Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

Ning Yu[1,2] Larry Davis[1] Mario Fritz[3]
[1]University of Maryland, College Park
[2]Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
[3]CISPA Helmholtz Center for Information Security
Saarland Informatics Campus, Germany
ningyu@mpi-inf.mpg.de lsd@cs.umd.edu fritz@cispa.saarland

Abstract

Recent advances in Generative Adversarial Networks (GANs) have shown increasing success in generating photorealistic images. But they also raise challenges to visual forensics and model attribution. We present the first study of learning GAN fingerprints towards image attribution and using them to classify an image as real or GAN-generated. For GAN-generated images, we further identify their sources. Our experiments show that (1) GANs carry distinct model fingerprints and leave stable fingerprints in their generated images, which support image attribution; (2) even minor differences in GAN training can result in different fingerprints, which enables fine-grained model authentication; (3) fingerprints persist across different image frequencies and patches and are not biased by GAN artifacts; (4) fingerprint finetuning is effective in immunizing against five types of adversarial image perturbations; and (5) comparisons also show our learned fingerprints consistently outperform several baselines in a variety of setups [1].

1. Introduction

In the last two decades, photorealistic image generation and manipulation techniques have rapidly evolved. Visual contents can now be easily created and edited without leaving obvious perceptual traces [72]. Recent breakthroughs in generative adversarial networks (GANs) [31, 52, 10, 32, 38, 19] have further improved the quality and photorealism of generated images. The adversarial framework of GANs can also be used in conditional scenarios for image translation [36, 70, 71] or manipulation in a given context [60, 61, 57, 12, 64], which diversifies media synthesis.

[1]Code, data, models, and supplementary material are available at GitHub.

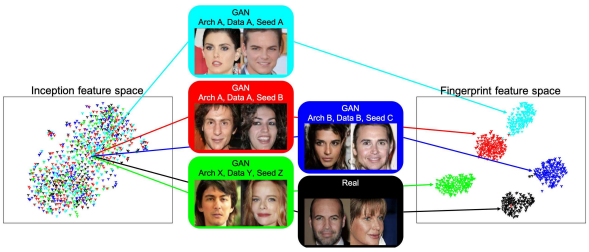


Figure 1. A t-SNE [43] visual comparison between our fingerprint features (right) and the baseline inception features [52] (left) for image attribution. Inception features are highly entangled, indicating the challenge to differentiate high-quality GAN-generated images from real ones. However, our result shows any single difference in GAN architectures, training sets, or even initialization seeds can result in distinct fingerprint features for effective attribution.

At the same time, however, the success of GANs has raised two challenges to the vision community: visual forensics and intellectual property protection.

GAN challenges to visual forensics. There is a widespread concern about the impact of this technology when used maliciously. This issue has also received increasing public attention, in terms of disruptive consequences to visual security, laws, politics, and society in general [6, 1, 3]. Therefore, it is critical to look into effective visual forensics against threats from GANs.

While recent state-of-the-art visual forensics techniques demonstrate impressive results for detecting fake visual media [16, 53, 27, 13, 22, 11, 35, 67, 68, 26], they have only focused on semantic, physical, or statistical inconsistency of specific forgery scenarios, e.g., copy-move manipulations[16, 26] or face swapping [67]. Forensics on GAN-generated images [44, 47, 59] shows good accuracy, but each method operates on only one GAN architecture by identifying its unique artifacts and results deteriorate when the GAN architecture is changed. It is still an open question of whether GANs leave stable marks that are commonly

shared by their generated images. That motivates us to investigate an effective feature representation that differentiates GAN-generated images from real ones.

GAN challenges to intellectual property protection. Similar to other successful applications of deep learning technology to image recognition [33] or natural language processing [30], building a product based on GANs is non-trivial [37, 4, 5]. It requires a large amount of training data, powerful computing resources, significant machine learning expertise, and numerous trial-and-error iterations for identifying optimal model architectures and their model hyperparameters. As GAN services become widely deployed with commercial potential, they will become increasingly vulnerable to pirates. Such copyright plagiarism may jeopardize the intellectual property of model owners and take future market share from them. Therefore, methods for attributing GAN-generated image origins are highly desirable for protecting intellectual property.

Given the level of realism that GAN techniques already achieve today, attribution by human inspection is no longer feasible (see the mixture of Figure 4). The state-of-the-art digital identification techniques can be separated into two categories: digital watermarking and digital fingerprint detection. Neither of them is applicable to GAN attribution. Previous work on watermarking deep neural networks [65, 62] depends on an embedded security scheme during "white-box" model training, requires control of the input, and is impractical when only GAN-generated images are accessible in a "black-box" scenario. Previous work on digital fingerprints is limited to device fingerprints [42, 21] or in-camera post-processing fingerprints [24], which cannot be easily adapted to GAN-generated images. That motivates us to investigate GAN fingerprints that attribute different GAN-generated images to their sources.

We present the first study addressing the two GAN challenges simultaneously by learning GAN fingerprints for image attribution: We introduce GAN fingerprints and use them to classify an image as real or GAN-generated. For GAN-generated images, we further identify their sources. We approach this by training a neural network classifier and predicting the source of an image. Our experiments show that GANs carry distinct model fingerprints and leave stable fingerprints in their generated images, which support image attribution.

We summarize our **contributions** as demonstrating the existence, uniqueness, persistence, immunizability, and visualization of GAN fingerprints. We address the following questions:

Existence and uniqueness: Which GAN parameters differentiate image attribution? We present experiments on GAN parameters including architecture, training data, as well as random initialization seed. We find that a difference in any one of these parameters results in a unique GAN fingerprint for image attribution. See Figure 1, Section 3.1 and 4.2.

Persistence: Which image components contain fingerprints for attribution? We investigate image components in different frequency bands and in different patch sizes. In order to eliminate possible bias from GAN artifact components, we apply a perceptual similarity metric to distill an artifact-free subset for attribution evaluation. We find that GAN fingerprints are persistent across different frequencies and patch sizes, and are not dominated by artifacts. See Section 3.2 and 4.3.

Immunizability: How robust is attribution to image perturbation attacks and how effective are the defenses? We investigate common attacks that aim at destroying image fingerprints. They include noise, blur, cropping, JPEG compression, relighting, and random combinations of them. We also defend against such attacks by finetuning our attribution classifier. See Section 4.4.

Visualization: How to expose GAN fingerprints? We propose an alternative classifier variant to explicitly visualize GAN fingerprints in the image domain, so as to better interpret the effectiveness of attribution. See Section 3.3 and 4.5.

Comparison to baselines. In terms of attribution accuracy, our method consistently outperforms three baseline methods (including a very recent one [45]) on two datasets under a variety of experimental conditions. In terms of feature representation, our fingerprints show superior distinguishability across image sources compared to inception features [52].

2. Related work

Generative Adversarial Networks (GANs). GANs [31, 52, 10, 32, 38, 19] have shown improved photorealism in image synthesis [40, 15, 69], translation [36, 70, 71], or manipulation [9, 60, 61]. We focus on unconditional GANs as the subject of our study. We choose the following four GAN models as representative candidates of the current state of the art: ProGAN [38], SNGAN [46], CramerGAN [14], and MMDGAN [17], considering their outstanding performances on face generation.

Visual forensics. Visual forensics targets detecting statistical or physics-based artifacts and then recognizing the authenticity of visual media without evidence from an embedded security mechanism [28, 27]. An example is a steganalysis-based method [29], which uses hand-crafted features plus a linear Support Vector Machine to detect forgeries. Recent CNN-based methods [13, 22, 18, 11, 35, 67, 68, 7, 23, 26] learn deep features and further improve tampering detection performance on images or videos. Rössler *et al*. [49, 50] introduced a large-scale face manipulation dataset to benchmark forensics classification and segmentation tasks, and demonstrated superior performance when using additional domain-specific knowledge. For forensics on

GAN-generated images, several existing works [44, 47, 59] show good accuracy. However, each method considers only one GAN architecture and results do not generalize across architectures.

Digital fingerprints. Prior digital fingerprint techniques focus on detecting hand-crafted features for either device fingerprints or postprocessing fingerprints. The device fingerprints rely on the fact that individual devices, due to manufacturing imperfections, leave a unique and stable mark on each acquired image, i.e., the photo-response non-uniformity (PRNU) pattern [42, 21]. Likewise, postprocessing fingerprints come from the specific in-camera post-processing suite (demosaicking, compression, etc.) during each image acquisition procedure [24]. Recently, Marra *et al*. [45] visualize GAN fingerprints based on PRNU, and show their application to GAN source identification. We replace their hand-crafted fingerprint formulation with a learning-based one, decoupling model fingerprint from image fingerprint, and show superior performances in a variety of experimental conditions.

Digital watermarking. Digital watermarking is a complementary forensics technique for image authentication [58, 39, 51]. It involves embedding artificial watermarks in images. It can be used to reveal image source and ownership so as to protect their copyright. It has been shown that neural networks can also be actively watermarked during training [65, 62]. In such models, a characteristic pattern can be built into the learned representation but with a trade-off between watermarking accuracy and the original performance. However, such watermarking has not been studied for GANs. In contrast, we utilize inherent fingerprints for image attribution without any extra embedding burden or quality deterioration.

3. Fingerprint learning for image attribution

Inspired by the prior works on digital fingerprints [42, 24], we introduce the concepts of GAN model fingerprint and image fingerprint. Both are simultaneously learned from an image attribution task.

Model fingerprint. Each GAN model is characterized by many parameters: training dataset distribution, network architecture, loss design, optimization strategy, and hyper-parameter settings. Because of the non-convexity of the objective function and the instability of adversarial equilibrium between the generator and discriminator in GANs, the values of model weights are sensitive to their random initializations and do not converge to the same values during each training. This indicates that even though two well-trained GAN models may perform equivalently, they generate high-quality images differently. This suggests the existence and uniqueness of GAN fingerprints. We define the model fingerprint per GAN instance as a reference vector, such that it consistently interacts with all its generated images. In a specifically designed case, the model fingerprint can be an RGB image the same size as its generated images. See Section 3.3.

Image fingerprint. GAN-generated images are the outcomes of a large number of fixed filtering and non-linear processes, which generate common and stable patterns within the same GAN instances but are distinct across different GAN instances. That suggests the existence of image fingerprints and attributability towards their GAN sources. We introduce the fingerprint per image as a feature vector encoded from that image. In a specifically designed case, an image fingerprint can be an RGB image the same size as the original image. See Section 3.3.

3.1. Attribution network

Similar to the authorship attribution task in natural language processing [56, 8], we train an attribution classifier that can predict the source of an image: real or from a GAN model.

We approach this using a deep convolutional neural network supervised by image-source pairs $\{(I, y)\}$ where $I \sim \mathbb{I}$ is sampled from an image set and $y \in \mathbb{Y}$ is the source ground truth belonging to a finite set. That set is composed of pre-trained GAN instances plus the real world. Figure 2(a) depicts an overview of our attribution network.

We implicitly represent image fingerprints as the final classifier features (the $1 \times 1 \times 512$ tensor before the final fully connected layer) and represent GAN model fingerprints as the corresponding classifier parameters (the $1 \times 1 \times 512$ weight tensor of the final fully connected layer).

Why is it necessary to use such an external classifier when GAN training already provides a discriminator? The discriminator learns a hyperplane in its own embedding space to distinguish generated images from real ones. Different embedding spaces are not aligned. In contrast, the proposed classifier necessarily learns a unified embedding space to distinguish generated images from different GAN instances or from real images.

Note that our motivation to investigate "white-box" GANs subject to known parameters is to validate the attributability along different GAN parameter dimensions. In practice, our method also applies to "black-box" GAN API services. The only required supervision is the source label of an image. We can simply query different services, collect their generated images, and label them by service indices. Our classifier would test image authenticity by predicting if an image is sampled from the desired service. We also test service authenticity by checking if most of their generated images have the desired source prediction.

3.2. Component analysis networks

In order to analyze which image components contain fingerprints, we propose three variants of the network.

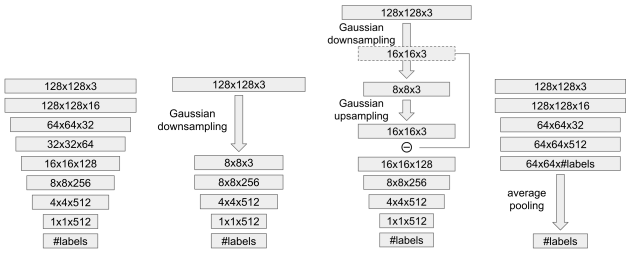


Figure 2. Different attribution network architectures. Tensor representation is specified by two spatial dimensions followed by the number of channels. The network is trained to minimize cross-entropy classification loss. (a) Attribution network. (b) Pre-downsampling network example that downsamples input image to 8×8 before convolution. (c) Pre-downsampling residual network example that extracts the residual component between 16×16 and 8×8 resolutions. (d) Post-pooling network example that starts average pooling at 64×64 resolution.

Pre-downsampling network. We propose to test whether fingerprints and attribution can be derived from different frequency bands. We investigate attribution performance w.r.t. downsampling factor. Figure 2(b) shows an architecture example that extracts low-frequency bands. We replace the trainable convolution layers with our Gaussian downsampling layers from the input end and systematically control at which resolution we stop such replacement.

Pre-downsampling residual network. Complementary to extracting low-frequency bands, Figure 2(c) shows an architecture example that extracts a residual high-frequency band between one resolution and its factor-2 downsampled resolution. It is reminiscent of a Laplacian Pyramid [20]. We systematically vary the resolution at which we extract such residual.

Post-pooling network. We propose to test whether fingerprints and attribution can be derived locally based on patch statistics. We investigate attribution performance w.r.t. patch size. Figure 2(d) shows an architecture example. Inspired by PatchGAN [36], we regard a "pixel" in a neural tensor as the feature representation of a local image patch covered by the receptive field of that "pixel". Therefore, post-pooling operations count for patch-based neural statistics. Earlier post-pooling corresponds to a smaller patch size. We systematically vary at which tensor resolution we start this pooling in order to switch between more local and more global patch statistics.

3.3. Fingerprint visualization

Alternatively to our attribution network in Section 3.1 where fingerprints are implicitly represented in the feature domain, we describe a model similar in spirit to Marra *et al.* [45] to explicitly represent them in the image domain. But in contrast to their hand-crafted PRNU-based representation, we modify our attribution network architecture and

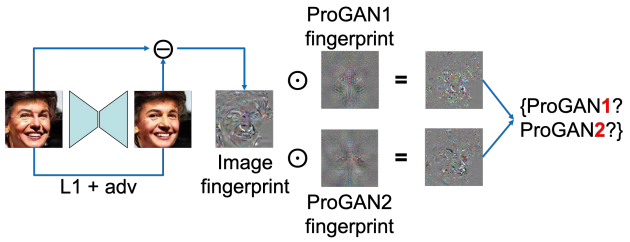


Figure 3. Fingerprint visualization diagram. We train an AutoEncoder and GAN fingerprints end-to-end. $\odot$ indicates pixel-wise multiplication of two normalized images.

learn fingerprint images from image-source pairs $(\{I, y\})$. We also decouple the representation of model fingerprints from image fingerprints. Figure 3 depicts the fingerprint visualization model.

Abstractly, we learn to map from input image to its fingerprint image. But without fingerprint supervision, we choose to ground the mapping based on a reconstruction task with an AutoEncoder. We then define the reconstruction residual as the image fingerprint. We simultaneously learn a model fingerprint for each source (each GAN instance plus the real world), such that the correlation index between one image fingerprint and each model fingerprint serves as softmax logit for classification.

Mathematically, given an image-source pair (I, y) where $y \in \mathbb{Y}$ belongs to the finite set $\mathbb{Y}$ of GAN instances plus the real world, we formulate a reconstruction mapping R from I to $R(I)$. We ground our reconstruction based on pixel-wise L_1 loss plus adversarial loss:

$$L_{pix}(I) = ||R(I) - I||_1 \tag{1}$$

$$L_{adv}(I) = D_{rec}\big(R(I)\big) - D_{rec}\big(I\big) + GP\big(R(I), I | D_{rec}\big) \tag{2}$$

where D_{rec} is an adversarially trained discriminator, and $GP(\cdot)$ is the gradient penalty regularization term defined in [32].

We then explicitly define image fingerprint F^I_{im} as the reconstruction residual $F^I_{im} = R(I) - I$.

We further explicitly define model fingerprint F^y_{mod} as freely trainable parameters with the same size as F^I_{im}, such that $corr(F^I_{im}, F^y_{mod})$, the correlation index between F^I_{im} and F^y_{mod}, is maximized over $\mathbb{Y}$. This can be formulated as the softmax logit for the cross-entropy classification loss supervised by the source ground truth:

$$L_{cls}(I, y) = -\log \frac{corr(F^I_{im}, F^y_{mod})}{\sum_{\hat{y} \in \mathbb{Y}} corr(F^I_{im}, F^{\hat{y}}_{mod})} \tag{3}$$

where $corr(A, B) = \hat{A} \odot \hat{B}$, $\hat{A}$ and $\hat{B}$ are the zero-mean, unit-norm, and vectorized version of images A and B, and $\odot$ is the inner product operation.

Our final training objective is

$$\min_{R, \{F^{\tilde{y}}_{mod} | \tilde{y} \in \mathbb{Y}\}} \max_{D_{rec}} \mathop{\mathbb{E}}_{\{(I,y)\}} (\lambda_1 L_{pix} + \lambda_2 L_{adv} + \lambda_3 L_{cls}) \tag{4}$$

(a) CelebA real data (b) ProGAN (c) SNGAN (d) CramerGAN (e) MMDGAN

Figure 4. Face samples from difference sources.

where $\lambda_1 = 20.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 1.0$ are used to balance the order of magnitude of each loss term, which are not sensitive to dataset and are fixed.

Note that this network variant is used to better visualize and interpret the effectiveness of image attribution. However, it introduces extra training complexity and thus is not used if we only focus on attribution.

4. Experiments

We discuss the experimental setup in Section 4.1. From Section 4.2 to 4.5, we explore the four research questions discussed in the Introduction.

4.1. Setup

Datasets . We employ CelebA human face dataset [41] and LSUN bedroom scene dataset [63], both containing $20,000$ real-world RGB images.

GAN models. We consider four recent state-of-the-art GAN architectures: ProGAN [38], SNGAN [46], CramerGAN [14], and MMDGAN [17]. Each model is trained from scratch with their default settings except we fix the number of training epochs to 240 and fix the output size of a generator to $128 \times 128 \times 3$.

Baseline methods. Given real-world datasets and four pre-trained GAN models, we compare with three baseline classification methods: k-nearest-neighbor (kNN) on raw pixels, Eigenface [55], and the very recent PRNU-based fingerprint method from Marra *et al*. [45].

Evaluation. We use classification accuracy to evaluate image attribution performance.

In addition, we use the ratio of inter-class and intra-class Fréchet Distance [25], denoted as FD ratio, to evaluate the distinguishability of a feature representation across classes. The larger the ratio, the more distinguishable the feature representation across sources. See supplementary material for more detail. We compare our fingerprint features to image inception features [52]. The FD of inception features is also known as FID for GAN evaluation [34]. Therefore, the FD ratio of inception features can serve as a reference to show how challenging it is to attribute high-quality GAN-generated images manually or without fingerprint learning.

4.2. Existence and uniqueness: which GAN parameters differentiate image attribution?

We consider GAN architecture, training set, and initialization seed respectively by varying one type of parameter and keeping the other two fixed.

Different architectures. First, we leverage all the real images to train ProGAN, SNGAN, CramerGAN, and MMDGAN separately. For the classification task, we configure training and testing sets with 5 classes: {*real*, *ProGAN*, *SNGAN*, *CramerGAN*, *MMDGAN*}. We randomly collect $100,000$ images from each source for classification training and another $10,000$ images from each source for testing. We show face samples from each source in Figure 4 and bedroom samples in the supplementary material. Table 1 shows that we can effectively differentiate GAN-generated images from real ones and attribute generated images to their sources, just using a regular CNN classifier. There do exist unique fingerprints in images that differentiate GAN architectures, even though it is far more challenging to attribute those images manually or through inception features [52].

Different GAN training sets. We further narrow down the investigation to GAN training sets. From now we only focus on ProGAN plus real dataset. We first randomly select a base real subset containing $100,000$ images, denoted as *real_subset_diff_0*. We then randomly select 10 other real subsets also containing $100,000$ images, denoted as *real_subset_diff_#i*, where $i \in \{1, 10, 100, 1000, 10000, 20000, 40000, 60000, 80000, 100000\}$ indicates the number of images that are not from the base subset. We collect such sets of datasets to explore the relationship between attribution performance and GAN training set overlaps.

For each *real_subset_diff_#i*, we separately train a ProGAN model and query $100,000$ images for classifier training and another $10,000$ images for testing, labeled as *ProGAN_subset_diff_#i*. In this setup of {*real*, *ProGAN_subset_diff_#i*}, we show the performance evaluation in Table 2. Surprisingly, we find that attribution performance remains equally high regardless of the amount of GAN training set overlap. Even GAN training sets that differ in just one image can lead to distinct GAN instances. That indicates that one-image mismatch during GAN training results in a different optimization step in one iteration

Table 1. Evaluation on {*real*, *ProGAN*, *SNGAN*, *CramerGAN*, *MMDGAN*}. The best performance is highlighted in **bold**.

		CelebA	LSUN
Accuracy (%)	kNN	28.00	36.30
	Eigenface [55]	53.28	-
	PRNU [45]	86.61	67.84
	Ours	**99.43**	**98.58**
FD ratio	Inception [52]	2.36	5.27
	Our fingerprint	**454.76**	**226.59**

Table 2. Evaluation on {*real*, *ProGAN_subset_diff_#i*}. The best performance is highlighted in **bold**.

		CelebA	LSUN
Accuracy (%)	kNN	11.46	10.72
	Eigenface [55]	27.98	-
	PRNU [45]	92.28	70.55
	Ours	**99.50**	**97.66**
FD ratio	Inception [52]	1.08	1.64
	Our fingerprint	**111.41**	**39.96**

Table 3. Evaluation on {*real*, *ProGAN_seed_v#i*}. The best performance is highlighted in **bold**. "Our visNet" row indicates our fingerprint visualization network described in Section 3.3 and evaluated in Section 4.5.

		CelebA	LSUN
Accuracy (%)	kNN	10.88	10.58
	Eigenface [55]	23.12	-
	PRNU [45]	89.40	69.73
	Ours	**99.14**	**97.04**
	Our visNet	97.07	96.58
FD ratio	Inception [52]	1.10	1.29
	Our fingerprint	**80.28**	**36.48**

and finally results in distinct fingerprints. That motivates us to investigate the attribution performance among GAN instances that were trained with identical architecture and dataset but with different random initialization seeds.

Different initialization seeds. We next investigate the impact of GAN training initialization on image attributability. We train 10 ProGAN instances with the entire real dataset and with different initialization seeds. We sample $100,000$ images for classifier training and another $10,000$ images for testing. In this setup of {*real*, *ProGAN_seed_v#i*} where $i \in \{1, ..., 10\}$, we show the performance evaluation in Table 3. We conclude that it is the difference in optimization (e.g., caused by different randomness) that leads to attributable fingerprints. In order to verify our experimental setup, we ran sanity checks. For example, two identical ProGAN instances trained with the same seed remain indistinguishable and result in random-chance attribution performance.

Table 4. Classification accuracy (%) of our network w.r.t. downsampling factor on low-frequency or high-frequency components of {*real*, *ProGAN_seed_v#i*}. "L-f" column indicates the low-frequency components and represents the performances from the pre-downsampling network. "H-f" column indicates the high-frequency components and represents the performances from the pre-downsampling residual network.

Downsample factor	Resolution	CelebA L-f	CelebA H-f	LSUN L-f	LSUN H-f
1	128^2	99.14	99.14	97.04	97.04
2	64^2	98.74	98.64	96.78	96.84
4	32^2	95.50	98.52	91.08	96.04
8	16^2	87.20	92.90	83.02	91.58
16	8^2	67.44	78.74	63.80	80.58
32	4^2	26.58	48.42	28.24	54.50

Table 5. Classification accuracy (%) of our network w.r.t. patch size on {*real*, *ProGAN_seed_v#i*}.

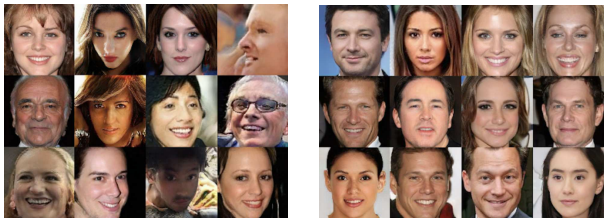
Pooling starts at	Patch size	CelebA	LSUN
4^2	128^2	99.34	97.44
8^2	108^2	99.32	96.30
16^2	52^2	99.30	95.94
32^2	24^2	99.24	88.36
64^2	10^2	89.60	18.26
128^2	3^2	13.42	17.10

4.3. Persistence: which image components contain fingerprints for attribution?

We systematically explore attribution performance w.r.t. image components in different frequency bands or with different patch sizes. We also investigate possible performance bias from GAN artifacts.

Different frequencies. We investigate if band-limited images carry effective fingerprints for attribution. We separately apply the proposed pre-downsampling network and pre-downsampling residual network for image attribution. Given the setup of {*real*, *ProGAN_seed_v#i*}, Table 4 shows the classification accuracy w.r.t. downsampling factors. We conclude that (1) a wider frequency band carries more fingerprint information for image attribution, (2) the low-frequency and high-frequency components (even at the resolution of 8×8) individually carry effective fingerprints and result in attribution performance better than random, and (3) at the same resolution, high-frequency components carry more fingerprint information than low-frequency components.

Different local patch sizes. We also investigate if local image patches carry effective fingerprints for attribution. We apply the post-pooling network for image attribution. Given the setup of {*real*, *ProGAN_seed_v#i*}, Table 5 shows the classification accuracy w.r.t. patch sizes. We conclude that for CelebA face dataset a patch of size 24×24 or larger carries sufficient fingerprint information for image attribution without deterioration; for LSUN, a patch of size 52×52



(a) Non-selected samples (b) Selected samples

Figure 5. Visual comparisons between (a) arbitrary face samples and (b) selected samples with top 10% Perceptual Similarity [66] to CelebA real dataset. We notice the selected samples have higher quality and fewer artifacts. They are also more similar to each other, which challenge more on attribution.

Table 6. Evaluation on the 10% selected images of {*real*, *ProGAN_seed_v#i*}. The best performance is highlighted in **bold**.

		CelebA	LSUN
Accuracy (%)	kNN	11.99	10.35
	Eigenface [55]	26.69	-
	PRNU [45]	93.50	74.49
	Ours	**99.93**	**98.16**
FD ratio	Inception [52]	1.04	1.22
	Our fingerprint	**15.63**	**6.27**

or larger carries a sufficient fingerprint.

Artifact-free subset. Throughout our experiments, the state-of-the-art GAN approaches are capable of generating high-quality images – but are also generating obvious artifacts in some cases. There is a concern that attribution might be biased by such artifacts. In order to eliminate this concern, we use Perceptual Similariy [66] to measure the 1-nearest-neighbor similarity between each testing generated image and the real-world dataset, and then select the 10% with the highest similarity for attribution. We compare face samples between non-selected and selected sets in Figure 5 and compare bedroom samples in the supplementary material. We notice this metric is visually effective in selecting samples of higher quality and with fewer artifacts.

Given the setup of 10% selected {*real*, *ProGAN_seed_v#i*}, we show the performance evaluation in Table 6. All the FD ratio measures consistently decreased compared to Table 3. This indicates our selection also moves the image distributions from different GAN instances closer to the real dataset and consequently closer to each other. This makes the attribution task more challenging. Encouragingly, our classifier, pre-trained on non-selected images, can perform equally well on the selected high-quality images and is hence not biased by artifacts.

4.4. Immunizability: how robust is attribution to image perturbation attacks and how effective are the defenses?

Attacks. We apply five types of attacks that perturb testing images [48]: *noise*, *blur*, *cropping*, *JPEG compression*, *relighting*, and random combination of them. The intention is to confuse the attribution network by destroying image fingerprints. Examples of the perturbations on face images are shown in Figure 6. Examples on bedroom images are shown in the supplementary material.

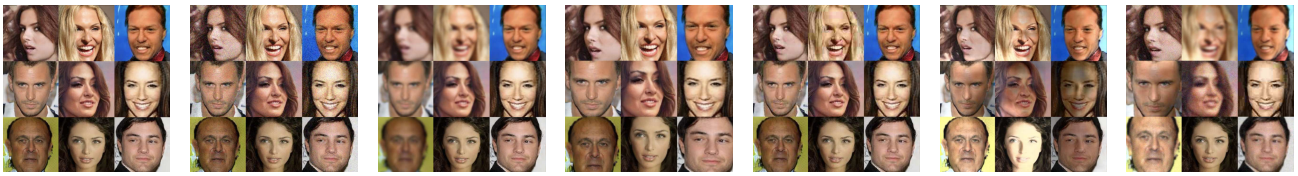
Noise adds i.i.d. Gaussian noise to testing images. The Gaussian variance is randomly sampled from $U[5.0, 20.0]$. *Blur* performs Gaussian filtering on testing images with kernel size randomly picked from $\{1,\ 3,\ 5,\ 7,\ 9\}$. *Cropping* crops testing images with a random offset between 5% and 20% of the image side lengths and then resizes back to the original. *JPEG compression* performs JPEG compression processing with quality factor randomly sampled from $U[10, 75]$. *Relighting* uses SfSNet [54] to replace the current image lighting condition with another random one from their lighting dataset. The combination performs each attack with a 50% probability in the order of *relighting*, *cropping*, *blur*, *JPEG compression*, and *noise*.

Given perturbed images and the setup of {*real*, *ProGAN_seed_v#i*}, we show the pre-trained classifier performances in the "Akt" columns in Table 7 and Table 8. All performances decrease due to attacks. In detail, the classifier completely fails to overcome *noise* and *JPEG compression* attacks. It still performs better than random when facing the other four types of attacks. The *relighting* attack is the least effective one because it only perturbs low-frequency image components. The barely unchanged fingerprints in high-frequency components enables reasonable attribution.

Defenses. In order to immunize our classifier against attacks, we finetune the classifier under the assumption that we know the attack category. Given perturbed images and the setup of {*real*, *ProGAN_seed_v#i*}, we show the finetuned classifier performance in the "Dfs" columns in Table 7 and Table 8. It turns out that the immunized classifier completely regains performance over *blur*, *cropping* and *relighting* attacks, and partially regains performance over the others. However, the recovery from *combination* attack is minimal due to its highest complexity. In addition, our method consistently outperforms the method of Marra *et al*. [45] under each attack after immunization, while theirs does not effectively benefit from such immunization.

4.5. Fingerprint visualization

Given the setup of {*real*, *ProGAN_seed_v#i*}, we alternatively apply the fingerprint visualization network (Section 3.3) to attribute images. We show the attribution performance in the "Our visNet" row in Table 3, which are com-



(a) No attack (b) Noise (c) Blur (d) Cropping (e) Compression (f) Relighting (g) Combination

Figure 6. Image samples for the attacks and defenses of our attribution network.

Table 7. Classification accuracy (%) of our network w.r.t. different perturbation attacks before or after immunization on CelebA {*real*, *ProGAN_seed_v#i*}. The best performance is highlighted in **bold**.

	CelebA											
	Noise		*Blur*		*Cropping*		*Compression*		*Relighting*		*Combination*	
	Atk	Dfs	Atk	Dfs	Atk	Dfs	Atk	Dfs	Atk	Dfs	Atk	Dfs
PRNU [45]	**57.88**	63.82	27.37	42.43	9.84	10.68	**26.15**	44.55	86.59	87.02	**19.93**	21.77
Ours	9.14	**93.02**	**49.64**	**97.20**	**46.80**	**98.28**	8.77	**88.02**	**94.02**	**98.66**	19.31	**72.64**

Table 8. Classification accuracy (%) of our network w.r.t. different perturbation attacks before or after immunization on LSUN bedroom {*real*, *ProGAN_seed_v#i*}. The best performance is highlighted in **bold**.

	LSUN											
	Noise		*Blur*		*Cropping*		*Compression*		*Relighting*		*Combination*	
	Atk	Dfs	Atk	Dfs	Atk	Dfs	Atk	Dfs	Atk	Dfs	Atk	Dfs
PRNU [45]	**39.59**	40.97	26.92	30.79	9.30	9.42	18.27	23.66	60.86	63.31	16.54	16.89
Ours	11.80	**95.30**	**74.48**	**96.68**	**86.20**	**97.30**	**24.73**	**92.40**	**62.21**	**97.36**	**24.44**	**83.42**

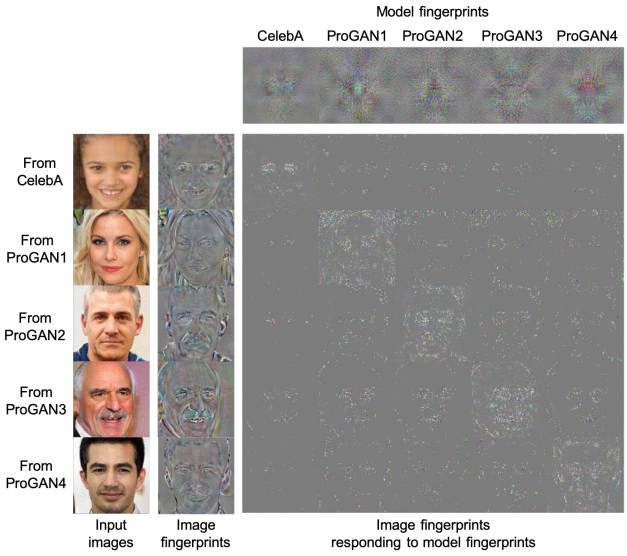


Figure 7. Visualization of model and image fingerprint samples. Their pairwise interactions are shown as the confusion matrix.

parable to that of the attribution model. Figure 7 visualizes face fingerprints. Bedroom fingerprints are shown in the supplementary material. It turns out that image fingerprints maximize responses only to their own model fingerprints, which supports effective attribution. To attribute the real-world image, it is sufficient for the fingerprint to focus only on the eyes. To attribute the other images, the fingerprints also consider clues from the background, which, compared to foreground faces, is more variant and harder for GANs to approximate realistically [2].

5. Conclusion

We have presented the first study of learning GAN fingerprints towards image attribution. Our experiments show that even a small difference in GAN training (e.g., the difference in initialization) can leave a distinct fingerprint that commonly exists over all its generated images. That enables fine-grained image attribution and model attribution. Further encouragingly, fingerprints are persistent across different frequencies and different patch sizes, and are not biased by GAN artifacts. Even though fingerprints can be deteriorated by several image perturbation attacks, they are effectively immunizable by simple finetuning. Comparisons also show that, in a variety of conditions, our learned fingerprints are consistently superior to the very recent baseline [45] for attribution, and consistently outperform inception features [52] for cross-source distinguishability.

Acknowledgement

This project was partially funded by DARPA MediFor program under cooperative agreement FA87501620191. We acknowledge the Maryland Advanced Research Computing Center for providing computing resources. We thank Hao Zhou for helping with the relighting experiments. We also thank Yaser Yacoob and Abhinav Shrivastava for constructive advice in general.

References

[1] Deep fakes: How they are made and how they can be detected. https://www.nbcnews.com/mach/video/deep-fakes-how-they-are-made-and-how-they-can-be-detected-1354417219989. 1

[2] How to recognize fake ai-generated images. https://medium.com/@kcimc/how-to-recognize-fake-ai-generated-images-4d1f6f9a2842. 8

[3] In the age of a.i., is seeing still believing? https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing. 1

[4] Model gallery. https://www.microsoft.com/en-us/cognitive-toolkit/features/model-gallery. 2

[5] The value of stolen data on the dark web. https://darkwebnews.com/dark-web/value-of-stolen-data-dark-web. 2

[6] You thought fake news was bad? deep fakes are where truth goes to die. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth. 1

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

[8] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger finder: Taking stylometry to the underground. In *Security and Privacy (SP), 2014 IEEE Symposium on*, pages 212–226. IEEE, 2014. 3

[9] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 2089–2093. IEEE, 2017. 2

[10] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 1, 2

[11] Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4970–4979, 2017. 1, 2

[12] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations*, 2019. 1

[13] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10. ACM, 2016. 1, 2

[14] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017. 2, 5

[15] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. Learning texture manifolds with the periodic spatial gan. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 469–477. JMLR. org, 2017. 2

[16] Paolo Bestagini, Simone Milani, Marco Tagliasacchi, and Stefano Tubaro. Local tampering detection in video sequences. In *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, pages 488–493. IEEE, 2013. 1

[17] Mikoaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 2, 5

[18] Luca Bondi, Silvia Lameri, David Guera, Paolo Bestagini, Edward J Delp, Stefano Tubaro, et al. Tampering detection and localization through clustering of camera-based cnn features. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1855–1864, 2017. 2

[19] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1, 2

[20] Peter Burt and Edward Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983. 4

[21] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukás. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3(1):74–90, 2008. 2, 3

[22] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164. ACM, 2017. 1, 2

[23] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 2

[24] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a cnn-based camera model fingerprint. *arXiv preprint arXiv:1808.08396*, 2018. 2, 3

[25] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. 5

[26] Luca DAmiano, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. A patchmatch-based dense-field algorithm for video copy–move detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(3):669–682, 2019. 1, 2

[27] Hany Farid. *Photo forensics*. MIT Press, 2016. 1, 2

[28] Jessica Fridrich. *Digital image forensics: there is more to a picture than meets the eye*. Springer New York, 2012. 2

[29] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 2

[30] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016. 2

[31] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1, 2

[32] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 1, 2, 4

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[35] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 1, 2

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1, 2, 4

[37] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678. ACM, 2014. 2

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1, 2, 5

[39] Gerhard C Langelaar, Iwan Setyawan, and Reginald L Lagendijk. Watermarking digital image and video data. a state-of-the-art overview. *IEEE Signal Processing Magazine*, 17(5):20–46, 2000. 3

[40] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 2

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5

[42] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006. 2, 3

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. 1

[44] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018. 1, 3

[45] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019. 2, 3, 4, 5, 6, 7, 8

[46] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2, 5

[47] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pages 43–47. ACM, 2018. 1, 3

[48] Seong Joon Oh, Max Augustin, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. In *Internation Conference on Representation Learning (ICLR)*, 2018. 7

[49] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 2

[50] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*, 2019. 2

[51] Lalit Kumar Saini and Vishal Shrivastava. A survey of digital watermarking techniques and its applications. *CoRR*, abs/1407.4735, 2014. 3

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2226–2234, 2016. 1, 2, 5, 6, 7, 8

[53] Husrev Taha Sencar and Nasir Memon. Digital image forensics. *Counter-Forensics: Attacking Image Forensics*, pages 327–366, 2013. 1

[54] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 7

[55] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *Optical Society of America*, 4(3):519–524, 1987. 5, 6, 7

[56] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, 2009. 3

[57] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017. 1

[58] Mitchell D Swanson, Mei Kobayashi, and Ahmed H Tewfik. Multimedia data-embedding and watermarking technologies. *Proceedings of the IEEE*, 86(6):1064–1087, 1998. 3

[59] 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. ACM, 2018. 1, 3

[60] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)*, 34(6):183–1, 2015. 1, 2

[61] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 1, 2

[62] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 269–277. ACM, 2017. 2, 3

[63] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[64] Ning Yu, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, and Michal Lukac. Texture mixer: A network for controllable synthesis and interpolation of texture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12164–12173, 2019. 1

[65] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 159–172. ACM, 2018. 2, 3

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 7

[67] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry 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. IEEE, 2017. 1, 2

[68] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1053–1061, 2018. 1, 2

[69] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. *ACM Transactions on Graphics (TOG)*, 37(4):49:1–49:13, 2018. 2

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 1, 2

[71] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 1, 2

[72] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018. 1