https://doi.org/10.2352/ISSN.2470-1173.2020.4.MWSF-119

The Effect of Class Definitions on the Transferability of Adversarial Attacks Against Forensic CNNs

Xinwei Zhao and Matthew C. Stamm; Drexel University; Philadelphia, PA, xz355@drexel.edu, mstamm@coe.drexel.edu

Abstract

In recent years, convolutional neural networks (CNNs) have been widely used by researchers to perform forensic tasks such as image tampering detection. At the same time, adversarial attacks have been developed that are capable of fooling CNN-based classifiers. Understanding the transferability of adversarial attacks, i.e. an attacks ability to attack a different CNN than the one it was trained against, has important implications for designing CNNs that are resistant to attacks. While attacks on object recognition CNNs are believed to be transferrable, recent work by Barni et al. has shown that attacks on forensic CNNs have difficulty transferring to other CNN architectures or CNNs trained using different datasets. In this paper, we demonstrate that adversarial attacks on forensic CNNs are even less transferrable than previously thought even between virtually identical CNN architectures! We show that several common adversarial attacks against CNNs trained to identify image manipulation fail to transfer to CNNs whose only difference is in the class definitions (i.e. the same CNN architectures trained using the same data). We note that all formulations of class definitions contain the unaltered class. This has important implications for the future design of forensic CNNs that are robust to adversarial and anti-forensic attacks.

Introduction

The integrity and authenticity of multimedia contents are top concerns in many scenarios, such as criminal investigation and news reporting [1]. Research has shown that many editing operations, such as resizing [2] or contrast enhancement [3], will leave unique traces behind. Many forensic algorithms have been developed to detect or identify editing operations [4–15]. In recent years, convolutional neural networks (CNNs) have been widely used by researchers to perform forensic tasks such as image tampering detection [9, 16–18] and source identification [19–21].

In some scenarios, an intelligent attacker may attempt to launch an adversarial attacks to fool forensic algorithms [22–26]. Many adversarial attacks have been found to be able to fool deep learning based algorithms [27–35]. Researchers have already demonstrated that fast gradient sign method (FGSM) [36] and generative adversarial network (GAN) [37, 38] based attacks can be used to fool forensic CNNs. Therefore, it is important to understand the capability and limitations of the adversarial attacks.

Transferability is one of the well-known problems pertaining to adversarial attacks [39–42]. Transferability issues occur when the attacker attempts to attack a different CNN than the one that were explicitly trained against. Since many attacks operate by pushing the adversarial examples across the boundaries of the target class, it is important for the attacks to be able to observe the gradient of the target classifier with respect to the input data. However, when the CNN used to train the attack cannot fully mimic the boundaries of the target CNN, the obtained adversarial examples may not be able to transfer. Two common reasons that can cause attacks' transferability issues are training data discrepancy and CNN architecture discrepancy.

Understanding the transferability of adversarial attacks has important security implications. If information can be discovered that negatively effects an attacks transferability, it can be used to defend CNNs against attack. Additionally, knowledge of attack transferability helps researchers understand how feasible real-world adversarial attacks could be. While previous research has shown that attacks against object recognition CNNs can transfer to attack CNNs with different architectures or trained using different data, recent research in multimedia forensics shows an opposite result. Specifically, work by Barni et al. has shown that attacks on forensic CNNs have difficulty transferring to attack other CNN architectures or CNNs trained using different datasets [43].

In this paper, we demonstrate that adversarial attacks on forensic CNNs are even less transferrable than previously thought even between virtually identical CNN architectures! Particularly, we discover that several common adversarial attacks against forensic CNNs fail to transfer between CNNs whose only difference is in the class definitions (i.e. the same CNN architectures trained using the same data). We note that all formulations of class definitions contain the unaltered class. To investigate the impact of class definitions on forensic CNNs, we assume that attacker knows every details of the forensic CNNs, including the training data and CNN architecture. The only missing information of the attacker is the class definition. Next, we defined three typical class definitions for image manipulation forensic CNNs by grouping individual manipulation or parameterization of individual manipulation. Then we use the attacked images that are produced by fooling one forensic CNN to fool the other CNNs whose only difference is the class definition. We defined the successful attack rates (SARs) and transferability scores (T-scores) to measure the success and transferability of adversarial attacks. By conducting an extensive amount of experiments, we found that adversarial attacks are difficult to transfer to other class definitions of the same CNN architecture. Moreover, a secondary finding of ours is that binary classification of forensic CNNs (i.e grouping all manipulation into one class) performs slightly more robust than the other two class definitions. This has important implications for the future design forensic CNNs that are robust to adversarial and anti-forensic attacks.

Background

We assume that an attacker applies some editing operations to an images and then launches an adversarial attack attempted to

bypass the detection. The investigator will use a forensic CNN to identify if the image presented was unaltered or not.

For a single forensic manipulation identification CNN, there exists different ways to form class definitions. For instance, an binary decisions of unaltered or manipulated, multi-class definitions of unaltered vs several individual manipulations, or multi-class definitions of unaltered vs. several parameterized versions of individual manipulations. Each of the above class definitions includes the unaltered class.

Near-perfect knowledge scenario

Previous research has shown the attacker's knowledge pertaining to the target investigator's algorithm determines how easy and successful attacks can be [37, 42]. Therefore, depending on the amount of knowledge accessible to attackers, it is common to categorize the scenarios into the perfect knowledge scenario and partial knowledge scenarios. The perfect knowledge scenario is when attackers can observe the every detail of the investigator's algorithm or they can obtain an identical copy of the investigator's algorithm. Under the perfect knowledge scenario, attackers can directly integrate the investigator's CNN into their attack and train the attack explicitly bypass the detection of the identification CNN. All other scenarios are categorized as partial knowledge scenarios. Under partial knowledge scenarios, attackers has no full access to the investigator's CNN. As a result, attackers have to ensure their trained attack is capable of fooling different CNNs than the CNN explicitly trained against. If an attack fails to fool different CNNs, transferability of the attack occurs. Two common reasons that cause that attack's transferability are the dependencies of training data and CNN architectures [39, 43].

To investigate the transferability of adversarial attacks induced by class definition, we formulate a special partial knowledge scenario, the near-perfect knowledge scenario. Under this scenario, the attacker knows every details of the investigator's CNN architecture and also will use identical training data as the investigator. The only missing information of the attacker is the class definition of the target CNN (i.e the attacker does not know how the investigator forms the output classes for the forensic identification CNN.).

Investigation procedure

To investigate the impact of class definition on transferability of adversarial attacker, we used the following procedure: 1) We categorized three different class definitions that could be used by forensic CNNs attempting to identify image editing. 2) We trained six different forensic CNNs to perform editing detection and achieve their baseline performance under each class definition. 3) We implemented two popular adversarial attacks and obtain their Successful Attack Rate (SAR) in the perfect knowledge scenario (without attempting transfer). 4) We evaluated each attacks ability to transfer to an identical CNN whose only difference is the class definition used in the near perfect knowledge scenario, then interpreted the results. A detailed description of our experimental procedure, as well as the metrics used to evaluate the attacks is provided below.

Class definitions

There are several ways to define the classes used by a forensic CNN created to identify image manipulation. While all class definitions include the unaltered class other classes may differ depending on if different manipulations, as well as different parameterizations of manipulations, are grouped together into one class. In this work, we consider the following three different CNN class definitions.

Manipulation detection: In this class definition, only two classes are used: manipulated and unaltered. Any type of editing is grouped together into the manipulated class. This class definition would be used if the investigator only wants to know if an image has been modified in any means.

Manipulation classification: In this multi-class case, one class is assigned to unaltered along with one class for each individual editing operation. All parameterizations of that editing operation are grouped together into a single class. This class definition would be used if the investigator not only wants to know if the image has been modified, but also wants to know the individual manipulation applied to the image.

Manipulation parameterization: In this multi-class case, one class is assigned to unaltered and separate classes are assigned to each pair of manipulation and parameterization (or range of parameterizations). For example, median filtering with a 3x3 window would be a separate class than median filtering with a 5x5 window. This class definition could be used if the investigator wants to know very detailed information about a possible forger or identify inconsistencies in editing within an image.

Image forensic CNNs

In this paper, we examined six well-known CNN architectures, including MISLnet [9], TransferNet [44], PHNet [45], SR-Net [46], DenseNet [47] and VGG-19 [48]. While some of the CNN architectures were initially used for computer vision or steganalysis tasks, they can be adapted to train for image forensics.

For each CNN architecture, we trained forensic CNNs using the above three class definitions. All CNNs were trained using the same dataset created from the Dresden Image Database (more detail is provided in the results section). Furthermore, CNNs with the same architecture were trained using the same hyperparameters for all class definitions.

Adversarial attacks

To fool a forensic CNN, images modified by an attack should be classified as unaltered by that (or other) CNNs. As a result, attacks used our work operate in a targeted fashion, where the unaltered class is always the attacks target.

We used two well-known adversarial attacks in our experiments: the iterative targeted fast gradient sign method (I-FGSM) attack and the generative adversarial network (GAN) based attack. These two attack methods are very commonly used in anti-forensics (as well as the broader ML community), and are described below.

Iterative targeted fast gradient sign method (targeted I-FGSM): It operates by iteratively adding a small noise to the original image I and to push the adversarial examples I_{adv} to the target classes (i.e unaltered class in this context). At each iteration, the gradient is calculated with respect to the attacked image produced from previous iteration. The equation of targeted I-FGSM attacks is,

$$I^0_{adv} = I \tag{1}$$

$$I^{n+1}_{adv} = I^n_{adv} - \varepsilon \times sign \nabla_{I^n_{adv}} J(I^n_{adv}, y_{unaltered}) \tag{2}$$

where n denotes the index of iteration, ε denotes a small number, $J(\cdot)$ denotes the loss function, and $y_unaltered$ denotes target class label.

Generative Adversarial Network (GAN)-Based Attack: GAN-based method operates by training a GAN network to obtain a generator and then uses the generator to produce an image that can mimic the statistics of unaltered images.

A traditional GAN [28] is trained using a min-max function,

$$\min_G \max_D \mathbb{E}_{I\sim p_r(I)}[\log D(I)] + \mathbb{E}_{I_{adv}\sim p_g(I_{adv})}[\log(1-D(I_{adv}))] \quad (3)$$

where G denotes the generator, D denotes the discriminator, $p_r(I)$ denotes the distribution of unaltered images, $p_g(I_{adv})$ denotes the distribution of adversarial images and $\mathbb{E}$ denotes the operation of taking expected value.

We adopted MISLGAN method which was has been initially designed for fooling camera model identification CNNs [34]. MISLGAN is consisted of three major components, a generator, a discriminator and a pre-trained forensic CNN. While the generator and the discriminator are trained in the same fashion as the traditional GAN, the pre-trained is introduced to force the generator to produce an image that can mimic the forensic information of the "unaltered" image. To attack the manipulation detection CNNs, we modified MISLGAN by removing the synthetic CFA module in the generator. Due to the page limitation of the paper, we advise the readers to find details about the architecture and loss formulation of MISLGAN in the original paper.

Evaluation metrics

We define the successful attack rate and transferability score to evaluate the performance and transferability of the attack against the classifiers.

Successful attack rate (SAR): To evaluate the performance of the anti-forensic crafted images against manipulation detection CNNs, we calculated the percentage that the adversarial images are classified as unaltered by each CNN, and we define this percentage as successful attack rate (SAR). CNNs that have a stronger resistance to an anti-forensic attack should have lower SARs.

Transferability score (T-Score): To evaluate an attacks transferability, we calculated transferability score as the SAR of the unknown classifier over the SAR of the known classifier. The known classifier is directly used when launching the attack and the unknown classifier is used for classifying the adversarial images created by the attack. As a result, when an attack has good transferability, the transferability score should be high. Otherwise, the transferability would be low. For example, when all adversarial images produced by fooling one forensic CNN can fool other unseen CNNs, the transferability score equals to 1. We would like to point out that the transferability score should be positive and can be higher than 1. It is because the adversarial attack may be more effective on the unknown classifiers than the known classifiers, typically when the known classifiers are more resistant to the attack.

Experiments

We conducted a series of experiments to evaluated the transferability of multiple attacks against several forensic CNN architectures. Our database is created using 84,810 full-size JPEG images taken by 27 camera models from the Dresden Image Database [49] (images are from 70 unique devices). We randomly selected 80% for training, 10% for validation and 10% for testing. Next, we divided the full images into non-overlapping 256 by 256 image patches for each set. As a result, we ensure that there are no image patches from the same set coming from the same image and share the same statistics. To create the manipulated image patches, we selected three manipulations and five parameters that span a reasonable range for each manipulation. Then we manipulated each image patch in the database and obtained 15 unique sets of manipulated image patches. Along with the unaltered image classes, we obtained 16 classes corresponding to unaltered vs parameterized manipulations (manipulation parameterizer). These images were also grouped into 4 classes of unaltered vs individual manipulations (manipulation classifier), and 2 classes of unaltered vs manipulated (manipulation detector). Table 1 shows the chosen manipulations and parameters we used to created manipulated image classes. Due to computational cost constraints, we limited ourselves to three manipulations with five parameterizations each. Since we used 5 parameters per manipulation to create forged images, we in total created over 1,000,000 full sized JPEG images which are in bar with the up-to-date data size for training CNNs.

Table 1: Editing operations and their associated parameters.

Manipulations	Parameters
Additive Gaussian white noise	$\mu = 0, \sigma = 0.5, 1, 1.5, 2, 2.5$
Gaussian blurring	$\sigma = 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5$
Median filtering	window size$= 3, 5, 7, 9, 11$

Baseline performance of forensic CNNs

We started by training forensic CNNs using six CNN architectures and three class definitions. Each CNN was trained from scratch using stochastic gradient decent optimizer for 43 epochs and would stop early if validation accuracies started decreasing. The learning rate started with 0.0005 and would decrease to half every 4 epochs. Batch size was 25. On average, we achieved 99.29% accuracy using manipulation detector, 98.52% for manipulation classifier, and 77.93% for manipulation parameterizer. These results are consistent with the state-of-art performance for manipulation detection. Table 2 demonstrates the classification accuracies achieved by trained manipulation detection CNNs. Each entry corresponds one pairing of CNN architecture and class definition.

Table 2: Baseline classification accuracies achieved by six CNN architectures and three class definitions.

CNN Architect.	Manip. Detector	Manip. Classifier	Manip. Parameterizer
MISLnet	99.84%	99.55%	86.24%
TransferNet	99.20%	98.04%	65.27%
PHNet	99.58%	98.94%	86.58%
SRNet	99.16%	99.36%	81.30%
*DenseNet*_BC	98.13%	95.66%	65.50%
VGG-19	99.87%	99.50%	82.67%
Average	99.29%	98.51%	77.93%

Launching adversarial attacks

We started by creating set of images used for evaluating the attacks. From the testing set, we randomly selected 6,000 manipulated image patches that equally come from 15 manipulated image classes to form the *attack set.* Then we used the two attack methods to attack each image patch in the attack set and targeted at "unaltered" class. As a result, we obtained 216,000 anti-forensically attacked images.

For targeted I-FGSM, we chose ε in equation to be 0.1 and the iteration to attack each image to be 100. For the GAN-based attack, we started by training a generator targeted at the "unaltered" class for each forensic CNNs, and then we used the trained

generator to attack each image patch in the attack set. To train the generator, we randomly selected 360,000 manipulated image patches from the training set that equally come from 15 manipulated image. We trained the GAN-based attacked using the parameters in the original MISLGAN paper authored by Chen et al [34].

Baseline performance of adversarial attacks

In this experiment, we would like to show the performance of the adversarial attacks against forensic CNNs when the attacks were trained directly to target at the "unaltered" class of each forensic CNN. It corresponds to the scenario when the attacker has the perfect knowledge of investigators training data and full CNNs (i.e. including CNN architecture and the class definition).

Table 3: Baseline performance of targeted I-FGSM against forensic CNNs.

	Successful Attack Rate		
CNN Architect.	**Manip. Detector**	**Manip. Classifier**	**Manip. Parameterizer**
MISLnet	1.00	1.00	1.00
TransferNet	0.99	1.00	1.00
PHNet	0.87	0.96	1.00
SRNet	0.88	0.78	1.00
DenseNet	0.63	0.98	0.91
VGG-19	0.85	1.00	0.98
Average	0.87	0.95	0.98

Table 4: Baseline performance of GAN-based attack against forensic CNNs.

	Successful Attack Rate		
CNN Architect.	**Manip. Detector**	**Manip. Classifier**	**Manip. Parameterizer**
MISLnet	0.55	0.95	0.84
TransferNet	0.81	0.84	0.98
PHNet	0.90	0.97	0.94
SRNet	0.88	0.90	0.82
DenseNet	0.90	0.94	0.94
VGG-19	0.71	0.97	0.96
Average	0.79	0.93	0.91

Table 3 and Table 4 show the SARs we obtained for fooling forensic CNNs using I-FGSM and GAN-based attacks. Each entry is the individual SAR when targeting at a particular pair of CNN architecture and class definition. One average, manipulation detectors can be fooled with 0.87 SAR using I-FGSM and 0.68 using GAN-based attack. Manipulation classifiers can be fooled with 0.95 SAR using I-FGSM attack and 0.90 SAR using GAN-based attack. And manipulation parameterizers can be fooled with 0.98 SAR using I-FGSM and 0.91 SAR using GAN-based attack. First we noticed that under the perfect knowledge scenarios, both attacks can fool forensic CNNs with high SARs. Second, we noticed that for both attacks SARs for fooling manipulation detectors are consistently lower than the other two class definitions. For example, targeted I-FGSM achieved 0.63 SAR on the manipulation detector using DenseNet architecture, compared to the over 0.90 SARs for fooling the other two class definitions. GAN-based attack achieved 0.55 SAR for fooling manipulation detector using MISLnet architecture, compared to over 0.85 SAR for fooling the other two class definitions. These results may imply that the manipulation detectors are more robust to adversarial attacks under the perfect knowledge scenario.

Transferability of adversarial attacks

In this experiment, we evaluated the performance of the adversarial attacks against forensic CNNs when only the class definition of the target CNNs is changed. For each CNN architecture, we used forensic CNNs built with other class definitions to classify the adversarial images produced by individual attack. For example, if the adversarial images were produced to fool a MISLnet manipulation detector, we used the manipulation classifiers and parameterizers of MISLnet to classify these attacked images.

Table 5: Transferability of targeted I-FGSM attack re-targeting on manipulation classifiers and parameterizers.

	Successful Attack Rate		Transferability Score	
CNN Architect.	**Manip. Classifier**	**Manip. Parameterizer**	**Manip. Classifier**	**Manip. Parameterizer**
MISLnet	0	0	0	0
TransferNet	0	0	0	0
PHNet	0	0	0	0
SRNet	0	0	0	0
DenseNet	0	0	0	0
VGG-19	0	0	0	0
Average	0	0	0	0

Table 6: Transferability of targeted I-FGSM attack re-targeting on manipulation classifiers and parameterizers.

	Successful Attack Rate		Transferability Score	
CNN Architect.	**Manip. Detector**	**Manip. Parameterizer**	**Manip. Detector**	**Manip. Parameterizer**
MISLnet	0	0	0	0
TransferNet	0	0	0	0
PHNet	0	0	0	0
SRNet	0	0	0	0
DenseNet	0	0	0	0
VGG-19	0	0	0	0
Average	0	0	0	0

Table 7: Transferability of targeted I-FGSM attack re-targeting on manipulation detectors and classifiers.

	Successful Attack Rate		Transferability Score	
CNN Architect.	**Manip. Detector**	**Manip. Classifier**	**Manip. Detector**	**Manip. Classifier**
MISLnet	0	0	0	0
TransferNet	0	0	0	0
PHNet	0	0	0	0
SRNet	0	0	0	0
DenseNet	0	0	0	0
VGG-19	0	0	0	0
Average	0	0	0	0

Table 8: Transferability of GAN-based attack re-targeting on manipulation classifiers and parameterizers.

	Successful Attack Rate		Transferability Score	
CNN Architect.	**Manip. Classifier**	**Manip. Parameterizer**	**Manip. Classifier**	**Manip. Parameterizer**
MISLnet	0.004	0.045	0.007	0.082
TransferNet	0.008	0.005	0.010	0.006
PHNet	0.275	0.120	0.306	0.133
SRNet	0.420	0.000	0.477	0.000
DenseNet	0.005	0.010	0.008	0.016
VGG-19	0.020	0.090	0.024	0.106
Average	0.122	0.045	0.139	0.057

Table 9: Transferability of GAN-based attack re-targeting on manipulation detectors and parameterizers.

	Successful Attack Rate		Transferability Score	
CNN Architect.	**Manip. Detector**	**Manip. Parameterizer**	**Manip. Detector**	**Manip. Parameterizer**
MISLnet	0.090	0.035	0.095	0.037
TransferNet	0.000	0.000	0.000	0.000
PHNet	0.000	0.055	0.000	0.057
SRNet	0.050	0.005	0.056	0.006
DenseNet	0.000	0.000	0.000	0.000
VGG-19	0.525	0.260	0.541	0.268
Average	0.111	0.059	0.115	0.060

Table 10: Transferability of GAN-based attack re-targeting on manipulation detectors and classifiers.

	Successful Attack Rate		Transferability Score	
CNN Architecture	**Manip. Detector**	**Manip. Classifier**	**Manip. Detector**	**Manip. Classifier**
MISLnet	0.365	0.035	0.435	0.042
TransferNet	0.000	0.000	0.000	0.000
PHNet	0.065	0.490	0.069	0.521
SRNet	0.350	0.440	0.427	0.537
DenseNet	0.535	0.135	0.588	0.148
VGG-19	0.235	0.185	0.240	0.189
Average	0.259	0.214	0.290	0.240

Table 5 - 10 show the successful attack rates and transferability scores achieved by the two adversarial attacks. The left side of each table shows the SARs of fooling one particular pairing of CNN architecture and class definition, and the right side of each

table shows the T-Scores of each class definition with respect to the trained class definition. Table 5-7 shows that for targeted I-FGSM attack, both SARs and T-scores are 0's when re-targeting on different class definitions. It means the targeted I-FGSM attack cannot transfer to other class definitions.

For GAN-based attack, the average SARs are less than 26% and the average T-scores are less than 0.30. Shown in Table 8-10, the GAN-based attacks can slightly transfer when trained with particular paring of class definitions and CNN architectures. Among the 36 transferability cases we tested, only 4 cases achieved over 0.5 T-scores and 20 cases are less then 0.1. The highest T-score was achieved when the GAN-based attack were trained to fool manipulation parameterizer using DenseNet architecture, then re-targeted at manipulation detectors. However, there is still over 40% SAR drop taken in account that class definition was the only changed factor. These results demonstrated that adversarial attacks cannot transfer well across class definitions. Changing class definitions would significantly mitigate impact from adversarial attacks.

Conclusion

In this paper, we investigated the impact of class definitions on the transferability of adversarial attacks.While previous research has shown that the adversarial attacks cannot transfer across different CNN architectures or training data, we discovered that adversarial attacks are less transferable than previously thought. Particularly, by only changing the class definition of a forensic CNN, we can significantly decrease the the performance of adversarial attacks. The finding holds consistent when using multiple adversarial attacks to attack many well-known CNN architectures. Besides, a secondary finding shows that some class definitions may be more robust to adversarial attacks than others. Particularly, the SARs are lower when fooling binary detection under the perfect knowledge scenario.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 1553610. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number PGSC-SC-111346-03. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and 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 Air Force Research Laboratory (AFRL) or the U.S. Government.

References

[1] 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.

[2] M. Kirchner, "Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue," in *Proceedings of the 10th ACM workshop on Multimedia and security*, 2008, pp. 11–20.

[3] M. C. Stamm and K. R. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010.

[4] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A sift-based forensic method for copymove attack detection and transformation recovery," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1099–1110, Sep. 2011.

[5] J. Fridrich, D. Soukal, and J. Lukáš, "Detection of copy-move forgery in digital images," in *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[6] S. Bayram, H. T. Sencar, and N. Memon, "An efficient and robust method for detecting copy-move forgery," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 1053–1056.

[7] X. Pan and S. Lyu, "Region duplication detection using image feature matching," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 857–867, Dec 2010.

[8] O. Mayer and M. C. Stamm, "Accurate and efficient image forgery detection using lateral chromatic aberration," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1762–1777, July 2018.

[9] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, Nov 2018.

[10] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 1003–1017, 2012.

[11] M. Chen, J. Fridrich, J. Lukáš, and M. Goljan, "Imaging sensor noise as digital x-ray for revealing forgeries," in *International Workshop on Information Hiding*. Springer, Berlin, Heidelberg, 2007, pp. 342–358.

[12] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *2015 IEEE International Workshop on Information Forensics and Security*. IEEE, 2015, pp. 1–6.

[13] H. Farid, "Exposing digital forgeries from jpeg ghosts," *IEEE transactions on information forensics and security*, vol. 4, no. 1, pp. 154–160, 2009.

[14] P. Ferrara, M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni, "Unsupervised fusion for forgery localization exploiting background information," in *2015 IEEE International Conference on Multimedia Expo Workshops*, June 2015, pp. 1–6.

[15] 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. 28, no. 1, pp. 31–45, Jan 2018.

[16] O. Mayer and M. C. Stamm, "Forensic similarity for digital images," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1331–1346, 2020.

[17] L. Bondi, S. Lameri, D. Gera, P. Bestagini, E. J. Delp, and S. Tubaro, "Tampering detection and localization through clustering of camera-based cnn features," in *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, July 2017, pp. 1855–1864.

[18] B. Li, H. Zhang, H. Luo, and S. Tan, "Detecting double jpeg compression and its related anti-forensic operations with cnn," *Multimedia Tools and Applications*, 01 2019.

[19] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in *Information Forensics and Security (WIFS)*. IEEE, 2016, pp. 1–6.

[20] D. Cozzolino and L. Verdoliva, "Noiseprint: a cnn-based camera

model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2019.

[21] 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, March 2017.

[22] S. Sharma, A. V. Subramanyam, M. Jain, A. Mehrish, and S. Emmanuel, "Anti-forensic technique for median filtering using l1-l2 tv model," in *2016 IEEE International Workshop on Information Forensics and Security*, Dec 2016, pp. 1–6.

[23] M. Fontani and M. Barni, "Hiding traces of median filtering in digital images," in *Signal Processing Conference, Proceedings of the 20th European*. IEEE, 2012, pp. 1239–1243.

[24] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 582–592, 2008.

[25] G. Cao, Y. Zhao, R. Ni, and H. Tian, "Anti-forensics of contrast enhancement in digital images," in *Proceedings of the 12th ACM Workshop on Multimedia and Security*, 2010, pp. 25–34.

[26] 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. 1050–1065, Sep. 2011.

[27] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[29] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. Association for Computing Machinery, 2017, pp. 506–519.

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

[31] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy*, May 2017, pp. 39–57.

[32] 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," pp. 387–402, 2013.

[33] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," 2012.

[34] C. Chen, X. Zhao, and M. C. Stamm, "Misgan: An anti-forensic camera model falsification framework using a generative adversarial network," in *2018 25th IEEE International Conference on Image Processing*, Oct 2018, pp. 535–539.

[35] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," *arXiv preprint arXiv:1702.02284*, 2017.

[36] D. Gera, Y. Wang, L. Bondi, P. Bestagini, S. Tubaro, and E. J. Delp, "A counter-forensic method for cnn-based camera model identification," in *Computer Vision and Pattern Recognition Workshops*. IEEE, July 2017, pp. 1840–1847.

[37] C. Chen, X. Zhao, and M. C. Stamm, "Generative adversarial attacks against deep-learning-based camera model identification," *IEEE Transactions on Information Forensics and Security*, 2019.

[38] 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.

[39] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016.

[40] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

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

[42] M. Barni, M. C. Stamm, and B. Tondi, "Adversarial multimedia forensics: Overview and challenges ahead," in *2018 26th European Signal Processing Conference*. IEEE, 2018, pp. 962–966.

[43] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, "On the transferability of adversarial examples against cnn-based image forensics," pp. 8286–8290, 2019.

[44] Y. Zhan, Y. Chen, Q. Zhang, and X. Kang, "Image forensics based on transfer learning and convolutional neural network," in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, ser. IH&MMSec '17, pp. 165–170.

[45] M. Boroumand and J. Fridrich, "Deep learning for detecting processing history of images," *Electronic Imaging*, vol. 2018, no. 7, pp. 213–1, 2018.

[46] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1181–1193, 2018.

[47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[49] T. Gloe and R. Böhme, "The dresden image database for benchmarking digital image forensics," *Journal of Digital Forensic Practice*, vol. 3, no. 2-4, pp. 150–159, 2010.

Author Biography

Xinwei Zhao received her B.S. degree in electrical engineering in 2012 from Shandong University of Science and Technology, Qingdao, Shandong, China, and her M.S in electrical engineering in 2015 from Drexel University, Philadelphia, PA, US. Currently, she is a Ph.D candidate working in the Multimedia and Information Security Lab at Drexel University. Her research interests include multimedia forensics and anti-forensics, and deep learning.

Matthew C. Stamm received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland at College Park, College Park, MD, USA, in 2004, 2011, and 2012, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA, USA. He leads the Multimedia and Information Security Lab where he and his team conduct research on signal processing, machine learning, and information security with a focus on multimedia forensics and anti-forensics.

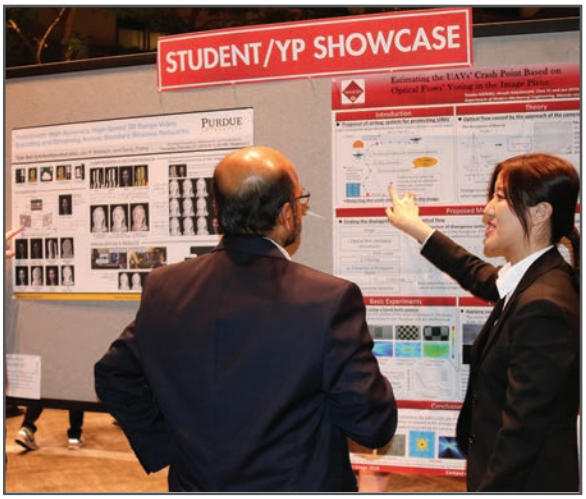
JOIN US AT THE NEXT EI!

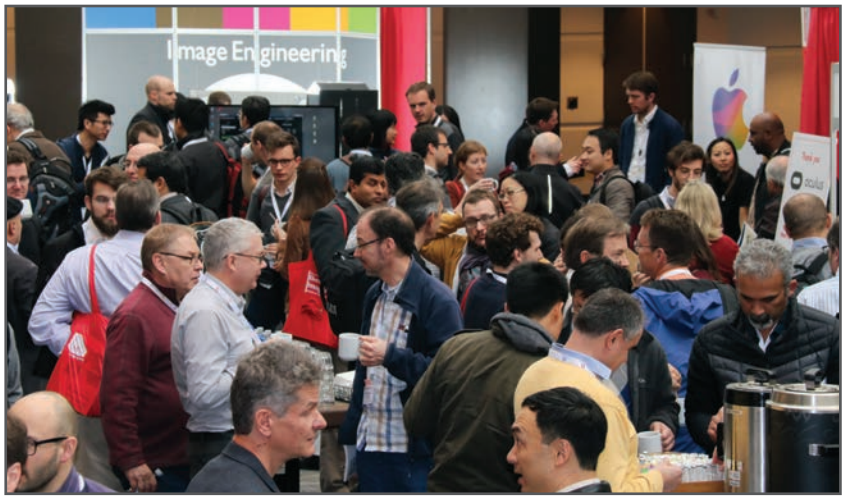
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org