Adversarial Multimedia Forensics: Overview and Challenges Ahead

Mauro Barni
Dept.of Information Engineering
University of Siena, Italy
Email: barni@dii.unisi.it

Matthew C. Stamm
Dept. of Electrical and Computer Engineering
Drexel University, Philadelphia, USA
Email: mstamm@drexel.edu

Benedetta Tondi
Dept. of Information Engineering
University of Siena, Italy
Email: benedettatondi@gmail.com

***Abstract*—In recent decades, a significant research effort has been devoted to the development of forensic tools for retrieving information and detecting possible tampering of multimedia documents. A number of counter-forensic tools have been developed as well in order to impede a correct analysis. Such tools are often very effective due to the vulnerability of multimedia forensics tools, which are not designed to work in an adversarial environment. In this scenario, developing forensic techniques capable of granting good performance even in the presence of an adversary aiming at impeding the forensic analysis, is becoming a necessity. This turns out to be a difficult task, given the weakness of the traces the forensic analysis usually relies on. The goal of this paper is to provide an overview of the advances made over the last decade in the field of adversarial multimedia forensics. We first consider the view points of the forensic analyst and the attacker independently, then we review some of the attempts made to simultaneously take into account both perspectives by resorting to game theory. Eventually, we discuss the hottest open problems and outline possible paths for future research.**

I. INTRODUCTION

The development of Counter-Forensic (CF) techniques has proceeded in parallel with the design of multimedia forensic tools. Counter-forensic techniques are often successful due to the weaknesses of the traces the forensic analysis relies on. This is worsened by the fact that the majority of multimedia forensic tools are designed neglecting the possibility that an adversary may actively work to make the forensic analysis fail [1]. In reaction, several anti-CF techniques have also been developed in the last years, the most common approach consisting in looking for the traces left by the CF tools, and develop new forensic algorithms explicitly thought to expose documents subjected to specific CF techniques.

Early CF techniques were rather simple, as they consisted in the application of some basic processing operators [2]–[4]. When the attacker has enough information about the forensic algorithm, more effective CF techniques can be devised. Following a terminology adopted in adversarial machine learning [5], we can distinguish between attacks with Perfect Knowledge (PK) when the attacker has complete information about the forensic algorithm, and attacks with Limited Knowledge (LK), when the attacker knows only some details about the forensic algorithm. In the great majority of the cases, CF techniques are designed to attack a specific algorithm (targeted attacks) without paying attention to the possible countermeasures adopted by the analyst, e.g. by neglecting the fact that the CF attack may itself leave traces that can be revealed by the analyst. On the other hand, anti-CF techniques are developed, often by targeting a specific CF technique, without taking into account the possibility that the attacker foresees the moves of the analyst. The search for CF traces can be carried out by relying on new features explicitly designed for this goal [6]–[10], or by using the same features of the original forensic technique to design an adversary-aware version of the classifier [11], [12]. An obvious problem with the above approach occurs when the attacker anticipates that the traces left by the CF tools may themselves be subjected to a forensics analysis. In this case, we fall in a situation wherein CF and anti-CF techniques are iteratively developed in a never-ending loop, whose outcome can hardly be foreseen [13], [14]. A possible approach to avoid this problem is to design the forensic techniques in such a way that they are intrinsically more resistant to CF attempts or to resort to game theory to model the interplay between the analyst and the attacker, and use the performance at the equilibrium to evaluate which party will win the arms race [15], [16]. Though rather theoretical in nature, these works provide a natural framework to cast multimedia forensics in and can provide very useful insight about the achievable security of a wide class of multimedia forensic tasks [17].

In the rest of this paper, we overview the CF and anti-CF techniques developed so far and outline the most interesting challenges ahead. We do so by focusing on image forensic techniques, since research in this area is more advanced with respect to video and audio forensics. More specifically, in Section II and Section III, we adopt, respectively, the point of view of the attacker and the forensic analyst, assuming that they operate independently. Then, in Section IV, we review some attempts made to simultaneously take into account both perspectives by resorting to game theory. Eventually, in Section V, we list some open problems and outline possible paths for future research.

II. ATTACKER'S VIEW

By following the terminology introduced in [18], we focus on *exploratory* attacks, that is, attacks carried out at test time, since the large majority of the CF methods proposed so far belong to this category. With regard to the kind of errors the attacker aims at, CF attacks are usually *integrity violation* attacks [18], as they aim at avoiding that the manipulation is detected, that is, at causing a missed detection event. In the rest of the paper we use the following formalism: we indicate with A the CF method adopted by the attacker and with ϕ the forensic algorithm used by the analyst, or, simply, the detector. ϕ depends on: i.) the type of algorithm, its structure and its parameters l_i (as well as the learning algorithm, for data-driven methods), all together denoted by $\mathcal{L} = \{l_1, l_2, ..\}$; ii.) the feature space $\mathcal{X}$; iii.) the training data $\mathcal{D}$ (for data-driven approaches only). Therefore, $\phi = \phi(\mathcal{L}, \mathcal{X}; \mathcal{D})$ ($\phi(\mathcal{L}, \mathcal{X})$, for the model-based case). In the sequel, we refer to ϕ as $\phi(\mathcal{L}, \mathcal{X})$, the dependence on the training data being explicitly stated only when needed.

A. Attacks with perfect knowledge

In the PK scenario, the attacker can build the attack by relying on the knowledge of the forensic algorithm ϕ, and then he can apply a targeted attack [1]. In this case, it is possible for the attacker to induce a false positive decision error by introducing a limited, ideally minimum, distortion. Generally speaking, the attacker needs to solve an optimisation problem looking for the image which is in *some sense* closest to the image under attack and for which the output of the forensic analysis is the wrong one. Although such optimization is not always easy to solve, the exact knowledge of ϕ often allows to carry out very powerful CF techniques in closed form. This is the case of the CF method in [19], and the more general one in [20], for countering the model-based detectors of double (multiple) JPEG compression based on the analysis of the First Significant Digits (FSD), or the approaches in [21] and [22] against median filtering and copy move detection. When the detector is more complicated, as it is often the case with machine learning (ML) approaches, the optimum attack can be implemented by relying on Gradient-Descent solutions [5], [23], [24] or other iterative techniques such as L-BFGS, recently adopted for generating adversarial examples against deep neural networks [25]. Multimedia forensics is recently moving towards the use of deep learning architectures. A targeted attack to fool CNN-based camera model identification algorithms, based on the Fast Gradient Sign Method [26], is proposed in [27].

A problem with many PK approaches is that the CF algorithm is directly applied in the feature domain and it is difficult to control the distortion introduced in the pixel domain, all the more that the dependence between the pixel and feature domain is often non-invertible, thus also raising the problem of mapping back the attack (e.g., in [19]). When first order features of pixel or invertible transformed domains (e.g. the DCT domain) are considered, the image distortion can be controlled by operating in the feature domain as it is the case in [28], [29]. A gradient-based attack directly applied in the pixel domain, which then does not require invertibility of pixel and feature domain, is provided in [24].

B. Attacks with limited knowledge

We start by introducing a taxonomy to classify the attacks within this category.

- **Universal attacks**
  The attacker *only* knows the feature space (or class of features) $\mathcal{X}$. Since he is not aware of the statistic used by the analyst, he carries out an attack which is effective against *any* detector ϕ' inside the class $\Phi = \{\phi(\mathcal{L}', \mathcal{X}), \forall \mathcal{L}'\}$.
- **Attacks based on a surrogate detector**
  The attacker has a partial knowledge of the algorithm ϕ; for instance, he might know the feature space but not all the parameters of the algorithm and/or the training data. In this case, the attacker generates a *surrogate* detector $\hat{\phi}$ by exploiting the available information and making an educated guess about the parameters he does not know. Then, he builds the CF attack by performing a targeted attack against $\hat{\phi}$, hoping that the attack will also work against the real detector (attack transferability). Formally, if we let for instance l_1, l_2 be the unknown parameters, then $\hat{\mathcal{L}} = \{\hat{l_1}, \hat{l}_2, l_3, l_4...\}$ where $\hat{l_1}$ and $\hat{l}_2$ are the attacker's guesses of l_1 and l_2 and $\hat{\phi} = \phi(\hat{\mathcal{L}}, \mathcal{X}; \mathcal{D})$. The effectiveness is then assessed against ϕ.
- **Laundering attacks**
  The attacker has only a very general and limited knowledge of the algorithm; then, he tries to erase the CF traces by applying some basic processing operations (e.g. noise addition, recompression, resampling or filtering). In this case, the attacker does not target any specific detector or class of detectors.

As examples of attacks belonging to the first category we mention the *universal* CF methods in [28], [30] and [29], developed against the class of detectors based on first order statistics in pixel and DCT domain respectively, and applied to counter the detection of contrast enhancement and double (multiple) JPEG compression.

An example of attack *based on a surrogate detector* is the fingerprint-copy attack for PRNU-based camera identification [31]: the real camera fingerprint K ($K \in \mathcal{L}$) is unknown to the attacker, who then bases the attack on an estimation $\hat{K}$ made from a set of available images. Attacks to ML detectors often fall into this category: in fact, even if it is safely assumed that the attacker knows the kind of classifier used (e.g., an SVM, or a neural network), and also its parameters, he rarely has access to the same dataset $\mathcal{D}$ used by the analyst to train the detector. However, the attacker may build another dataset $\hat{\mathcal{D}}$, sampled from the same distribution, and use it in place of the real one, thus attacking an home-made replica of the detector $\phi(\mathcal{L}, \mathcal{X}; \hat{\mathcal{D}})$, see for instance [5], [24], [32]. Another LK attack for the case where the attacker knows only the feature space $\mathcal{X}$ and guesses both $\mathcal{L}$ and $\mathcal{D}$ is provided in [23]. It is worth stressing that such attacks work well under the assumption of attack transferability. Noticeably, standard ML tools are known to be sensitive to the problem of database mismatch, then, relying on home-made replica of ML classifiers is not always effective to build an attack which works against the real classifiers. This is less the case with deep learning architectures where the attack transferability assumption works well under a wide variety of scenarios [26].

We categorize as *laundering-type*, early CF techniques against detection of resampling [2], single and double JPEG compression [4], [33], contrast adjustments [3], median filtering [34], and splicing detection [35], just to mention a few.[1] Though very simplistic, the application of a post-processing operation has recently been shown to be very effective also against general SVM-based manipulation detectors trained on rich image representations [36]. A noticeable strength of such CF attacks with respect to most PK attacks is that they are much easier to implement; by applying a basic processing, in fact, the attacker can easily control the distortion introduced into the image.

III. ANALYST'S VIEW

We classify the solutions proposed so far to counter CF attacks according to the perspective adopted by the analyst, which can be tailored against a specific CF method or more general. In particular, we distinguish between **adversary-aware systems** and **intrinsically more secure detectors**.

A. Adversary-aware systems

The analyst, aware of the CF method the system is subject to, develops a new algorithm capable to expose the attack, by looking for the traces left by the CF tool. This is the most common approach used so far. In most cases, this goal is achieved by resorting to new, tailored, features. Then, a new algorithm ϕ_A is explicitly designed to reveal if the document underwent the CF attack, and used in conjunction with the original, unaware, algorithm ϕ. Such a view is adopted in [6], [8], [9], to address the adversarial detection of JPEG compression and median filtering. Among other examples, we mention the algorithm proposed in [37] for defeating the fingerprint-copy attack to PRNU-based camera identification and the one in [38] against the keypoint removal and injection attack to copy-move detectors. In other cases, the new algorithm is obtained by using the same features of the original algorithm ϕ and designing an adversary-aware version of the algorithm ϕ_A, which is then used in place of ϕ. This method is particularly suited for ML approaches, where the original detector is re-trained also with examples of attacked images to learn the statistical traces left by the CF algorithm. In this way, the analyst obtains a refined detector $\phi_A = \phi(\mathcal{L}, \mathcal{X}; \mathcal{D} \cup \mathcal{D}_A)$, where $\mathcal{D}_A$ is the set of attacked images used for training. In general, this approach is viable when the feature space is discriminative enough, i.e., it is capable to distinguish original, manipulated, as well as attacked images. Examples of this approach can be found in [11], [12] for adversarial double compression detection, and in [39] for a variety of manipulation detection problems with the JPEG laundering attacks. Exploiting the superior capabilities of deep architectures to learn good feature representations, adversary-aware training can also be used in image recognition applications to improve CNN robustness to adversarial examples [26].

We observe that by following the above approach, the analyst tries to exit the PK scenario, since it is (implicitly) assumed that the attacker keeps attacking the original algorithm ϕ. In other words, the analyst uses a system thought to reveal the traces introduced by an attacker which attacks a different system, namely the unaware algorithm, thus overlooking the game-theoretic nature of the problem (see Section IV).

B. Intrinsically more secure detectors

The analysts designs a system which is intrinsically more resistant to CF attempts, i.e. a system which is more difficult to attack even in the PK case. In this case, then, differently from the previous one, the analyst does not specialize the algorithm to work against a particular CF tool. Improved intrinsic security can be achieved in several ways. A possibility is to use higher order statistics; formally, the algorithm is refined by considering larger feature spaces $\mathcal{X}'$ ($\mathcal{X}' \supset \mathcal{X}$). This is done for instance for the detection of contrast enhancement [40], double JPEG [41] and local tampering [42], where resorting to second-order statistics allows the analyst to expose CF attacks and re-establish a correct analysis. Another approach consists in fusing the outputs of several forensic algorithms looking for different traces [43].

More in general, approaches belonging to this category look for solutions that work under a worst-case or a kind of most-powerful attack (MPA) A^*, namely, an attack that causes the largest damage when applied to the original (unaware) algorithm. Examples of MPA-aware detectors are provided in [11], [12], where the algorithm is refined by training on $\mathcal{D} \cup \mathcal{D}_{A^*}$. Another possibility is to resort to intrinsically *more secure* features, as done in general literature about ML security, by optimizing in some way the feature set, for instance by looking for the best feature set (in a large feature space) against a PK attack [32], or searching for intrinsically *more secure* architectures [44]. Randomizing the feature selection according to a secret key, thus preventing the attacker from gaining full knowledge of the system, is another way to design a more secure algorithm; such a strategy has been proven to be effective against PK attacks to SVM-based detectors [45].

IV. GAME THEORETIC VIEW

As we have seen in Section III, an intelligent analyst can design an adversary-aware detector ϕ_A anticipating the presence of a CF attack A. Under the PK scenario, however, an intelligent attacker can alter his attack to avoid detection by ϕ_A. The analyst, in turn, can again adjust his detector, leading to a dynamic interplay between the analyst and the attacker. Game theory can be used to study this dynamic interplay, and to identify optimal attack and detection strategies for the attacker and the analyst [13], [14].

The forensic scenarios described above are typically formulated as two player games [46], where the analyst's utility is defined as the probability of detecting a forgery and the attacker's utility is defined as the probability that the forgery is not detected. Since an increase in one player's utility leads to a corresponding decrease in the other player's utility, these games are known as zero sum games, i.e. games in which the sum of players' utilities is zero.

Game theory can be used to analyze the PK scenario where a CF attack A designed against an analyst's detector ϕ also

[1] Such attacks are often referred to as targeted attacks in the literature. However, we do not include them in the PK category, since the knowledge of the detector is only marginally exploited in these works. In most cases, the specific detector is only used to prove the attack effectiveness.

leaves behind its own detectable traces [7], [15]. An analyst can then form a refined detector ϕ_A by fusing the detection results from ϕ and a second detector ϕ' designed to detect A. The attacker can modulate the strength of A in an attempt to avoid detection while the analyst can alter the decision thresholds associated with ϕ and ϕ'. This setup has been used to identify the Nash equilibrium (NE) of the game in a scenario wherein the adversary aims at hiding the evidence of segment addition or deletion in a video sequence [7]. Game theory has also been used to analyze detection strategies and CF attacks in forensic source identification. In this scenario, a forensic analyst wishes to determine if a sequence originates from a known source X, while an attacker wishes to modify a sequence drawn from a different source Y in such a way that the analyst believes that it originated from X. This has important applications in PRNU-based camera model identification, where an adversary can attempt to falsify the PRNU pattern in a set of images. The asymptotic NE can be used to approximate the optimal detection and CF strategies of the attacker and the analyst for finite length sequences [16]. The set of source distributions that can not be distinguished reliably in the presence of an attack, can be identified when the analyst and adversary share the same training sequence, and when they utilize different sequences to empirically approximate a source's distribution [47]. Further analysis has been performed for the case when the attacker can also corrupt the analyst's training data [48].

V. Looking Ahead

Recently, deep learning techniques have significantly shifted the way researchers develop new forensic algorithms. Convolutional neural networks (CNNs) capable of automatically learning forensic feature extractors have been developed to address several problems in forensics such as manipulation detection [49]–[51] and camera model identification [50], [52]. While deep learning techniques are going to revolutionize multimedia forensics, they also open up new vulnerabilities that can be exploited by an attacker. It will be critical for researchers to understand new CF attacks that are enabled by deep learning and to search for ways to mitigate their effects. While a key advantage of CNNs is their ability to learn forensic features directly from data, an intelligent attacker can use this to his advantage. Since the space of possible inputs to a CNN is substantially larger than the set of images used to train it, an attacker can create modified images that fall into an 'unseen' space and force the CNN to misclassify. One method of accomplishing this involves introducing adversarial perturbations into an image. These perturbations are typically learned by computing the gradient of the loss function with respect to the input as done in the Fast Gradient Sign Method [26] and DeepFool attacks [53], or by using an iterative method such as the Jacobian-Based Saliency Map Attack [54]. As mentioned in Section II, a first CF attack based on this approach was recently proposed to fool CNN-based camera model identification algorithms [27].

Another significant threat is posed by Generative Adversarial Networks (GANs) [55]. GANs are a learning framework developed to create generative models capable of statistically mimicking the distribution of training data. This is done by iteratively training a discriminator to differentiate between real and generated samples of data and training the generator to produce samples capable of fooling the discriminator. GANs have been used by the computer vision community to produce visually realistic images [56] and even synthesized faces [57]. While the automatic creation of visually realistic images itself poses a forensic challenge, an even greater threat lies in the possibility that GANs can be used to create generators capable of producing forensically realistic images. Specifically, an attacker may be able to use a GAN to train a generator capable of falsifying forensic traces. A GAN capable of removing forensic traces left by median filtering [58] has already been developed, and it is very likely that more GAN-based CF attacks will be developed in the near future. Understanding the capabilities and limitations of deep learning-based attacks, and developing forensic measures to defend against or detect these attacks as they emerge will likely prove an important and difficult challenge for the future.

Acknowledgment

This work has been partially supported by a research sponsored by DARPA and AFRL under agreement number FA8750-16-2-0173 and PGSC-SC-111346-03. The U.S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and AFRL or the U.S. Government.

References

[1] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*, H. T. Sencar and N. Memon, Eds. Springer Berlin / Heidelberg, 2012.

[2] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 582–292, December 2008.

[3] G. Cao, Y. Zhao, R. Ni, and H. Tian, "Anti-forensics of contrast enhancement in digital images," in *Proc. ACM Workshop on Multimedia and Security*. New York, NY, USA: ACM, 2010, pp. 25–34.

[4] M. C. Stamm and K. J. R. Liu, "Anti-forensics of digital image compression," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 50–65, September 2011.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 387–402.

[6] S. Lai and R. Böhme, "Countering counter-forensics: The case of JPEG compression," in *Information hiding*. Springer, 2011, pp. 285–298.

[7] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Temporal forensics and anti-forensics for motion compensated video," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 4, pp. 1315–1329, Aug. 2012.

[8] G. Valenzise, M. Tagliasacchi, and S. Tubaro, "Revealing the traces of JPEG compression anti-forensics," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 2, pp. 335–349, 2013.

[9] H. Zeng, T. Qin, X. Kang, and L. Liu, "Countering anti-forensics of median filtering," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 2704–2708.

[10] O. Mayer and M. C. Stamm, "Countering anti-forensics of lateral chromatic aberration," in *ACM Workshop on Information Hiding & Multimedia Security*, 2017, pp. 15–20.

[11] M. Barni, Z. Chen, and B. Tondi, "Adversary-aware, data-driven detection of double JPEG compression: How to make counter-forensics harder," in *IEEE Int. Workshop on Information Forensics and Security*. IEEE, 2016, pp. 1–6.

[12] M. Barni, E. Nowroozi, and B. Tondi, "Higher-order, adversary-aware, double JPEG-detection via selected training on attacked samples," in *Proc. European Signal Processing Conference*, 2017, pp. 281–285.

[13] M. Barni and F. Pérez-González, "Coping with the enemy: advances in adversary-aware signal processing," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 26-31 May 2013, pp. 8682–8686.

[14] M. C. Stamm, M. Wu, and K. J. R. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, 2013.

[15] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Forensics vs anti-forensics: a decision and game theoretic framework," in *IIEEE Int. Conf. Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

[16] M. Barni and B. Tondi, "The source identification game: an information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, March 2013.

[17] ——, "Source distinguishability under distortion-limited attack: An optimal transport perspective," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145–2159, 2016.

[18] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 43–58.

[19] C. Pasquini, P. Comesaña-Alfaro, F. Pérez-González, and G. Boato, "Transportation-theoretic image counterforensics to first significant digit histogram forensics," in *EEE Int. Conf. Acoustics, Speech and Signal Processing*, May 2014, pp. 2699–2703.

[20] P. Comesana and F. Perez-Gonzalez, "The optimal attack to histogram-based forensic detectors is simple(x)," in *IEEE International Workshop on Information Forensics and Security*, Dec 2014, pp. 137–142.

[21] M. Fontani and M. Barni, "Hiding traces of median filtering in digital images," in *Proc. European Signal Processing Conference*. IEEE, 2012, pp. 1239–1243.

[22] T.-T. Do, E. Kijak, T. Furon, and L. Amsaleg, "Deluding image recognition in sift-based cbir systems," in *Proceedings of the 2Nd ACM Workshop on Multimedia in Forensics, Security and Intelligence*, ser. MiFor '10. New York, NY, USA: ACM, 2010, pp. 7–12. [Online]. Available: http://doi.acm.org/10.1145/1877972.1877977

[23] F. Marra, G. Poggi, F. Roli, C. Sansone, and L. Verdoliva, "Counter-forensics in machine learning based forgery detection." in *Media Watermarking, Security, and Forensics*, 2015, p. 94090L.

[24] Z. Chen, B. Tondi, X. Li, R. Ni, Y. Zhao, and M. Barni, "A gradient-based pixel-domain attack against SVM detection of global image manipulations," in *IEEE International Workshop on Information Forensics and Security*. Rennes, France: IEEE, 4-7 December 2017.

[25] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[27] D. Güera, Y. Wang, L. Bondi, P. Bestagini, S. Tubaro, and E. J. Delp, "A counter-forensic method for cnn-based camera model identification," in *IEEE CVPRW*. IEEE, 2017, pp. 1840–1847.

[28] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. ACM Multimedia and Security Workshop*, Coventry, UK, 6-7 September 2012, pp. 97–104.

[29] ——, "Universal counterforensics of multiple compressed jpeg images," in *Int. Workshop on Digital Watermarking*. Springer, 2014, pp. 31–46.

[30] P. Comesana-Alfaro and F. Pérez-González, "Optimal counterforensics for histogram-based forensics," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process*, 2013.

[31] T. Gloe, M. Kirchner, A. Winkler, and R. Bohme, "Can we trust digital image forensics?" in *ACM Multimedia 2007, Augsburg, Germany*, September 2007, pp. 78–86.

[32] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial feature selection against evasion attacks," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 766–777, 2016.

[33] P. Sutthiwan and Y. Q. Shi, "Anti-forensics of double JPEG compression detection," in *Shi Y.Q., Kim HJ., Perez-Gonzalez F. (eds) Digital Forensics and Watermarking. IWDW 2011. Lecture Notes in Computer Science, vol 7128*. Springer, 2012.

[34] Z.-H. Wu, M. C. Stamm, and K. R. Liu, "Anti-forensics of median filtering," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3043–3047.

[35] O. Mayer and M. C. Stamm, "Anti-forensics of chromatic aberration," in *Media Watermarking, Security, & Forensics*, vol. 9409. SPIE, 2015, p. 94090M.

[36] H. Li, W. Luo, X. Qiu, and J. Huang, "Identification of various image operations using residual-based features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2016.

[37] M. Goljan, J. Fridrich, and M. Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Trans. Information Forensics and Security*, vol. 6, no. 1, pp. 227–236, March 2011.

[38] A. Costanzo, I. Amerini, R. Caldelli, and M. Barni, "Forensic analysis of sift keypoint removal and injection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 9, pp. 1450–1464, Sept 2014.

[39] M. Boroumand and J. Fridrich, "Scalable processing history detector for jpeg images," *Electronic Imaging*, vol. 2017, no. 7, pp. 128–137, 2017.

[40] A. De Rosa, M. Fontani, M. Massai, A. Piva, and M. Barni, "Second-order statistics analysis to cope with contrast enhancement counter-forensics," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1132–1136, 2015.

[41] C. Chen, Y. Q. Shi, and W. Su, "A machine learning based scheme for double JPEG compression detection," in *19th International Conference on Pattern Recognition, 2008. ICPR 2008.* IEEE, 2008, pp. 1–4.

[42] X. Pan, X. Zhang, and S. Lyu, "Exposing image forgery with blind noise estimation," in *Proceedings of the thirteenth ACM multimedia workshop on Multimedia and security*. ACM, 2011, pp. 15–20.

[43] M. Fontani, A. Bonchi, A. Piva, and M. Barni, "Countering anti-forensics by means of data fusion," in *Media Watermarking, Security, and Forensics 2014*, vol. 9028. International Society for Optics and Photonics, 2014, p. 90280Z.

[44] B. Biggio *et al.*, "One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time," in *International Workshop on Multiple Classifier Systems*. Springer, 2015, pp. 168–180.

[45] Z. Chen, B. Tondi, X. Li, R. Ni, Y. Zhao, and M. Barni, "Secure detection of image manipulation by means of random feature selection," *submitted to IEEE Transactions on Information Forensics and Security*, 2014, arXiv preprint arXiv:1802.00573, 2017.

[46] R. B. Myerson, *Game theory*. Harvard university press, 2013.

[47] M. Barni and B. Tondi, "Binary hypothesis testing game with training data," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4848–4866, Aug. 2014.

[48] ——, "Adversarial source identification game with corrupted training," *accepted for publication, IEEE Trans. Information Theory, available on arXiv:1703.09244*, 2017.

[49] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *ACM Workshop on Info. Hiding & Multimedia Security*, 2016, pp. 5–10.

[50] ——, "Design principles of convolutional neural networks for multimedia forensics," *Electronic Imaging*, vol. 2017, no. 7, pp. 77–86, 2017.

[51] M. Barni *et al.*, "Aligned and non-aligned double jpeg detection using convolutional neural networks."

[52] L. Bondi, L. Baroffio, D. Gera, P. Bestagini, E. J. Delp, and S. Tubaro, "First steps toward camera model identification with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 259–263, Mar. 2017.

[53] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proc. IEEE CVPR*, no. EPFL-CONF-218057, 2016.

[54] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symp. Security & Privacy*, 2016, pp. 372–387.

[55] I. Goodfellow *et al.*, "Generative adversarial nets," in *Adv. in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[56] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint*, 2016.

[57] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[58] D. Kim, H. U. Jang, S. M. Mun, S. Choi, and H. K. Lee, "Median filtered image restoration and anti-forensics using adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 278–282, Feb. 2018.