Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.859
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Enhanced Adversarial Attack for Avoidance of Fake Image Detection

Kutub Uddin[a)] and Byung Tae Oh[a)‡]

Abstract

Image forensics is one of the most emerging topics in multimedia forensics to ensure the integrity of image content. Anti-forensic (AF) attacks, particularly generative adversarial network (GAN)-based attacks on fake images, can make forensic methods vulnerable. However, the effectiveness of AF attacks is limited to certain training conditions such as datasets, forensic methods, and attack types. Even though an AF attack is applied to misguide the forensic methods, forensic methods can be again updated using the AF dataset, which continues an infinite loop. This paper proposes an improved AF attack that can misguide all forensic methods. We update the forensic methods multiple times with multiple AF datasets and build an AF model that learns different forensic methods updated at different times. The experiments show that the proposed AF attack successfully deceives all forensic methods.

Keyword : Multimedia forensics, anti-forensic attacks, deep learning, generative adversarial network

Ⅰ. Introduction

Nowadays, multimedia information, particularly images and videos, has become an important part of our lives owing to its many applications. People are more interested in visual communication such as video conferencing. We also share images and videos of our daily lives over social platforms. An expert or an inexpert forger can manipulate these images and videos for malicious intentions [1]. Therefore, several forensic methods [2]-[5] have been introduced to detect manipulation traces to authenticate images and videos. However, anti-forensic (AF) attacks on manipulated images hide the manipulation fingerprints and misguide forensic methods. Sometimes, forensic methods can be updated based on the AF dataset. Consequently, AF attack fails to misguide forensic methods.

This paper proposes an improved AF attack using generative adversarial network (GAN) models to deceive all forensic methods. The major contributions are listed as follows:

a) Korea Aerospace University, Electornics and Information Engineering
‡ Corresponding Author : Byung Tae Oh
E-mail: byungoh@kau.ac.kr
Tel: +82-2-300-0409
ORCID: https://orcid.org/0000-0003-1437-2422

※ This research was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of ICT (NRF-2022R1A2C1005769), and by the GRRC program of Gyeonggi Province [2023-B02, Development of Intelligent Interactive Media and Space Convergence Application System].
· Manuscript September 27, 2023; Revised November 29,2023; Accepted November 29,2023.

We analyzed the effects of AF attacks on forensic methods and found that forensic methods could be further updated based on AF datasets. Therefore, the updated forensic methods can easily detect AF images. Similarly, AF attacks could be updated based on updated forensic methods. The updated AF attack can successfully deceive the updated forensic methods. This process continues in an infinite loop.

Thus, we updated forensic methods multiple times on multiple AF datasets. Then, we built an improved AF attack that randomly selects one of the forensic methods updated on an AF dataset and learned them to misguide.

We evaluated the proposed AF attack on two different deepfake datasets. The proposed method can successfully deceive forensic methods updated based on multiple AF datasets.

Ⅱ. Related Works

This section illustrates image forensic and AF attacks that work against each other to improve the performance.

1. Image Forensic

Over the past years, several forensic tools have been introduced to detect fake images such as noise addition, compression, filtering, and deepfake images. Wu et al. [2] proposed an end-to-end deep learning-based method for noise addition detection in online images. Hussain et al. [3] extracted discrete cosine transform (DCT) features using the DCT layer before convolution layers to provide end-to-end JPEG double compression detection. Dong et al. [4] integrated a magnifying layer to enlarge small-size images. Then, they applied a high-pass filter to transform them into multi-directional residuals for median filtering fingerprint detection.

Compared to noise addition, compression, and filtering, deepfake detection is a more practical and up-to-date technique for image forensics. Several methods have been proposed to detect deepfake images. Mara et al. [5] studied prior image forgery detectors against GAN-based image-to-image translation. Nataraj et al. [6] extracted a co-occurrence matrix on three color channels and used deep learning to detect image-toimage translations and facial expression swapping. Barni et al. [7] extended the co-occurrence-based approach to a cross-band co-occurrence matrix to explore the inconsistency between natural and deepfake images. Wang et al. [8] introduced a universal deepfake detection method in which they trained a deep learning-based classifier on a specific deepfake dataset and applied it to other deepfake datasets.

2. Anti-Forensic of Manipulated Image to Deceive Forensic Methods

Forensic methods [2]-[8] targeted detecting natural and fake images without considering any AF attacks. AF attacks on fake images alter the fingerprints and deceive the forensic methods. Unfortunately, forensic methods detect fake images as natural images. Several AF methods, particularly GAN based AF attacks [9]-[13] have been applied in numerous fields including compression, filtering, and deepfake domains. Kim et al. [9] designed a GAN model for median filtering restoration in which they applied high-pass filtering before the discriminator to capture high-frequency information. Uddin et al. [10] applied a GAN-based AF attack on double JPEG images to deceive double JPEG detectors. Zhao et al. [11]-[12] introduced a transferable GAN-based AF attack and applied it to median filtering and deepfake images to deceive median filtering and deepfake detectors. Zhang et al. [13] proposed a similar approach as a transferable GAN-based AF attack with local perturbation generation using regularization loss.

Ⅲ. Proposed Method

This section describes the motivation, GAN-based AF attack, and proposed adversarial attack.

1. Motivation

Forensic methods focus on detecting natural and fake images while AF attacks try to conceal or alter manipulation fingerprints from fake images to deceive forensic methods. If a transferable GAN-based AF is applied to deceive state-of-the-art forensic methods, forensic methods can again be updated based on the AF datasets. This process continues in an infinite loop. In this case, forensic methods updated at time t can detect only AF attacks at time $t-1$. Similarly, an AF attack updated at time $t+1$ can deceive only forensic methods updated at time t. Considering this limitation, we propose an improved AF attack that can deceive forensic methods updated at any time.

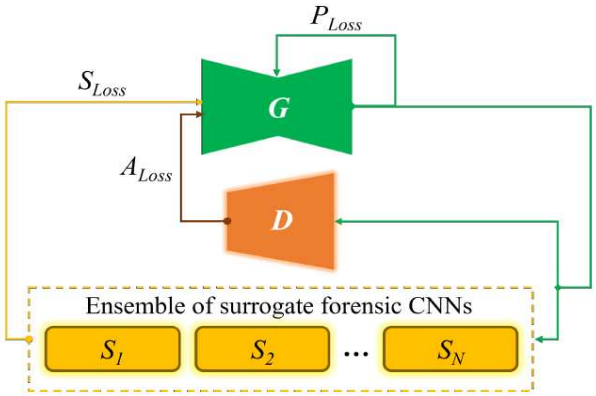

Fig. 1. GAN-based AF attacks on digital images

2. Generative Adversarial Network for AF Attack

GAN-based AF attacks can be grouped into conventional [9] and transferable [11]. Conventional GAN-based AF attack learns the underlying properties of real images to restore the fake images to make them similar to real images. On the other hand, a transferable AF attack learns the underlying properties of state-of-the-art forensic methods and generates new images close to fake images but deceives the forensic methods. Fig. 1 illustrates the working principle of two different GAN-based AF attacks. Conventional GAN-based AF attack is accomplished by training a GAN model with perceptual (P_{Loss}) and adversarial (A_{Loss}) losses defined as follows:

$$P_{Loss} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} | I(i,j) - G(I'(i,j)) | \quad (1)$$

where H and W indicate the height and width of the real and generated images, respectively. I and I' indicate real and fake images, respectively. G is the generator network in the GAN model.

$$A_{Loss} = \log(1 - D(G(I'))) \quad (2)$$

For transferable GAN-based AF attacks, we add a loss called surrogate loss (S_{Loss}), defined as follows:

$$S_{Loss} = -\sum_{i \in S}\sum_{j \in L} \log(C_i(G(I_j'))) \quad (3)$$

where S and L indicate the number of surrogate CNNs and the total number of samples, respectively. C_i is the $i-$th classification model of S.

3. Improved Adversarial Attack

Let us consider I_0 and I_0' to be real and fake images captured by a camera and generated by a GAN model. Then forensic methods are trained on I_0 and I_0' to detect them. Forensic methods can effectively detect real and fake images. However, if a GAN-based AF attack, particularly a transferable GAN-based AF attack, is applied on I_0', a new image (I_1') is generated as follows:

$$I_1' = G_{TAF}(I_0' ; P_{TAF}) \quad (4)$$

where GTAF is the generator model used to accomplish transferable AF attack with parameter PTAF.

Forensic methods trained on I_0 and I_0' can not detect I_1' as fake image. Unfortunately, it is detected as a real image. If the forensic methods are updated again using I_1', then they can detect I_1' as a fake image. Similarly, the AF attack is updated again using immediately updated forensic methods as an ensemble of surrogates and successfully deceives them. This process continues an infinite loop, as shown in Fig. 2.

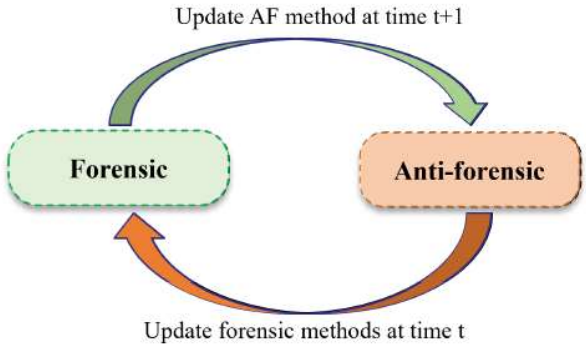

Fig. 2. Updating of forensic and AF methods in an infinite loop

If forensic methods and AF attacks are updated N times to compete against each other, then AF attacks (AF_t) updated at time t can deceive only forensic methods (FR_{t-1}) updated at time $t-1$. Sometimes, AFt fails to deceive the rest of the forensic methods.

We build an improved AF attack to deceive all forensic methods updated at any time. First, we update forensic methods N times with AF datasets, defined as follows:

$$FR_N = Update\,(AF_{N-1} ; P_{FR}) \quad (5)$$

where FRN is updated based on AF dataset AF_{N-1} with parameters, PFR. For the proposed method, we updated forensic methods five times ($N=5$). Update() function indicates training forensic methods at the time N with AF dataset from time $N-1$.

Now, we have N times updated forensic methods. To build an improved AF attack ($AF_{Improved}$) as shown in Fig. 3, we randomly select one of N times updated forensic methods at each iteration and updated the AF attack accordingly defined as follows:

$$AF_{Improved} = Update\,(FR_i ; P_{TAF}) \quad (6)$$

where i is any random integer and $i=1,2,\ldots,5$. Update() function indicates training $AF_{Improved}$ by randomly selecting one of N times updated forensic methods.

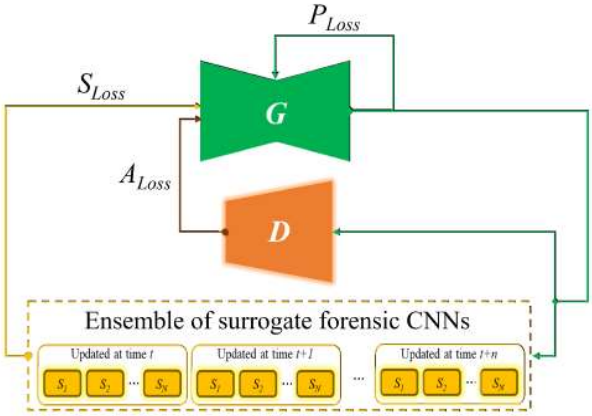

Fig. 3. Proposed improved AF attack

The proposed $AF_{Improved}$ attack can successfully deceive all the forensic methods. To train $AF_{Improved}$, we use perceptual, adversarial, and surrogate losses, respectively. The perceptual and adversarial losses are the same as defined in (1) and (2). We modified the surrogate loss as follows:

$$S_{Improved} = -\sum_{i \in S_N} \sum_{j \in L} \log(C_i(G(I_j'))) \quad (7)$$

where S_N is the randomly selected surrogate forensic method.

IV. Experimental Results

This section details the proposed method's implementation, dataset preparation, and performance evaluation.

1. Environmental Setup

We performed the whole experiment on Linux-20.4 with an NVIDIA GeForce GTX 24576MiB GPU. We used the PyTorch framework to implement forensic and AF methods. We adopted five deep learning models, such as InceptionV3 (C1) [14], Xception (C2) [15], DenseNet121 (C3) [16], MISLNet (C4) [17], and SRNet (C5) [18] to evaluate and compare forensic and AF methods. We trained the forensic methods until 50 epochs with a learning rate of 0.001 and batch size of 32. We used stochastic gradient descent (SGD) to optimize the forensic methods. The learning rate of forensic methods dropped by 10% after every 20 epochs. We trained the AF model until 15 epochs with a learning rate of 0.0001 for both generator and discriminator networks. We set a batch size of 32. The learning rate of the AF model drops by 10% after every five epochs. We used Adam and SGD optimizers to optimize generator and discriminator models.

2. Dataset Preparation

We conducted more practical forgery detection scenarios, such as deepfake detection to evaluate the proposed method. We used the Flickr-faces high quality (FFHQ) [19] and largescale scene understanding (LSUN) [20] datasets. The FFHQ and LSUN dataset contain human faces and different objects. We collected a total of 90,000 images for each dataset in which half of the data is real and half of the data is fake. We split the dataset into 60,000 images for training and 30,000 images for testing. We unified the spatial resolution of each dataset as 256×256.

3. Performance Evaluation of the Proposed Method on the FFHQ Dataset

First, we computed the baseline forensic ($FR_{Baseline}$) method to show the effectiveness of the deep learning model to detect fake images in the FFHQ dataset. We reported average detection result for C1, C2, C3, C4, and C5 in Table 1. The average detection accuracy is approximately 97.44% for the FFHQ dataset.

Table 1. Baseline detection results of FR methods on the FFHQ dataset

Method	Acc.
$FR_{Baseline}$	97.4%

Then we evaluated the effectiveness of forensic and AF methods that works against each other on the FFHQ dataset. We trained forensic (FR) and AF methods five times. Forensic methods are updated using AF datasets, while AF methods are updated using FR as an ensemble of surrogates in a transferable fashion. Table 2 lists the detection results of FR (FR1, FR2, FR3, FR4, and FR5) and AF (AF1, AF2, AF3, AF4, and AF5) updated five times.

Table 2. Comparisons of detection results for FR and AF methods that work against each other on the FFHQ dataset.

Method	AF1	AF2	AF3	AF4	AF5
FR1	48.9%	97.4%	73.5%	97.8%	97.1%
FR2	99.9%	50.1%	95.4%	50.1%	76.8%
FR3	83.3%	98.6%	48.9%	93.7%	95.3%
FR4	50.5%	50.1%	99.9%	50.0%	89.2%
FR5	84.4%	51.2%	78.1%	99.9%	50.0%

AF1 can successfully deceive FR1 because it is updated using FR1 as an ensemble of surrogates as given in bold

in Table 2. AF1 can also deceive FR4 but it can not deceive FR2, FR3, and FR5. The same trend appeared for AF2, AF3, AF4, and AF5. That is, a specific AF method that is updated using a specific FR as an ensemble of surrogates can deceive only that FR or some of FR methods but not all.

To solve this problem, we built an improved AF (AFImproved). Table 3 provides the detection results of the proposed AFImproved method on the FFHQ dataset. The proposed AFImproved successfully deceived all the forensic methods and achieved average detection accuracies of approximately 50% for all cases.

Table 3. Comparisons of detection results of the proposed $AF_{Improved}$ attack on the FFHQ dataset

Method	FR1	FR2	FR3	FR4	FR5
AFImproved	48.9%	50.1%	48.8%	50.1%	51.7%

4. Performance Evaluation of the Proposed Method on the LSUN Dataset

To show the generalizability of the proposed AFImproved, we applied it to the LSUN dataset. Table 4 gives the average detection results of FRBaseline on the LSUN dataset. The average detection accuracy is approximately 97.8%.

Table 4. Baseline detection results of FR methods on the LSUN daset

Method	Acc.
FRBaseline	97.8%

Similar to the FFHQ dataset, a specific AF method can deceive only a specific FR method used as an ensemble of surrogates, as given in Table 5 vacancy for the LSUN dataset. AF1 can only deceive FR1, but it can not deceive FR2, FR3, FR4, and FR5. In contrast, the proposed AFImproved achieved average accuracies of approximately 50% for the LSUN dataset, as in Table 6.

Table 5. Comparisons of detection results for FR and AF methods that work against each other on the LSUN dataset

Method	AF1	AF2	AF3	AF4	AF5
FR1	50.6%	97.5%	97.7%	97.5%	97.5%
FR2	99.8%	51.1%	51.9%	51.1%	57.2%
FR3	80.3%	98.0%	49.1%	84.8%	52.1%
FR4	81.6%	50.6%	99.8%	53.5%	89.8%
FR5	76.5%	77.6%	79.8%	98.7%	50.5%

Table 6. Comparisons of detection results of the proposed $AF_{Improved}$ attack on the LSUN dataset

Method	FR1	FR2	FR3	FR4	FR5
$AF_{Improved}$	53.2%	49.9%	53.4%	49.9%	54.7%

V. Conclusion

The development of deep learning has been significantly applied in image forensics and AF to work against each other. Forensic methods try to detect fake images, while AF methods misguide forensic methods. GAN-based AF attacks, particularly GAN-based transferable AF attacks can successfully deceive the forensic methods. However, forensic methods updated on the AF dataset can detect AF attacks. AF attack can deceive only the surrogate or some of the surrogate models used to update it. Therefore, we built an improved AF attack that can misguide all forensic methods. We trained forensic methods several times on different AF datasets. Then we used those pre-trained forensic methods to train an improved AF attack. The proposed improved AF attack is evaluated on two benchmark deepfake datasets and achieved promising results.

References

[1] A. Piva, An overview on image forensics, Int. Scholarly Research Notices, 2013.
doi: https://doi.org/10.1155/2013/496701

[2] H. Wu, J. Zhou, J. Tian, J. Liu, Robust image forgery detection over online social network shared image, in: Proceedings of the IEEE Conference on Comp. Vision and Pattern Recog., New Orleans, Louisiana, USA, ISBN: 978-1-6654-6946-3, 2022, pp. 13440-13449.

[3] I. Hussain, S. Tan, B. Li, X. Qin, D. Hussain, J. Huang, A novel deep learning framework for double JPEG compression detection of small size blocks, Journal of Visual Commun. and Image Represen. 80(2021) 103269.
doi: https://doi.org/10.1016/j.jvcir.2021.103269

[4] W. Dong, H. Zeng, Y. Peng, X. Gao, A. Peng, A deep learning approach with data augmentation for median filtering forensics, Multi. Tools and Appli. 81(2022) 11087-105.
doi: https://doi.org/10.1007/s11042-022-12040-w

[5] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Detection of gan-generated fake images over social networks, in: Proceedings of the IEEE Conference on Mult. Inf. Process. and Retrieval, Miami, FL, USA, ISBN: 978-1-5386-1857-8, 2018, pp. 384-389.

[6] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, B. S. Manjunath, Detecting GAN generated fake images using co-occurrence matrices, arXiv preprint arXiv:1903.06836, 2019.
doi: https://doi.org/10.48550/arXiv.1903.06836

[7] M. Barni, K. Kallas, E. Nowroozi, B. Tondi, CNN detection of GANgenerated face images based on cross-band co-occurrences analysis, in: Proceedings of the IEEE Int. Workshop on Inf. Foren. and Secur., New York City, NY, USA, ISBN 978-1-7281-9930-6, 2020, pp. 1-6.
doi: https://doi.org/10.1109/WIFS49906.2020.9360905

[8] S.Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, CNN-generated images are surprisingly easy to spot... for now, in: Proceedings of the Conference on Comp. Vision and Pattern Recog., Seattle, Online, USA, ISBN 978-1-7281-7168-5, 2020, pp. 8695-8704.

[9] D. Kim, H.U. Jang, S.M. Mun, S. Choi, H.K. Lee, Median filtered image restoration and anti-forensics using adversarial networks, IEEE Signal Process. Letters 25(2017) 78-82.
doi: https://doi.org/10.1109/LSP.2017.2782363

[10] K. Uddin, Y. Yang, B.T. Oh, Anti-forensic against double JPEG compression detection using adversarial generative network, in: Proceedings of the Korean Society of Broadcast Engineers Conference, Seoul, South Korea, 2019, pp. 58-60.

[11] X. Zhao, C. Chen, M.C. Stamm, A transferable anti-forensic attack on forensic CNNs using a generative adversarial network, arXiv preprint arXiv:2101.09568, 2021.
doi: https://doi.org/10.48550/arXiv.2101.09568

[12] X. Zhao, M.C. Stamm, Making GAN-generated images difficult to spot: a new attack against synthetic image detectors, arXiv preprint arXiv:2104.12069, 2021.
doi: https://doi.org/10.48550/arXiv.2104.12069

[13] H. Zhang, B. Chen, J. Wang, A local perturbation generation method for gan-generated face anti-forensics, IEEE Trans. on Cir. and Sys. for Video Tech., 2023, 33(2): 661-676.
doi: https://doi.org/10.1109/TCSVT.2022.3207310

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Comp. Vision and Pattern Recog., Las Vegas, Nevada,USA, ISBN 978-1-4673-8852-8, 2016, pp. 2818-2826.

[15] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Comp. Vision and Pattern Recog., Honolulu, Hawaii, USA, ISBN 978-1-5386-0458-8, 2017, pp. 1251-1258.

[16] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Comp. Vision and Pattern Recog., Honolulu, Hawaii, USA, ISBN 978-1-5386-0458-8, 2017, pp. 4700-4708.

[17] B. Bayar, M. C. Stamm, Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection, IEEE Trans. Inf. Forensic Secur., 2018, 13(11): 2691-706.
doi: https://doi.org/10.1109/TIFS.2018.2825953

[18] M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images, IEEE Trans. Inf. Forensic Secur., 2018, 14(5): 1181-93.
doi: https://doi.org/10.1109/TIFS.2018.2871749

[19] T. Karras, S. Laine, and T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE Conference on Comp. Vision and Pattern Recog., Long Beach, California, USA, ISBN 978-1-6654-4899-4, 2019, pp. 4401-4410.

[20] F. Yu, A. Seff, Y Zhang, S Song, T Funkhouser, J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365, 2015.
doi: https://doi.org/10.48550/arXiv.1506.03365

Introduction Authors

Kutub Uddin

- 2017 : Bachelor of Science, University of Chittagong
- 2020 : Master of Science, Korea Aerospace University
- 2020 ~ present : Ph. D. candidate, Korea Aerospace University
- ORCID : http://orcid.org/0000-0003-4365-682X
- Research interest : Image processing, Image forensic

Byung Tae Oh

- 2003 : Bachelor of Science, Yonsei University
- 2007 : Master of Science, University of Southern California
- 2009 : Ph. D., University of Southern California
- 2009 ~ 2013 : Samsung Advanced Institute of Technology
- 2013 ~ present : Professor, Korea Aerospace University
- ORCID : http://orcid.org/0000-0003-1437-2422
- Research interest : Image processing, Image forensic