Supplemental Material

A. Training of Detection Networks

The following section details the training procedures and describes any modifications made to the detection methods.

Fingerprint-based Network: The official implementation of the fingerprint-based attribution network [28] (see Figure A1) was used with the default settings, i.e., batch size 32 and the Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$, and learning rate 0.001. The network was trained using softmax cross-entropy loss between logits and labels. However, the code was modified to train with *epochs* instead of a predefined number of *ticks*. Consequently, the validation accuracy was computed after the last tick of each epoch, and only the model with the highest validation accuracy was saved. Training was halted when the validation accuracy had not improved for 30 consecutive epochs.

Xception Network: The Xception network [3] (see Figure A2) was trained following a similar procedure as described in previous studies [21, 18]. First, the final fully-connected layer was replaced to only output probabilities for 2 classes (real and synthetic) instead of 1,000 classes. Then, fine-tuning on domain-specific data was done for 3 epochs after initializing the weights in the final fully-connected layer, while keeping the ImageNet [6] weights in earlier layers frozen. Finally, the network was trained for 20 additional epochs after unfreezing all weights. Here, the model with the highest validation accuracy, computed after every epoch, was chosen for testing. Training was performed using batch size 16 and the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and learning rate 0.0002. The network was trained using cross-entropy loss.

ForensicTransfer Network: Since no official implementation was available, ForensicTransfer [4] had to be re-implemented to the best of the authors' ability. As shown in Figure A3, the spatial resolution is reduced from 256×256

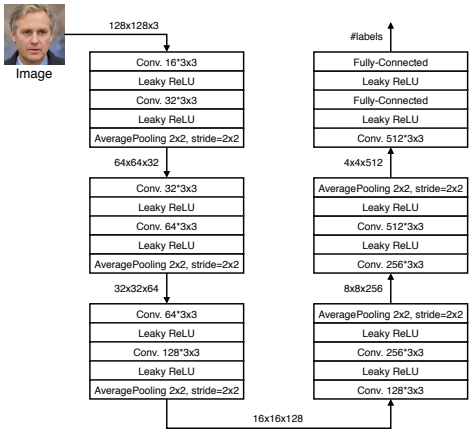


Figure A1: Fingerprint-based image attribution architecture [28].

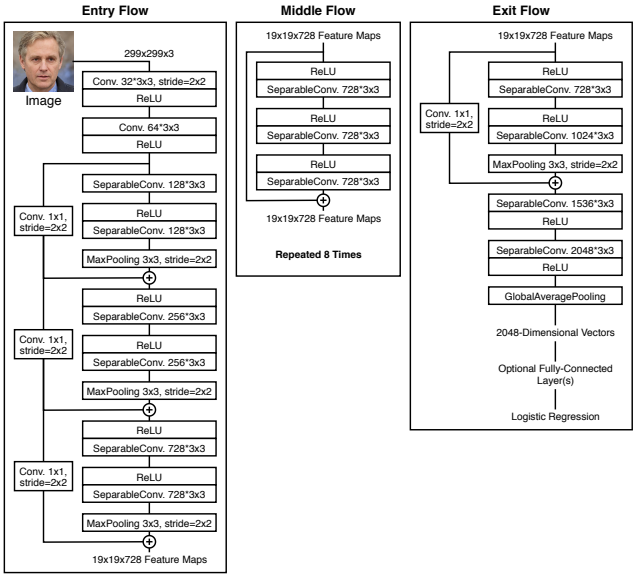


Figure A2: The Xception architecture [3]. Each convolutional and separable convolutional layer is followed by a batch normalization layer, which is not included in the figure. The architecture uses residual connections [7] around each module, except for the first and last ones.

to 240×240 before the input is fed to the encoder. Unlike the original paper, reflection padding was used to preserve the resolution when applying the high-pass filter to obtain the residual image. This resulted in a $16 \times 16 \times 128$ latent representation instead of $15 \times 15 \times 128$. As described in Section 3.3, this latent representation $\mathbf{h} = \{\mathbf{h}_0, \mathbf{h}_1\}$ is split into the disjoint parts $\mathbf{h}_0$ and $\mathbf{h}_1$ before being fed to the selection layer. The per-sample loss equals the weighted sum of the activation loss $\mathcal{L}_{ACT}$ and reconstruction loss $\mathcal{L}_{REC}$:

$$\mathcal{L} = \mathcal{L}_{ACT} + \gamma \mathcal{L}_{REC}, \tag{1}$$

where $\gamma = 0.1$. The activation loss is given by

$$\mathcal{L}_{ACT} = |a_1 - l| + |a_0 - (1 - l)|\,, \tag{2}$$

where

$$a_k = \frac{1}{N_k} \|\mathbf{h}_k\|_1\,. \tag{3}$$

Here, $l \in \{0, 1\}$ is the ground truth class label and N_k is the number of elements in $\mathbf{h}_k$ of class $k \in \{0, 1\}$. The reconstruction loss is given by

$$\mathcal{L}_{REC} = \frac{1}{N} \|\mathbf{x} - \mathcal{D}(\mathbf{h})\|_1\,, \tag{4}$$

where N is the number of elements in the encoder input $\mathbf{x}$, and $\mathcal{D}(\cdot)$ is the decoder. At test time, predictions are made based on the activations in the 128 feature maps. A sample

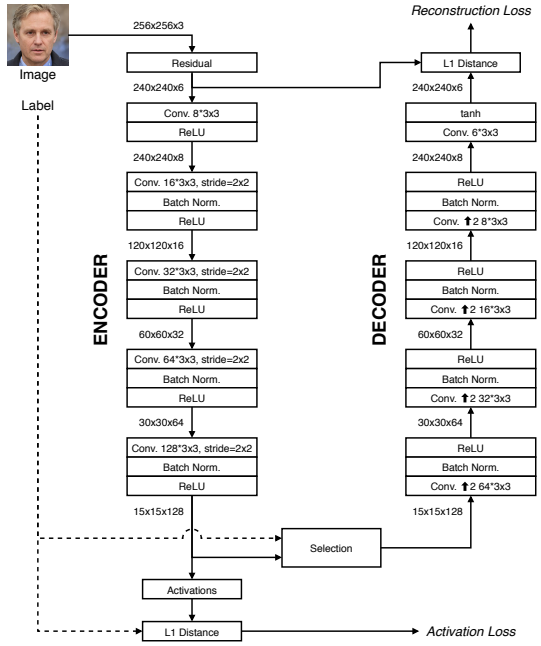


Figure A3: The autoencoder-based ForensicTransfer architecture with a custom weighted loss function [4]. In the decoder, 2×2 nearest-neighbor interpolation is used to recover the original spatial resolution of the input.

is classified as synthetic ($k = 1$) if $a_1 > a_0$, and real otherwise. Some performance metrics require class probabilities in addition to class labels. Therefore, ForensicTransfer was extended to also output class probabilities at test time, where the real and synthetic probabilities were given by $\frac{a_0}{(a_0+a_1)}$ and $\frac{a_1}{(a_0+a_1)}$, respectively.

Following the original paper, the network was trained using batch size 64 and the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$, and learning rate 0.001. Training was halted when the validation loss had not decreased for 30 consecutive epochs. Hence, the model with the lowest validation loss was chosen for testing.

Fine-tuning on Target Data: Before fine-tuning on out-of-distribution images as described in Section 4.3, the default learning rate of each detection network was halved, while the number of epochs specified for the stopping criterion was doubled. The latter meant that the fingerprint-based network and the ForensicTransfer network were trained until the validation accuracy/loss had not improved for 60 consecutive epochs, while the Xception network was trained for 40 epochs in total.

B. Individual Dataset Results and Complementary Evaluation Metrics

Some of the experiments presented in the main paper were conducted using many different dataset combinations. Consequently, the values of the evaluation metrics were sometimes averaged across all test sets to provide an overview of the general performance of each detection method. In this section, the performance on each individual test set is presented, upon which average values in the main paper were calculated. All results in the main paper are also complemented with additional evaluation metrics, namely precision, recall, F1-score, and average precision.

Perturbation Robustness on Individual Test Sets: Tables B1–B12 show the performance on each individual test set used in the robustness experiments described under *Random Perturbations* in Section 4.2. In those experiments, detection models were trained on real and synthetic datasets, and tested on held-out samples that had been augmented by randomly sampling values for the parameters controlling noise, blur, compression, and resizing. Here, Tables B1–B6 show the performance of models trained on unperturbed samples, while Tables B7–B12 show the performance of models trained on randomly augmented samples.

Strictly Controlled Perturbations: Figure B1 shows the results of the robustness experiment described under *Strictly Controlled Perturbations* in Section 4.2. In that experiment, detection models were trained on FFHQ and StyleGAN2, and tested on held-out samples augmented using gradually increasing perturbation intensities. Here, the performance is presented with respect to multiple evaluation metrics in addition to accuracy and AUROC.

Generalizability on Individual Test Sets: Tables B13–B14 show the performance on each individual test set considered in the first generalization experiment described under *No Fine-tuning* in Section 4.3. In that experiment, detection models were tested on out-of-distribution samples from unseen datasets, without any fine-tuning. Here, the performance is presented with respect to multiple evaluation metrics in addition to accuracy and AUROC. Table B13 shows the performance of models trained on unperturbed samples, while Table B14 shows the performance of models trained on randomly augmented samples.

Generalizability After Fine-tuning: Figures B2–B3 show the results of the generalization experiments described under *Target Fine-tuning* and *Source and Target Fine-tuning* in Section 4.3. In those experiments, detection models originally trained on FFHQ / ProGAN (source) were tested on CelebA-HQ / StyleGAN2 (target) after being fine-tuned on samples from the target distribution. Here, the performance is presented with respect to multiple evaluation metrics in addition to accuracy and AUROC. Figure B2 shows the performance of models fine-tuned on target samples, while Figure B3 shows the performance of models fine-tuned on both source and target samples.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.979	**1.000**	**1.000**	0.984	**1.000**	**1.000**	0.981	**1.000**	**1.000**	0.981	**1.000**	**1.000**	0.998	**1.000**	**1.000**	0.998	**1.000**	**1.000**
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.985	**1.000**	**1.000**	0.988	**1.000**	**1.000**	0.986	**1.000**	**1.000**	0.986	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.999	**1.000**	**1.000**
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.995	**1.000**	**1.000**	0.995	**1.000**	**1.000**	0.995	**1.000**	**1.000**	0.995	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
FFHQ	ProGAN	FFHQ	ProGAN	0.986	**1.000**	0.999	0.987	**1.000**	**1.000**	0.986	**1.000**	**1.000**	0.986	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.999	**1.000**	**1.000**
FFHQ	StyleGAN	FFHQ	StyleGAN	0.987	**1.000**	0.995	0.974	**1.000**	**1.000**	0.981	**1.000**	0.997	0.980	**1.000**	0.997	0.998	**1.000**	**1.000**	0.998	**1.000**	**1.000**
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.991	**1.000**	**1.000**	0.991	**1.000**	**1.000**	0.991	**1.000**	**1.000**	0.991	**1.000**	**1.000**	0.999	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
			Average	0.987	**1.000**	0.999	0.987	**1.000**	**1.000**	0.987	**1.000**	**1.000**	0.987	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.999	**1.000**	**1.000**

Table B1: Performance of detection models trained on unperturbed datasets and tested on held-out unperturbed samples. The fingerprint-based network, Xception network, and ForensicTransfer network are abbreviated as F.P., X.C., and F.T., respectively.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.978	**0.992**	0.500	0.981	0.027	**1.000**	**0.980**	0.514	0.500	**0.980**	0.053	0.667	**0.997**	0.930	0.545	**0.997**	0.926	0.535
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	**0.985**	0.951	0.500	0.987	0.997	**1.000**	**0.986**	0.973	0.500	**0.986**	0.973	0.667	**0.999**	**0.999**	0.442	**0.999**	0.998	0.453
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.995	**0.996**	0.500	0.994	0.787	**1.000**	**0.995**	0.892	0.500	**0.995**	0.879	0.667	**1.000**	0.995	0.596	**1.000**	0.995	0.604
FFHQ	ProGAN	FFHQ	ProGAN	0.985	**0.994**	0.000	**0.986**	0.018	0.000	**0.986**	0.509	0.500	**0.986**	0.035	0.000	**0.998**	0.971	0.514	**0.998**	0.971	0.506
FFHQ	StyleGAN	FFHQ	StyleGAN	0.986	**0.989**	0.508	0.973	0.984	**0.985**	0.979	**0.986**	0.516	0.979	**0.986**	0.670	0.998	**0.999**	0.597	0.998	**0.999**	0.603
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.991	**1.000**	0.000	**0.990**	0.110	0.000	**0.991**	0.555	0.500	**0.991**	0.199	0.000	**0.999**	0.981	0.481	**0.999**	0.983	0.488
			Average	**0.987**	**0.987**	0.335	**0.985**	0.487	0.664	**0.986**	0.738	0.503	**0.986**	0.521	0.445	**0.999**	0.979	0.529	**0.999**	0.979	0.532

Table B2: Performance of detection models trained on unperturbed datasets and tested on held-out samples augmented with Gaussian noise.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.979	**0.983**	0.534	0.983	**1.000**	**1.000**	0.981	**0.991**	0.564	0.981	**0.991**	0.697	0.998	**1.000**	0.925	0.998	**1.000**	0.910
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.984	**1.000**	0.988	**0.987**	0.615	0.175	**0.986**	0.808	0.586	**0.986**	0.762	0.297	**0.999**	**0.999**	0.894	**0.999**	**0.999**	0.896
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.995	**1.000**	0.998	**0.995**	0.965	0.163	**0.995**	0.982	0.582	**0.995**	0.982	0.281	**1.000**	**1.000**	0.912	**1.000**	**1.000**	0.902
FFHQ	ProGAN	FFHQ	ProGAN	**0.985**	0.954	0.505	0.986	**1.000**	**1.000**	**0.985**	0.976	0.510	**0.985**	0.976	0.671	**0.999**	**0.999**	0.882	**0.999**	**0.999**	0.852
FFHQ	StyleGAN	FFHQ	StyleGAN	0.985	**1.000**	**1.000**	**0.973**	0.797	0.005	**0.979**	0.899	0.502	**0.979**	0.887	0.010	0.998	**1.000**	0.937	0.998	**1.000**	0.946
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.991	**1.000**	**1.000**	**0.991**	0.932	0.091	**0.991**	0.966	0.546	**0.991**	0.965	0.167	0.999	**1.000**	0.811	**1.000**	**1.000**	0.833
			Average	0.987	**0.990**	0.838	**0.986**	0.885	0.406	**0.986**	0.937	0.548	**0.986**	0.927	0.354	0.999	**1.000**	0.894	0.999	**1.000**	0.890

Table B3: Performance of detection models trained on unperturbed datasets and tested on held-out samples augmented with Gaussian blur.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.977	**1.000**	0.000	**0.956**	0.144	0.000	**0.966**	0.572	0.500	**0.966**	0.252	0.000	**0.994**	0.868	0.488	**0.994**	0.900	0.500
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.984	0.999	**1.000**	**0.986**	0.215	0.000	**0.985**	0.607	0.500	**0.985**	0.353	0.000	**0.999**	0.941	0.507	**0.999**	0.949	0.524
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.996	**1.000**	0.000	**0.991**	0.104	0.000	**0.993**	0.552	0.500	**0.993**	0.188	0.000	**1.000**	0.824	0.419	**1.000**	0.879	0.454
FFHQ	ProGAN	FFHQ	ProGAN	**0.984**	0.908	0.933	0.982	**0.995**	0.017	**0.983**	0.947	0.508	**0.983**	0.949	0.033	**0.998**	0.984	0.859	**0.998**	0.976	0.861
FFHQ	StyleGAN	FFHQ	StyleGAN	0.986	**1.000**	0.000	**0.973**	0.946	0.000	**0.980**	0.973	0.500	**0.980**	0.972	0.000	0.998	**1.000**	0.490	0.998	**1.000**	0.495
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.992	**1.000**	0.000	**0.986**	0.893	0.000	**0.989**	0.946	0.500	**0.989**	0.943	0.000	**0.999**	0.996	0.387	**0.999**	0.997	0.446
			Average	**0.987**	0.985	0.322	**0.979**	0.550	0.003	**0.983**	0.766	0.501	**0.983**	0.610	0.006	**0.998**	0.936	0.525	**0.998**	0.950	0.547

Table B4: Performance of detection models trained on unperturbed datasets and tested on held-out samples augmented with JPEG compression.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.979	0.999	**1.000**	0.984	**0.998**	0.782	0.982	**0.998**	0.891	0.982	**0.998**	0.878	0.998	**1.000**	0.998	0.998	**1.000**	0.998
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.985	**1.000**	**1.000**	**0.988**	0.946	0.834	**0.986**	0.973	0.917	**0.986**	0.972	0.909	0.999	**1.000**	0.999	0.999	**1.000**	**1.000**
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.995	**1.000**	**1.000**	**0.995**	0.968	0.794	**0.995**	0.984	0.897	**0.995**	0.984	0.885	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
FFHQ	ProGAN	FFHQ	ProGAN	**0.986**	0.950	0.930	0.986	**1.000**	0.955	**0.986**	0.973	0.942	**0.986**	0.974	0.942	**0.999**	**0.999**	0.983	**0.999**	**0.999**	0.984
FFHQ	StyleGAN	FFHQ	StyleGAN	0.986	0.999	**1.000**	0.974	**0.984**	0.363	0.980	**0.991**	0.681	0.980	**0.991**	0.533	0.998	**1.000**	0.888	0.998	**1.000**	0.917
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.991	**1.000**	**1.000**	**0.991**	0.971	0.634	**0.991**	0.985	0.817	**0.991**	0.985	0.776	0.999	**1.000**	0.984	**1.000**	**1.000**	0.987
			Average	0.987	**0.991**	0.988	**0.986**	0.978	0.727	**0.987**	0.984	0.858	**0.987**	0.984	0.821	0.999	**1.000**	0.975	0.999	**1.000**	0.981

Table B5: Performance of detection models trained on unperturbed datasets and tested on held-out samples augmented with resizing ($75 \times 75 \leq$ resolution $\leq 299 \times 299$, before cropping).

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	**0.977**	0.950	0.555	**0.959**	0.292	0.635	**0.968**	0.638	0.562	**0.968**	0.446	0.592	**0.995**	0.794	0.597	**0.995**	0.816	0.609
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	**0.984**	0.976	0.588	**0.985**	0.499	0.455	**0.984**	0.744	0.568	**0.984**	0.661	0.513	**0.999**	0.845	0.575	**0.999**	0.877	0.560
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.995	**0.996**	0.596	**0.990**	0.444	0.430	**0.992**	0.721	0.569	**0.992**	0.614	0.499	**1.000**	0.839	0.606	**1.000**	0.892	0.649
FFHQ	ProGAN	FFHQ	ProGAN	**0.982**	0.839	0.592	**0.983**	0.590	0.433	**0.982**	0.738	0.568	**0.982**	0.693	0.500	**0.998**	0.794	0.592	**0.998**	0.793	0.626
FFHQ	StyleGAN	FFHQ	StyleGAN	0.987	**0.990**	0.603	**0.972**	0.749	0.331	**0.980**	0.871	0.557	**0.980**	0.853	0.428	**0.997**	0.984	0.581	**0.998**	0.985	0.626
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.991	0.999	**1.000**	**0.985**	0.431	0.122	**0.988**	0.716	0.561	**0.988**	0.603	0.218	**0.999**	0.856	0.536	**0.999**	0.896	0.631
			Average	**0.986**	0.958	0.656	**0.979**	0.501	0.401	**0.982**	0.738	0.564	**0.982**	0.645	0.458	**0.998**	0.852	0.581	**0.998**	0.877	0.617

Table B6: Performance of detection models trained on unperturbed datasets and tested on held-out samples randomly augmented with combinations of Gaussian noise, Gaussian blur, JPEG compression, and resizing.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.977	**0.999**	0.975	0.981	**1.000**	0.974	0.979	**1.000**	0.975	0.979	**1.000**	0.975	0.997	**1.000**	0.997	0.996	**1.000**	0.997
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.986	**1.000**	0.966	0.989	**1.000**	0.996	0.987	**1.000**	0.980	0.987	**1.000**	0.981	0.999	**1.000**	0.999	0.999	**1.000**	0.999
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.992	**1.000**	0.981	0.993	**1.000**	0.997	0.992	**1.000**	0.989	0.993	**1.000**	0.989	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
FFHQ	ProGAN	FFHQ	ProGAN	0.983	**1.000**	0.985	0.981	**1.000**	0.974	0.982	**1.000**	0.979	0.982	**1.000**	0.979	0.998	**1.000**	0.998	0.998	**1.000**	0.998
FFHQ	StyleGAN	FFHQ	StyleGAN	0.982	**0.998**	0.969	0.981	**0.999**	0.921	0.981	**0.999**	0.946	0.981	**0.999**	0.945	0.998	**1.000**	0.986	0.998	**1.000**	0.989
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.987	**1.000**	0.963	0.992	**1.000**	0.952	0.989	**1.000**	0.957	0.989	**1.000**	0.957	0.999	**1.000**	0.992	0.999	**1.000**	0.993
			Average	0.985	**1.000**	0.973	0.986	**1.000**	0.969	0.985	**1.000**	0.971	0.985	**1.000**	0.971	0.999	**1.000**	0.995	0.998	**1.000**	0.996

Table B7: Performance of detection models trained on randomly augmented datasets and tested on held-out unperturbed samples that had not been augmented at all.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.976	**0.999**	0.834	0.980	**0.999**	0.660	0.978	**0.999**	0.764	0.978	**0.999**	0.737	0.997	**1.000**	0.855	0.996	**1.000**	0.859
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.986	**0.999**	0.930	0.988	**0.999**	0.988	0.987	**0.999**	0.957	0.987	**0.999**	0.958	0.999	**1.000**	0.994	0.999	**1.000**	0.994
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.991	**1.000**	0.977	0.993	**1.000**	0.978	0.992	**1.000**	0.977	0.992	**1.000**	0.977	**1.000**	**1.000**	0.997	**1.000**	**1.000**	0.997
FFHQ	ProGAN	FFHQ	ProGAN	0.983	**1.000**	0.959	0.981	**0.999**	0.903	0.982	**1.000**	0.932	0.982	**1.000**	0.930	0.998	**1.000**	0.978	0.998	**1.000**	0.982
FFHQ	StyleGAN	FFHQ	StyleGAN	0.982	**0.999**	0.904	0.980	**0.999**	0.877	0.981	**0.999**	0.892	0.981	**0.999**	0.891	0.998	**1.000**	0.956	0.998	**1.000**	0.962
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.987	**0.999**	0.898	0.991	**1.000**	0.923	0.989	**1.000**	0.909	0.989	**1.000**	0.910	0.999	**1.000**	0.970	0.999	**1.000**	0.971
			Average	0.984	**0.999**	0.917	0.986	**0.999**	0.888	0.985	**1.000**	0.905	0.985	**1.000**	0.901	0.999	**1.000**	0.958	0.998	**1.000**	0.961

Table B8: Performance of detection models trained on randomly augmented datasets and tested on held-out samples that had only been augmented with Gaussian noise.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.976	**0.999**	0.919	0.981	**1.000**	0.987	0.978	**1.000**	0.950	0.978	**1.000**	0.952	0.997	**1.000**	0.994	0.996	**1.000**	0.995
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.985	**0.999**	0.991	0.988	**0.999**	0.970	0.987	**0.999**	0.981	0.987	**0.999**	0.981	0.999	**1.000**	0.999	0.999	**1.000**	0.998
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.991	**0.999**	0.988	0.994	**1.000**	0.991	0.992	**1.000**	0.989	0.992	**1.000**	0.989	**1.000**	**1.000**	0.999	**1.000**	**1.000**	0.999
FFHQ	ProGAN	FFHQ	ProGAN	0.982	**1.000**	0.870	0.981	**1.000**	0.985	0.982	**1.000**	0.919	0.982	**1.000**	0.924	0.998	**1.000**	0.985	0.998	**1.000**	0.983
FFHQ	StyleGAN	FFHQ	StyleGAN	0.982	**0.997**	0.934	0.981	**0.999**	0.927	0.981	**0.998**	0.931	0.981	**0.998**	0.930	0.998	**1.000**	0.980	0.998	**1.000**	0.982
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.985	**1.000**	0.924	0.992	**1.000**	0.958	0.989	**1.000**	0.940	0.989	**1.000**	0.941	0.999	**1.000**	0.987	0.999	**1.000**	0.987
			Average	0.984	**0.999**	0.938	0.986	**1.000**	0.970	0.985	**1.000**	0.952	0.985	**1.000**	0.953	0.999	**1.000**	0.991	0.998	**1.000**	0.991

Table B9: Performance of detection models trained on randomly augmented datasets and tested on held-out samples that had only been augmented with Gaussian blur.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.972	**0.999**	0.905	0.974	**0.999**	0.425	0.973	**0.999**	0.690	0.973	**0.999**	0.578	0.996	**1.000**	0.863	0.995	**1.000**	0.856
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.986	**0.999**	0.941	0.987	**0.999**	0.921	0.987	**0.999**	0.931	0.987	**0.999**	0.931	0.999	**1.000**	0.981	0.999	**1.000**	0.978
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.993	**1.000**	0.969	0.993	**1.000**	0.942	0.993	**1.000**	0.956	0.993	**1.000**	0.955	**1.000**	**1.000**	0.993	**1.000**	**1.000**	0.993
FFHQ	ProGAN	FFHQ	ProGAN	0.980	**1.000**	0.923	0.979	**0.999**	0.946	0.980	**0.999**	0.933	0.980	**0.999**	0.934	0.997	**1.000**	0.984	0.997	**1.000**	0.985
FFHQ	StyleGAN	FFHQ	StyleGAN	0.982	**0.999**	0.894	0.981	**0.999**	0.874	0.981	**0.999**	0.885	0.981	**0.999**	0.884	0.998	**1.000**	0.954	0.998	**1.000**	0.959
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.987	**0.999**	0.923	0.990	**1.000**	0.899	0.989	**0.999**	0.912	0.989	**0.999**	0.911	0.999	**1.000**	0.972	0.999	**1.000**	0.973
			Average	0.983	**0.999**	0.926	0.984	**0.999**	0.835	0.984	**0.999**	0.885	0.984	**0.999**	0.866	0.998	**1.000**	0.958	0.998	**1.000**	0.957

Table B10: Performance of detection models trained on randomly augmented datasets and tested on held-out samples that had only been augmented with JPEG compression.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.976	**0.999**	0.912	0.981	**1.000**	0.929	0.979	**1.000**	0.920	0.979	**1.000**	0.920	0.997	**1.000**	0.975	0.996	**1.000**	0.977
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.986	**1.000**	0.970	0.989	**0.999**	0.975	0.988	**0.999**	0.973	0.988	**0.999**	0.973	0.999	**1.000**	0.996	0.999	**1.000**	0.996
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.992	**1.000**	0.962	0.993	**1.000**	0.984	0.993	**1.000**	0.973	0.993	**1.000**	0.973	**1.000**	**1.000**	0.997	**1.000**	**1.000**	0.998
FFHQ	ProGAN	FFHQ	ProGAN	0.984	**1.000**	0.880	0.981	**0.999**	0.930	0.982	**1.000**	0.902	0.982	**1.000**	0.904	0.998	**1.000**	0.966	0.998	**1.000**	0.964
FFHQ	StyleGAN	FFHQ	StyleGAN	0.982	**0.998**	0.939	0.981	**0.998**	0.901	0.982	**0.998**	0.921	0.981	**0.998**	0.920	0.998	**1.000**	0.975	0.998	**1.000**	0.979
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.986	**0.999**	0.928	0.991	**1.000**	0.939	0.989	**1.000**	0.933	0.989	**1.000**	0.933	0.999	**1.000**	0.983	0.999	**1.000**	0.984
			Average	0.984	**0.999**	0.932	0.986	**0.999**	0.943	0.986	**1.000**	0.937	0.985	**1.000**	0.937	0.999	**1.000**	0.982	0.998	**1.000**	0.983

Table B11: Performance of detection models trained on randomly augmented datasets and tested on held-out samples that had only been augmented with resizing ($75 \times 75 \leq$ resolution $\leq 299 \times 299$, before cropping).

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	CelebA-HQ	ProGAN	0.971	**0.995**	0.817	0.976	**0.997**	0.575	0.973	**0.996**	0.723	0.973	**0.996**	0.675	0.996	**1.000**	0.821	0.995	**1.000**	0.835
CelebA-HQ	StyleGAN	CelebA-HQ	StyleGAN	0.986	**0.999**	0.918	0.987	**0.998**	0.875	0.986	**0.999**	0.899	0.986	**0.999**	0.896	0.999	**1.000**	0.964	0.999	**1.000**	0.967
CelebA-HQ	StyleGAN2	CelebA-HQ	StyleGAN2	0.992	**0.999**	0.947	0.992	**0.999**	0.931	0.992	**0.999**	0.940	0.992	**0.999**	0.939	**1.000**	**1.000**	0.987	**1.000**	**1.000**	0.987
FFHQ	ProGAN	FFHQ	ProGAN	0.980	**0.996**	0.815	0.979	**0.998**	0.855	0.979	**0.997**	0.830	0.979	**0.997**	0.834	0.998	**1.000**	0.913	0.998	**1.000**	0.913
FFHQ	StyleGAN	FFHQ	StyleGAN	0.982	**0.996**	0.870	0.979	**0.998**	0.890	0.980	**0.997**	0.879	0.980	**0.997**	0.880	0.998	**1.000**	0.950	0.998	**1.000**	0.955
FFHQ	StyleGAN2	FFHQ	StyleGAN2	0.986	**0.999**	0.886	0.990	**0.999**	0.921	0.988	**0.999**	0.901	0.988	**0.999**	0.903	0.999	**1.000**	0.968	0.999	**1.000**	0.969
			Average	0.983	**0.997**	0.876	0.984	**0.998**	0.841	0.983	**0.998**	0.862	0.983	**0.998**	0.855	0.998	**1.000**	0.934	0.998	**1.000**	0.938

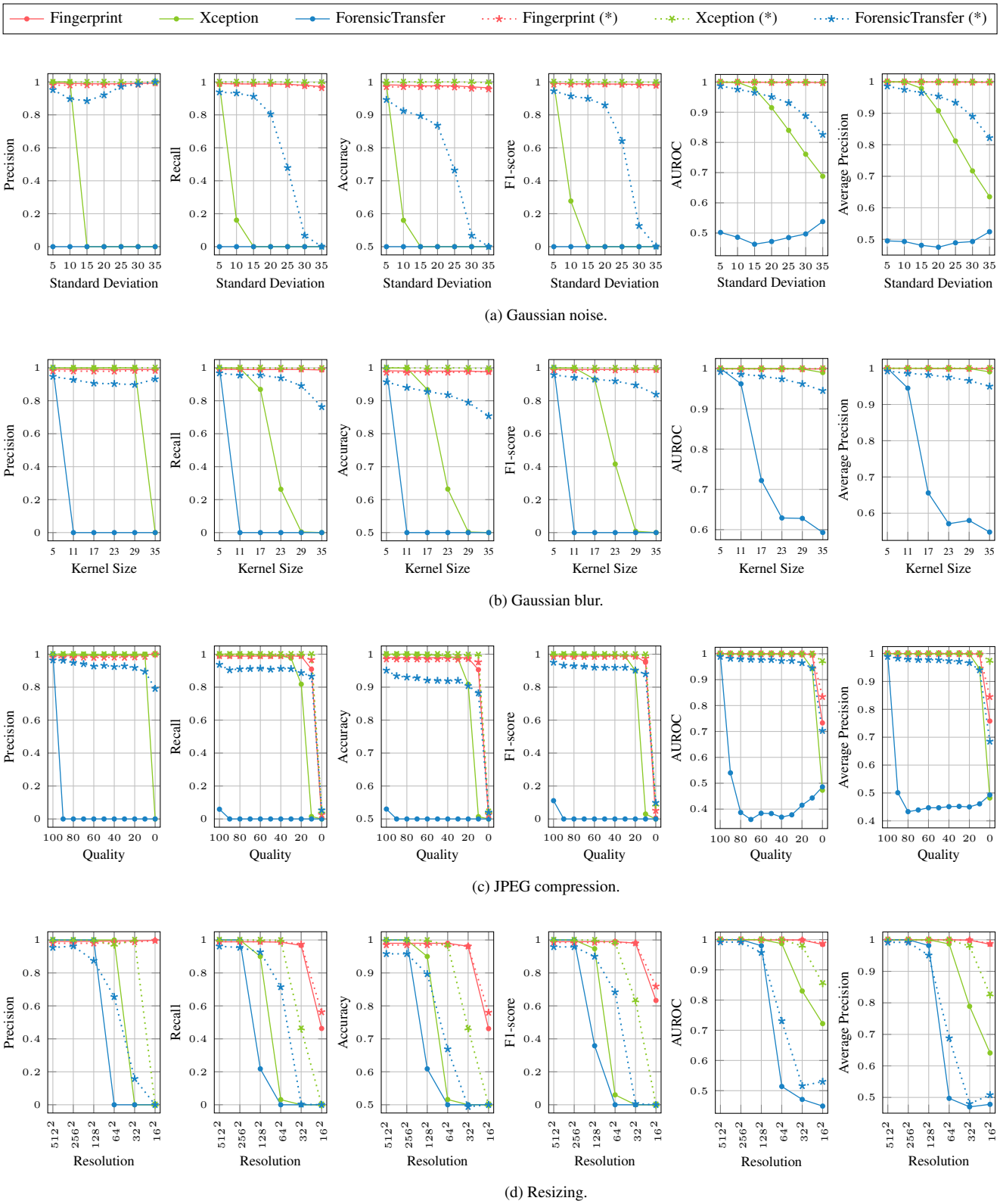
Table B12: Performance of detection models trained on randomly augmented datasets and tested on held-out samples randomly augmented with combinations of Gaussian noise, Gaussian blur, JPEG compression, and resizing.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	FFHQ	StyleGAN	0.192	**0.722**	0.539	0.006	0.031	**1.000**	0.491	0.510	**0.573**	0.011	0.060	**0.701**	0.394	**0.797**	0.666	0.418	**0.737**	0.561
CelebA-HQ	ProGAN	FFHQ	StyleGAN2	0.018	**0.740**	0.539	0.000	0.034	**1.000**	0.488	0.511	**0.573**	0.001	0.065	**0.701**	0.265	**0.821**	0.818	0.361	**0.762**	0.711
CelebA-HQ	StyleGAN	FFHQ	ProGAN	0.078	0.008	**0.098**	0.015	0.003	**0.090**	**0.417**	0.318	0.134	0.025	0.004	**0.094**	**0.205**	0.099	0.108	**0.344**	0.321	0.326
CelebA-HQ	StyleGAN	FFHQ	StyleGAN2	**0.829**	0.626	0.549	0.876	0.613	**1.000**	**0.848**	0.623	0.589	**0.852**	0.619	0.709	0.923	0.672	**0.933**	**0.906**	0.596	0.885
CelebA-HQ	StyleGAN2	FFHQ	ProGAN	0.053	0.002	**0.062**	0.005	0.001	**0.031**	**0.461**	0.224	0.281	0.008	0.001	**0.041**	0.235	0.053	**0.259**	0.352	0.314	**0.360**
CelebA-HQ	StyleGAN2	FFHQ	StyleGAN	**0.891**	0.618	0.676	0.673	0.897	**0.978**	**0.796**	0.672	0.755	0.767	0.732	**0.800**	0.892	0.737	**0.927**	0.897	0.667	**0.921**
FFHQ	ProGAN	CelebA-HQ	StyleGAN	**0.015**	0.000	0.000	**0.004**	0.000	0.000	**0.370**	0.244	0.365	**0.006**	0.000	0.000	**0.080**	0.048	0.004	**0.316**	0.314	0.312
FFHQ	ProGAN	CelebA-HQ	StyleGAN2	**0.001**	0.000	0.000	**0.000**	**0.000**	**0.000**	**0.368**	0.244	0.365	**0.001**	0.000	0.000	**0.043**	0.026	**0.043**	0.311	0.310	**0.313**
FFHQ	StyleGAN	CelebA-HQ	ProGAN	0.601	**0.892**	0.000	0.018	**0.033**	0.000	0.503	**0.515**	0.500	0.035	**0.064**	0.000	0.539	0.831	**0.842**	0.535	0.797	**0.853**
FFHQ	StyleGAN	CelebA-HQ	StyleGAN2	0.984	0.294	**1.000**	0.714	0.002	**0.988**	0.851	0.499	**0.994**	0.828	0.003	**0.994**	0.976	0.526	**1.000**	0.977	0.500	**1.000**
FFHQ	StyleGAN2	CelebA-HQ	ProGAN	0.688	**1.000**	0.000	**0.026**	0.002	0.000	**0.507**	0.501	0.500	**0.049**	0.004	0.000	0.579	0.685	**0.982**	0.576	0.682	**0.983**
FFHQ	StyleGAN2	CelebA-HQ	StyleGAN	0.973	**1.000**	**1.000**	0.422	0.003	**0.471**	0.705	0.501	**0.736**	0.588	0.005	**0.640**	0.902	0.412	**0.985**	0.913	0.493	**0.989**
			Average	0.444	**0.492**	0.372	0.230	0.135	**0.463**	**0.567**	0.447	0.530	0.264	0.130	**0.390**	0.503	0.476	**0.631**	0.576	0.541	**0.685**

Table B13: Performance of detection models trained on unperturbed datasets and tested on out-of-distribution samples from unseen unperturbed datasets.

Training		Test		Precision			Recall			Accuracy			F1-score			AUROC			Average Precision		
Real	Synthetic	Real	Synthetic	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.	F.P.	X.C.	F.T.
CelebA-HQ	ProGAN	FFHQ	StyleGAN	0.148	**0.882**	0.575	0.006	0.002	**0.794**	0.486	0.501	**0.604**	0.011	0.003	**0.667**	0.349	**0.759**	0.609	0.397	**0.707**	0.546
CelebA-HQ	ProGAN	FFHQ	StyleGAN2	0.007	0.000	**0.583**	0.000	0.000	**0.821**	0.484	0.500	**0.617**	0.000	0.000	**0.682**	0.216	0.236	**0.619**	0.346	0.352	**0.541**
CelebA-HQ	StyleGAN	FFHQ	ProGAN	**0.081**	0.025	0.045	0.016	0.001	**0.036**	0.419	**0.487**	0.141	0.026	0.001	**0.040**	0.208	**0.283**	0.042	0.345	**0.367**	0.311
CelebA-HQ	StyleGAN	FFHQ	StyleGAN2	0.834	**0.872**	0.568	0.897	0.180	**0.991**	**0.860**	0.577	0.619	**0.865**	0.299	0.722	**0.935**	0.762	0.821	**0.929**	0.756	0.776
CelebA-HQ	StyleGAN2	FFHQ	ProGAN	**0.079**	0.008	0.031	0.010	0.000	**0.016**	0.447	**0.480**	0.255	0.018	0.001	**0.021**	**0.256**	0.160	0.087	**0.359**	0.332	0.318
CelebA-HQ	StyleGAN2	FFHQ	StyleGAN	0.862	**0.932**	0.645	0.723	0.545	**0.919**	**0.804**	0.752	0.706	**0.787**	0.688	0.758	0.884	**0.900**	0.836	0.889	**0.904**	0.825
FFHQ	ProGAN	CelebA-HQ	StyleGAN	**0.017**	0.000	0.002	**0.006**	0.000	0.002	0.328	**0.457**	0.098	**0.009**	0.000	0.002	**0.082**	0.066	0.013	**0.317**	0.314	0.308
FFHQ	ProGAN	CelebA-HQ	StyleGAN2	**0.001**	0.000	**0.001**	0.000	0.000	**0.001**	0.325	**0.457**	0.098	0.000	0.000	**0.001**	**0.055**	0.014	0.019	**0.313**	0.308	0.309
FFHQ	StyleGAN	CelebA-HQ	ProGAN	0.579	**0.918**	0.278	**0.028**	0.020	0.020	0.504	**0.509**	0.484	**0.054**	0.039	0.038	0.549	**0.742**	0.393	0.542	**0.723**	0.423
FFHQ	StyleGAN	CelebA-HQ	StyleGAN2	**0.973**	0.963	0.916	**0.739**	0.046	0.579	**0.859**	0.522	0.763	**0.840**	0.088	0.710	**0.971**	0.831	0.886	**0.971**	0.822	0.891
FFHQ	StyleGAN2	CelebA-HQ	ProGAN	0.695	**0.800**	0.482	0.056	0.003	**0.077**	**0.516**	0.501	0.497	0.104	0.006	**0.133**	0.580	**0.614**	0.481	0.583	**0.614**	0.488
FFHQ	StyleGAN2	CelebA-HQ	StyleGAN	0.955	**0.996**	0.853	**0.519**	0.194	0.482	**0.747**	0.597	0.699	**0.672**	0.325	0.616	0.897	**0.927**	0.814	0.909	**0.939**	0.819
			Average	0.436	**0.533**	0.415	0.250	0.083	**0.395**	**0.565**	0.528	0.465	0.282	0.121	**0.366**	0.499	**0.525**	0.468	0.575	**0.595**	0.546

Table B14: Performance of detection models trained on randomly augmented datasets and tested on out-of-distribution samples from unseen unperturbed datasets.



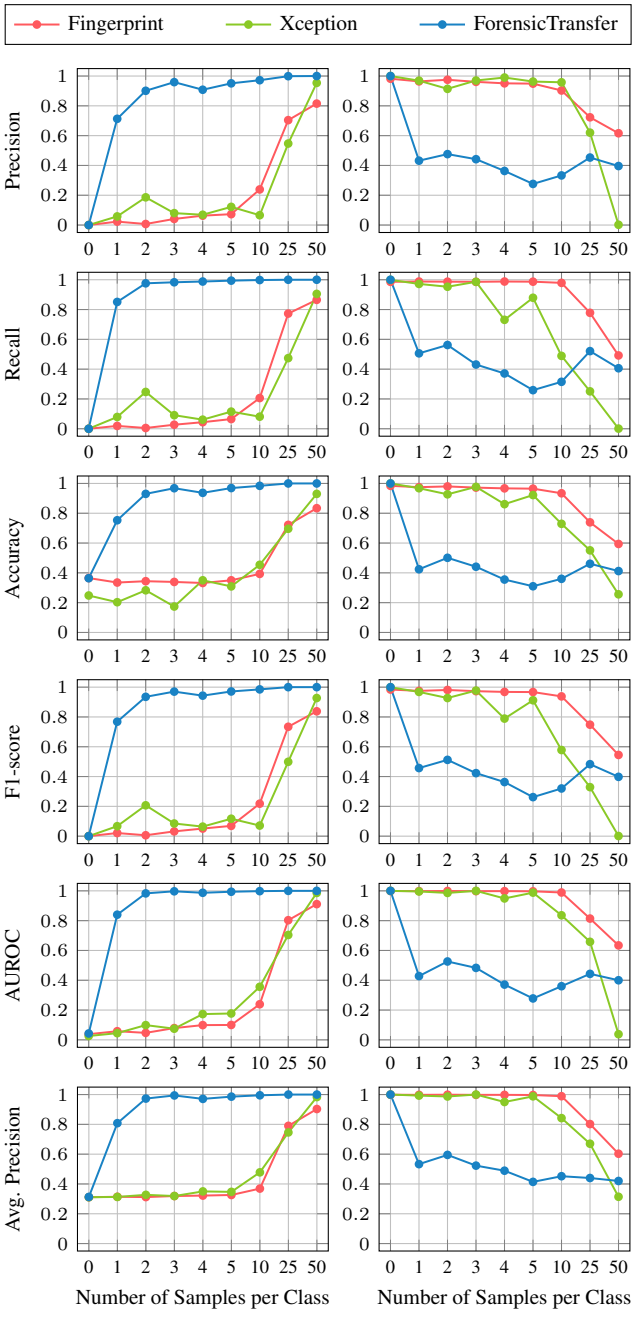
(a) Gaussian noise.

(b) Gaussian blur.

(c) JPEG compression.

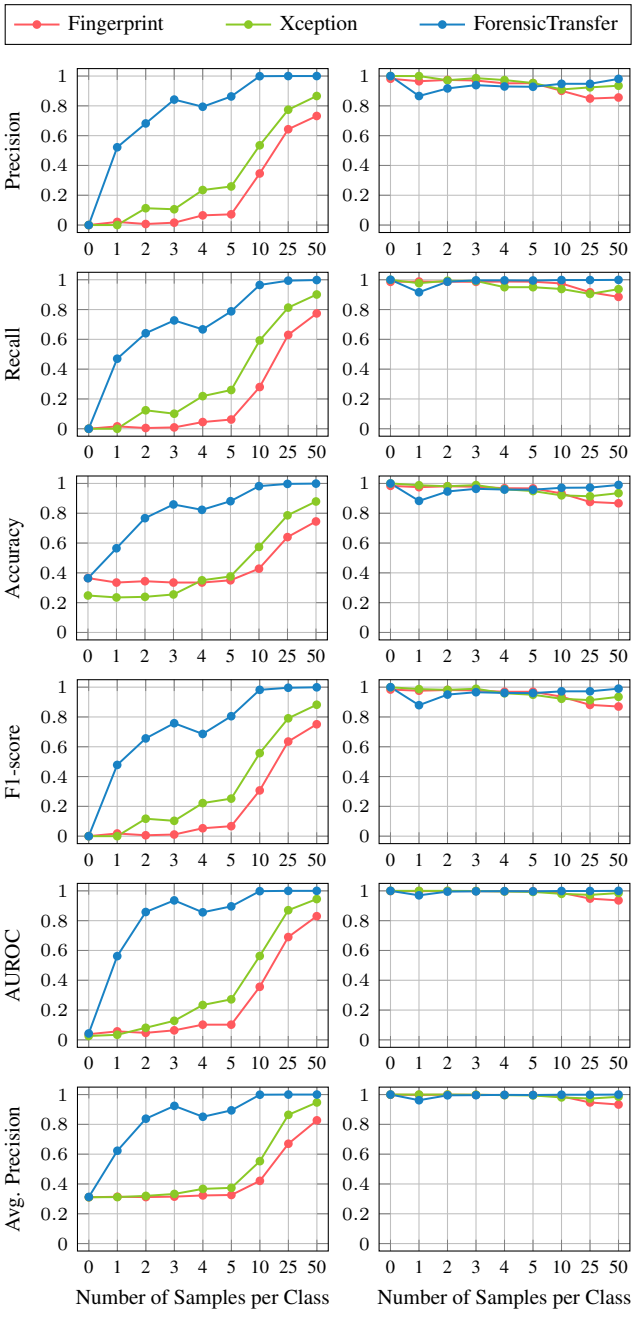
(d) Resizing.

Figure B1: Performance of models trained on FFHQ / StyleGAN2 and tested on held-out samples augmented in various ways. Training was performed separately on unperturbed and randomly augmented (*) images. The scale of the axis is not uniform for *Resolution*.



(a) Target. (b) Source.

Figure B2: Source and target performance of models trained on FFHQ / ProGAN (source) and fine-tuned on CelebA-HQ / StyleGAN2 (target), averaged over 10 runs. The metrics are plotted with respect to the number of target training samples per class used for fine-tuning. All detection models were trained and tested on unperturbed images. Note that the scale of the axis is not uniform for *Number of Samples per Class*.



(a) Target. (b) Source.

Figure B3: Source and target performance of models trained on FFHQ / ProGAN (source) and fine-tuned on both CelebA-HQ / StyleGAN2 (target) and FFHQ / ProGAN (source), averaged over 10 runs. The metrics are plotted with respect to the number of target training samples per class used for fine-tuning. All models were trained and tested on unperturbed images. Note that the scale of the axis is not uniform for *Number of Samples per Class*.