distinctive spatial patterns highlighting regions of low detector confidence, particularly in peripheral facial areas. This multi-modal visualization provides comprehensive insights into both what detectors focus on (saliency) and where they lack confidence (uncertainty), enabling more informed deployment decisions.

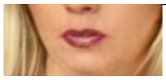
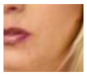
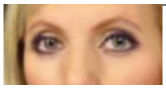
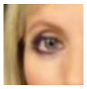
Region	M	Eval	DF	F2F	FSh	FSw	NT	All	DF	F2F	FSh	FSw	NT	All
	R	acc	99.45	98.20	99.30	99.10	97.76	97.86	96.25	95.25	97.00	90.83	88.83	93.65
	R$_B$	acc	98.37	97.69	99.17	97.03	92.17	95.00	95.25	93.33	98.00	91.33	92.75	93. 25
		PU	0.133	0.077	0.047	0.103	0.124	0.058	0.096	0.132	0.054	0.187	0.144	0.154
		MU	0.056	0.045	0.026	0.058	0.079	0.035	0.050	0.062	0.029	0.092	0.067	0.076
	R	acc	99.30	96.46	97.86	95.02	95.77	97.56	93.33	92.08	96.16	90.83	88.58	89.64
	R$_B$	acc	97.63	95.27	97.99	94.37	88.03	94.45	94.75	89.58	95.91	86.16	87.41	88.99
		PU	0.085	0.099	0.053	0.101	0.145	0.080	0.140	0.192	0.112	0.267	0.192	0.217
		MU	0.051	0.053	0.034	0.069	0.095	0.049	0.077	0.108	0.061	0.152	0.105	0.121
	R	acc	99.54	97.27	98.84	98.05	95.81	97.32	95.42	91.5	97.58	93.33	88.16	91.88
	R$_B$	acc	96.04	96.65	98.21	94.67	84.47	92.48	94.33	90.75	97.75	91.16	87.58	91.37
		PU	0.101	0.111	0.087	0.115	0.154	0.089	0.114	0.143	0.060	0.170	0.193	0.162
		MU	0.063	0.072	0.063	0.069	0.093	0.063	0.057	0.073	0.031	0.085	0.103	0.083
	R	acc	98.95	95.53	98.86	95.88	91.68	95.08	93.58	80.41	96.00	84.16	73.66	83.56
	R$_B$	acc	96.64	91.68	98.19	92.58	79.80	83.78	88.91	78.83	93.42	84.16	78.5	83.39
		PU	0.093	0.127	0.044	0.106	0.140	0.250	0.168	0.277	0.167	0.254	0.266	0.247
		MU	0.055	0.076	0.030	0.067	0.088	0.054	0.082	0.138	0.091	0.130	0.131	0.125

Table 5. Region-based analysis of accuracy and predictive (PU) and model uncertainty (MU) for deepfake detection (left half) and source detection (right half) on FF, using Resnet18 (R) and BNN Resnet18 (R$_B$).

6. Ablation Studies

Tab. 6 shows the impact of parameter values on BNN performance. Exp.1, Exp.2, Exp.3, Exp.4 and Exp.5 refer to the parameter settings of $n = \{40, 10, 40, 40, 40\}, \delta_{moped} = \{0.1, 0.1, 0.5, 0.1, 0.1\}, kl_{factor} = \{1, 1, 1, 0.5, 0.1\}$, respectively for number of MC samples n, moped delta value δ_{moped}, and scaling coefficient for KL loss kl_{factor}. The results show that increasing n from 10 to 40 marginally improves uncertainty quality without substantial accuracy changes. Smaller kl_{factor} causes degradation for NT and the quality of uncertainty measures in general deepfake detection. Finally, increasing δ_{moped} causes a significant drop for BNN performance so it should be fine-tuned.

Dataset	Eval	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5
All	acc	96.72	96.94	87.79	96.60	97.09
	PU	0.042	0.052	0.227	0.058	0.050
	MU	0.025	0.026	0.120	0.036	0.029
Neural Textures	acc	93.61	93.79	84.50	93.91	90.92
	PU	0.127	0.119	0.301	0.114	0.129
	MU	0.072	0.063	0.175	0.075	0.080

Table 6. Impact of the hyperparameters on BNN performance.

Tab. 7 shows that increasing dr (dropout ratio) causes significant drop in MC dropout performance as measured on NN and F2F. On the other hand, a fine-tuned dropout ratio may even result in an improved accuracy.

Lastly, we measure the robustness of Bayesian detectors on adversarial samples. Tab. 8 reports accuracy of the attacked BNN ResNet18 before and after adversarial generation on five generator subsets, reducing the detection accuracy by 93.53% on the average.

Data	NT			F2F		
Eval	dr=0.2	0.3	0.5	dr=0.2	0.3	0.5
acc	97.75	97.17	50.27	98.54	98.99	49.84
MU	0.015	0.026	0.030	0.013	0.011	0.023

Table 7. dr impact on Resnet18 performance with NT and F2F.

Generator	DF	F2F	FSw	NT	FSh
Baseline	95.81	95.69	89.58	93.64	97.26
After attack	0.30	0.03	3.07	0.94	0

Table 8. Adversarial robustness of BNN Resnet18 detector.

7. Implementation Details

The first four detectors consume raw data whereas the last two detectors exploit intermediate representations. FakeCatcher [16] extracts photoplethysmography (PPG) maps from videos, representative of spatial, temporal, and spectral signal behavior of heart rates. We follow their construction of PPG maps, except setting $\omega = 64$. Motion-based detector [24] extracts dual representations to represent sub-muscular motion by deep and phase-based motion magnification. We follow their construction of motion tensors with the optimum suggested parameters. Intermediate representations for three types of detectors (raw, PPG-based, and motion-based) are sampled in Fig. 4. For network counterparts, we use VGG19 [61] and C3D [69] respectively, as suggested in [17] and [24].

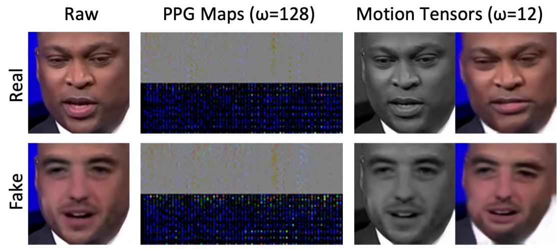


Figure 4. Data representations created by different deepfake detectors for a pair of real and fake videos from DF in FF.

For our BNN implementations using Bayesian repo [42], prior parameters are set as $\mu_{prior} = 0$, $\sigma_{prior} = 1$, $\mu i_{posterior} = 0$, and $\rho i_{posterior} = -3$. *moped* is enabled with $\delta_{moped} = 0.1$ selecting *reparameterization* type.

During training, models are trained with Adam optimizer [37] with a learning rate (LR) of 0.0001 for all architectures, except C3D. C3D LR is initiated as 0.001 and dynamic LR is applied with 0.1 scaling after each 10 epoch of overall 100 epochs. All other models are trained for 200 epochs. All weights are initiated using pretrained models on Imagenet [56] from torchvision [48], except C3D model which is pretrained on UCF101 dataset [62].

The lowest validation loss model is selected as the best model. Since accuracy is computed using n MC samples, our definition of the best model may not always correspond to the model with the highest accuracy. MC sampling enables variations at the output that may cause some noise in accuracy. Predictive and model uncertainties represent the average uncertainty measures of the test splits.

For saliency construction, batch size is set as 1 and the cut-offs are set as 20%, 20%, and 10% experimentally for saliency, Bayesian saliency, and uncertainty maps.

8. Limitations and Discussion

Our comprehensive analysis establishes several critical insights that fundamentally challenge current deepfake detection evaluation paradigms. (1) Biological detectors demonstrate markedly superior uncertainty calibration compared to blind approaches, maintaining stable performance during Bayesian conversion while exhibiting substantially lower uncertainty levels. This suggests that incorporating domain-specific physiological priors enhances not only detection accuracy but also prediction reliability, a crucial consideration for deployment. (2) The strong correlation between uncertainty measures and generalization performance across unseen generators establishes uncertainty quantification as a fundamental requirement rather than auxiliary information. Systems exhibiting high uncertainty should trigger additional verification procedures, preventing overconfident misclassifications in critical applications. (3) Uncertainty patterns encode generator-specific signatures that enable forensic analysis beyond binary classification. These findings suggest that uncertainty-aware systems could provide valuable attribution capabilities for investigating deepfake origins and understanding manipulation techniques employed. (4) Complex architectures without appropriate inductive biases exhibit poor uncertainty calibration, highlighting the importance of domain-informed design. The dramatic performance degradation suggests that model complexity alone does not guarantee reliable uncertainty estimation. (5) Our adversarial evaluation reveals concerning vulnerabilities across all detector types, with performance drops exceeding 99% under simple gradient-based attacks. The relationship between baseline uncertainty and adversarial susceptibility could open new pathways for uncertainty-aware systems to implement adaptive security measures, increasing verification for elevated uncertainty levels.

Unfortunately, Bayesian inference introduces substantial computational costs through multiple forward passes and weight sampling. Future work should explore more efficient uncertainty estimation techniques, potentially through distillation or approximation methods. While our analysis spans multiple generators, evaluation on completely novel synthesis paradigms (e.g., diffusion-based deepfakes) remains an open question requiring continued investigation.

9. Conclusion

We propose an in-depth analysis of deepfake detectors, generators, and source-detectors from an uncertainty perspective; including region-based detection experiments, novel uncertainty maps, blind and biological detector comparisons, and revelations between detector architectures and generator artifacts. Uncertainty analysis in the deepfake landscape is a new but essential dimension before releasing these detectors for public use. We have demonstrated that underconfident certain models are superior to overconfident uncertain models in terms of generalization. Our results indicate that generator artifacts can guide both detection and source detection, in image, region, and pixel levels.

As future work, we would like to build source detectors incorporating uncertainty maps, as they are shown to be information-rich for this task. As generative models and their applications become more ubiquitous and embedded in our lives, we support the responsible dissemination of tools that foster explainability, transparency, trust, and risk-awareness to ensure their use for future social good.

References

[1] An incredible series of videos swaps famous hollywood faces to demonstrate how convincing 'deepfake' tech has gotten. https://www.businessinsider.com/deepfakes-of-famous-movies-youtube-channel-2019-5, . Accessed: 2020-05-27. 1

[2] Deepfake porn nearly ruined my life. https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/, . Accessed: 2020-05-27. 1

[3] Deepfakes. https://github.com/deepfakes/faceswap, . Accessed: 2020-03-16. 2, 3

[4] Faceswap-gan. https://github.com/shaoanlu/faceswap-GAN, . Accessed: 2020-03-16. 2, 3

[5] Faceswap. https://github.com/MarekKowalski/FaceSwap, . Accessed: 2020-03-16. 2, 3

[6] Faceforensics benchmark. https://kaldir.vc.in.tum.de/faceforensics_benchmark/. Accessed: 2023-09-20. 3

[7] Nsa, u.s. federal agencies advise on deepfake threats. https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/. Accessed: 2023-09-20. 3

[8] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 2

[9] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020. 2

[10] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1205–1207, 2019. 2

[11] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15477–15493, 2023. 2

[12] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 2

[13] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro. Aligned and non-aligned double jpeg detection using convolutional neural networks. *J. Vis. Comun. Image Represent.*, 49(C):153–163, 2017. 2

[14] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015. 3

[15] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 2

[16] Umur Aybars Çiftçi, İlke Demir, and Lijun Yin. Fake-Catcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis & Machine Intelligence (PAMI)*, 2020. 1, 2, 3, 4, 5, 6, 7

[17] Umur Aybars Çiftçi, İlke Demir, and Lijun Yin. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020. 2, 8

[18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[19] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 6

[20] David Chu, İlke Demir, Kristen Eichensehr, Jacob G Foster, Mark L Green, Kristina Lerman, Filippo Menczer, Cailin O'Connor, Edward Parson, Lars Ruthotto, et al. White paper: Deep fakery – an action plan. Technical Report http://www.ipam.ucla.edu/wp-content/uploads/2020/01/Whitepaper-Deep-Fakery.pdf, Institute for Pure and Applied Mathematics (IPAM), University of California Los Angeles, Los Angeles, CA, 2020. 1

[21] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *Image Analysis and Processing – ICIAP 2022*, pages 219–229, Cham, 2022. Springer International Publishing. 2, 3, 4, 6

[22] İlke Demir and Umur Aybars Çiftçi. Where do deep fakes look? synthetic face detection via gaze tracking. In *ACM Symposium on Eye Tracking Research and Applications*, New York, NY, USA, 2021. Association for Computing Machinery. 2

[23] İlke Demir and Umur Aybars Çiftçi. Mixsyn: Compositional image synthesis with fuzzy masks and style fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7460–7469, 2024. 2

[24] İlke Demir and Umur Aybars Çiftçi. How do deepfakes move? motion magnification for deepfake source detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4780–4790, 2024. 2, 3, 4, 5, 7, 8

[25] Yuzhen Ding, Nupur Thakur, and Baoxin Li. Does a gan leave distinct model-specific fingerprints? In *BMVC*, 2021. 2

[26] Michael W. Dusenberry, Ghassen Jerfel, Yeming Wen, Yi-An Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian

neural nets with rank-1 factors. *Proceedings of the 37th International Conference on Machine Learning*, 2020. 3

[27] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 3

[28] Ian J 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. 1, 2

[29] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 2

[30] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 2

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4, 5, 6

[32] Brian Hosler, Davide Salvi, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C. Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1013–1022, 2021. 2

[33] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504, 2021. 1

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[35] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 3

[36] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. Fake face detection methods: Can they be generalized? In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2018. 2

[37] DP Kingma. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*, 2014. 8

[38] P. Korshunov and S. Marcel. Speaker inconsistency detection in tampered video. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2375–2379, 2018. 2

[39] Neslihan Kose, Ranganath Krishnan, Akash Dhamasia, Omesh Tickoo, and Michael Paulitsch. Reliable multimodal trajectory prediction via error aligned uncertainty optimization. In *Proceedings of the European Conference on Computer Vision (ECCV) workshops*, pages 443–458, 2022. 3

[40] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems*, pages 18237–18248, 2020. 3

[41] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in bayesian deep neural networks with empirical bayes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4477–4484, 2020. 3

[42] Ranganath Krishnan, Pi Esposito, and Mahesh Subedar. Bayesian-torch: Bayesian neural network layers for uncertainty estimation. `https://github.com/IntelLabs/bayesian-torch`, 2022. 3, 8

[43] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017. 3

[44] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2, 3

[45] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020. 2

[46] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 2

[47] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*, 2020. 3

[48] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. 8

[49] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511, 2019. 2

[50] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), 2021. 2

[51] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019. 2

[52] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019. 3

[53] KR Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2, 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[55] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 4

[56] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 8

[57] Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, and Otmar Hilliges. Gazenerf: 3d-aware gaze redirection with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9685, 2023. 2

[58] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3, 6

[59] Sophie R. Saremsky, Umur A. Çiftçi, Emily A. Greene, and İlke Demir. Adversarial deepfake generation for detector misclassification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2, 3

[60] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4

[61] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8

[62] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[64] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S. Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 81–87, New York, NY, USA, 2018. ACM. 2

[65] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Realtime expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), 2015. 2

[66] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 2, 3

[67] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), 2019. 2, 3

[68] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 2

[69] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 8

[70] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. *Proceedings of the 37th International Conference on Machine Learning*, 119:9690–9700, 2020. 3

[71] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[72] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 2

[73] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 3

[74] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019. 2

[75] Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[76] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019. 2

[77] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 3

[78] Y. Zhang, L. Zheng, and V. L. L. Thing. Automated face swapping and its detection. In *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, pages 15–19, 2017. 2

[79] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839, 2017. 2