

Adversarial Attack on Deep Learning-Based Splice Localization

Andras Rozsa
Verisk Analytics
Jersey City, NJ

andras.rozsa@verisk.com

Zheng Zhong
Verisk Analytics
Jersey City, NJ

zheng.zhong@verisk.com

Terrance E. Boulton
UCCS VAST Lab
Colorado Springs, CO

tboulton@vast.uccs.edu

Abstract

Regarding image forensics, researchers have proposed various approaches to detect and/or localize manipulations, such as splices. Recent best performing image-forensics algorithms greatly benefit from the application of deep learning, but such tools can be vulnerable to adversarial attacks. Due to the fact that most of the proposed adversarial example generation techniques can be used only on end-to-end classifiers, the adversarial robustness of image-forensics methods that utilize deep learning only for feature extraction has not been studied yet. Using a novel algorithm capable of directly adjusting the underlying representations of patches we demonstrate on three non end-to-end deep learning-based splice localization tools that hiding manipulations of images is feasible via adversarial attacks. While the tested image-forensics methods, EXIF-SC, SpliceRadar, and Noiseprint, rely on feature extractors that were trained on different surrogate tasks, we find that the formed adversarial perturbations can be transferable among them regarding the deterioration of their localization performance.

1. Introduction

There is an increasing demand for tools capable of detecting manipulated digital media, including fake pictures. Researchers proposed various approaches to detect and/or localize image manipulations. While detection only provides a verdict on the authenticity of the picture in question, localization shows the tampered pixels of the manipulated image and, therefore, gives us a better insight about the nature and goal of the modification. Recent advances in image forensics are, at least, partially due to the application of machine learning as some of the best performing splice localization systems are deep learning-based approaches [1, 15, 11, 7, 28, 24, 26].

Although the utilization of deep learning yields improvements in image forensics, deep learning models can be manipulated. Szegedy *et al.* [25] demonstrated that machine

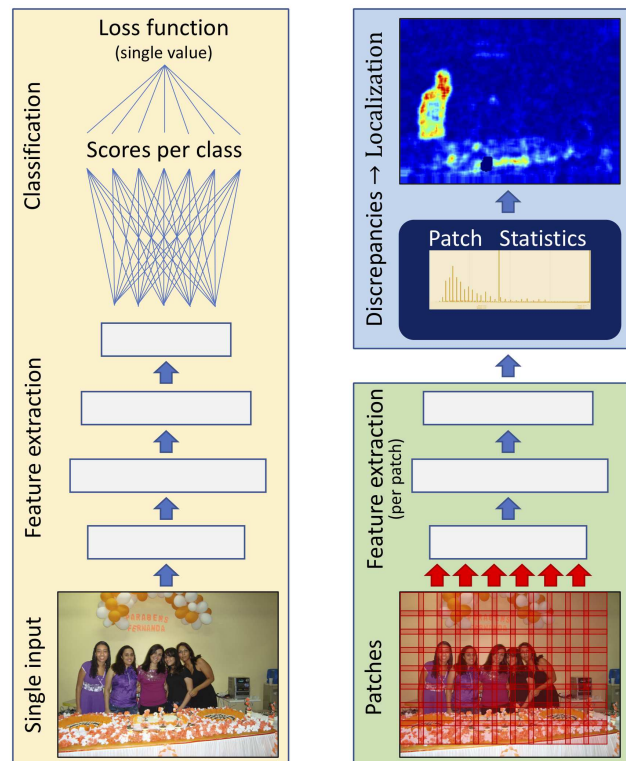


Figure 1: ADVERSARIAL EXAMPLE GENERATION. Adversarial attacks on networks performing recognition tasks, as shown on the left, are simple as there is one loss function that combines the scores from the classes and it can be back-propagated to the input image. Splice localization tools, displayed on the right, break their input image into 100s or 1000s of patches that often overlap. Then a feature representation is extracted for each patch, from which group statistics identify discrepancies yielding the localization of manipulated regions. In such systems there is no single loss function to target and there is also no differential path backwards through the statistical processes. This paper develops an alternative attack by directly manipulating the patch representations.

learning models, including the state-of-the-art deep neural networks, can misclassify slightly perturbed, otherwise correctly classified inputs. The existence of such adversarial

examples raises fundamental questions about the generalization properties and the real-world applications of machine learning models.

Researchers developed various counter-forensics tools to impede traditional multimedia-forensics methods – Barni *et al.* [3] provides a detailed overview on the subject area – but the adversarial robustness of the latest deep learning-based manipulation localization approaches has not been analyzed yet. While the latest image-forensics methods utilize deep learning models for extracting feature representations to localize manipulations, most adversarial attacks are capable of forming adversarial perturbations only on end-to-end recognition systems [25, 12, 17, 4, 19, 18, 8]. Consequently, as visually demonstrated in Figure 1, such approaches are not suitable for performing adversarial attacks on non end-to-end systems.

In this paper, we show that non end-to-end deep learning-based splice localization tools can be manipulated via adversarial attacks. We perform experiments on three of the best performing tools that are available, EXIF-SC [15], SpliceRadar [11], and Noiseprint [7]. Our experiments on three datasets demonstrate that the splice localization performance of the targeted deep learning-based image-forensics tools can significantly deteriorate due to our novel adversarial attack. Namely, by imperceptibly and simultaneously perturbing each image patch, we can hide manipulations of spliced images from the targeted image-forensics methods. We find that adversarial perturbations between the three image-forensics tools can be transferable, with different effectiveness. For instance, the adversarial examples generated on EXIF-SC tend to greatly decrease the localization performance of SpliceRadar. Interestingly, this transferability appears to be asymmetric as it does not occur in the other direction. Overall, our experimental results suggest that Noiseprint is the most robust among the three tested image-forensics tools.

2. Related Work

Many image-forensics techniques, especially manipulation detection and localization approaches, rely on artifacts that are introduced by the image formation pipeline. The idea is that manipulations, although they may not leave any visual clues, alter such underlying artifacts of images [20].

EXIF-SC proposed by Huh *et al.* [15] utilizes EXIF metadata for manipulation detection and localization. Given a pair of image patches, a Siamese network has been trained to predict the probability that patches share the same value for each of the 83 EXIF metadata attributes. Each branch of the network produces a 4096-dimensional feature vector which represents the EXIF metadata information of the particular patch. Following that, a two-layer Multi-Layer Perceptron (MLP) is applied to combine the EXIF metadata consistency predictions to an overall consistency.

Finally, mean shift [5] is used for finding the most consistent mode among all consistency maps of the selected patches.

SpliceRadar of Ghosh *et al.* [11] relies on low-level image statistics related to camera models and disregards high-level features connected to semantics. The feature extractor has been trained via a surrogate task of camera model identification. After an image patch is fed into a learned and constrained convolutional layer mimicking the filters proposed in [10], mutual information computed between the patch and its corresponding intermediate feature representation is used as a penalty to further suppress semantics. During inference, a 100-dimensional feature representation is obtained for each image patch. Similar to [6], a Gaussian Mixture Model (GMM) with two components is applied to localize genuine and forged regions of the particular image based upon the feature representation of the patches.

Noiseprint proposed by Cozzolino *et al.* [7] extracts noise residuals – called noiseprints – from images and utilizes those patterns as camera model fingerprints. Based on a convolutional image denoiser proposed in [27], the authors train a Siamese network to classify whether a pair of patches originate from the same camera model and location. Each branch of the network yields a 2-D feature map which represents the camera noise residuals. A distance-based logistic loss is applied during training to ensure small distances between noiseprints that belong to same camera model/location and large distances otherwise. During inference, noiseprint is extracted via the trained image denoiser for the particular image. With the assumption that genuine and forged patches originate from different camera models and, therefore, yield discrepancies on noise residuals of the corresponding patches, Noiseprint, similar to [6], utilizes a GMM to separate genuine and manipulated regions.

The feasibility and effectiveness of adversarial attacks as counter-forensics on non end-to-end deep learning-based manipulation localization approaches, such as EXIF-SC, SpliceRadar, or Noiseprint, have not been analyzed yet. Güera *et al.* [13] showed that an end-to-end Convolutional Neural Network (CNN), which was trained to identify the camera source of images, can be spoofed via adversarial attacks, such as the Fast Gradient Sign (FGS) method [12] or the Jacobian-based Saliency Map Attack (JSMA) [19]. Barni *et al.* [2] assessed the transferability of adversarial examples formed via the iterative FGS [17] and JSMA methods using two CNN architectures that were trained to detect image resizing and median filtering, respectively. Note that the two previously mentioned related works experiment with adversarial attacks on end-to-end networks. Also, those image-forensics algorithms perform manipulation detection rather than localization.

To perform adversarial attacks on EXIF-SC, SpliceRadar, or Noiseprint, adversaries need to directly manipulate the internal feature representations these image-forensics

methods rely on. The approach of Sabour *et al.* [23] yields adversarial examples that cause misclassifications and also mimic the internal representations of targeted inputs. They use the computationally expensive L-BFGS optimization technique, which limits the application of their technique. The Layerwise Origin-Target Synthesis (LOTS) of Rozsa *et al.* [21, 22] is similar as it is also capable of directly manipulating internal representations, but it is computationally more efficient. Inkawich *et al.* [16] used an approach very similar to LOTS to show its effectiveness of forming transferable adversarial examples between end-to-end deep learning-based recognition systems. To experiment with the targeted image-forensics methods, we adapt LOTS to simultaneously manipulate the feature representations of overlapping patches representing the spliced images.

3. Approach

While the targeted image-forensics algorithms solve the same localization task on manipulated images very differently, from a bird's-eye view, their architectures show some similarities. Namely, they extract feature representations from patches and utilize those for splice localization. While SpliceRadar and Noiseprint simply perform clustering based upon the extracted feature representations using a two-component Gaussian Mixture Model (GMM), EXIF-SC derives consistency maps for each patch using their feature representations and combines those to identify discrepancies. Consequently, if an adversary is able to modify pixels of patches and make feature representations of those similar, these image-forensics tools will fail to differentiate authentic and manipulated regions.

To be able to perform adversarial attacks on the targeted image-forensics methods, we adapt LOTS [21, 22] to simultaneously manipulate the feature representations of image patches that these splice localization tools rely on. Considering a manipulated image X and its corresponding set of patches P , we start with identifying authentic patches, denoted as A , from P . We define a patch authentic if and only if each of its pixels are authentic. Next, we specify the target representation t to be the mean feature representation of authentic patches. Our goal is to make both authentic and manipulated patches mimic t with their feature representations. To do so, for each patch we utilize a Euclidean loss defined on its feature representation $f(P_i)$ and the target representation t , and apply its gradient with respect to the pixels of patch P_i to move its feature representation closer to t . These patch-gradients can be defined as

$$\eta_i(P_i, t) = \nabla_{P_i} \left(\frac{1}{2} \|t - f(P_i)\|^2 \right), \quad i = 0, \dots, L, \quad (1)$$

where L is the number of patches for image X . Note that patches (P) and feature representations of patches ($f(P_i)$) are obtained as defined by the targeted image-forensics tool.

The targeted image-forensics methods can rely on multiple, sometimes overlapping, patches to localize manipulated regions of images. Therefore, we simply combine the calculated patch-gradients $\eta_i(P_i, t)$ according to their corresponding locations on the particular image X to obtain $G(P, t)$. Note that G has the same dimensions as X . We calculate and apply the combined gradients iteratively. Therefore, our algorithm can be formalized as

$$X_0 = X, \quad X_{n+1} = X_n - \alpha \frac{G(P^{(n)}, t)}{\|G(P^{(n)}, t)\|_\infty}, \quad (2)$$

where $P^{(n)}$ refers to the patches of X_n , and α defines the step-size. After applying the scaled and combined gradients to X_n , we clip pixels to be in the range of $[0, 255]$. Note that the number of patches and their locations remain the same, while pixel-values for $P^{(n)}$ can vary between iterations.

In our experiments, we run this algorithm for a predefined number of iterations with a fixed step-size α , and choose the perturbed image X_n with patches having the smallest average distance from target t . After identifying the best, we round pixel-values and save the resulting perturbed image.

4. Experiments

In this section, after describing the experimental setup, we present the results of our adversarial attacks on three non end-to-end deep learning-based image-forensics methods.

4.1. Experimental setup

To help reproducibility and future comparisons, we aim at revealing all details about our experiments and justifying various decisions we made for the experimental setup.

Targeted image-forensics algorithms. Due to the nature of white-box attacks, our selection is limited to approaches with available source code. Consequently, we choose EXIF-SC [15], SpliceRadar [11], and Noiseprint [7]. As the Columbia dataset contains smaller images than DSO-1, we slightly adjust SpliceRadar for the former dataset by changing the stride of patches from 48 to 24 pixels. This modification allows SpliceRadar to perform better due to the increased number of patches via the smaller stride. Regarding the adapted LOTS, while attacking EXIF-SC and SpliceRadar, we aim at mimicking the mean feature representations of authentic patches with the 4096- and 100-dimensional feature vectors of patches, respectively. The patch sizes and their locations are identical to those used for inference. For Noiseprint – based on preliminary experiments – we utilize 8×8 pixel, non-overlapping patches for both defining the target as the mean feature representation of authentic patches and then moving feature representations of all patches closer to the specified target. Other than

LOCALIZATION METHOD (results on spliced images)		COLUMBIA				DSO-1				NC16			
		F1	MCC	AUC	MAP	F1	MCC	AUC	MAP	F1	MCC	AUC	MAP
BL	EXIF-SC	0.911	0.878	0.980	0.952	0.568	0.517	0.845	0.526	0.377	0.359	0.796	0.340
	SpliceRadar	0.785	0.698	0.877	0.758	0.690	0.656	0.912	0.657	0.399	0.382	0.813	0.349
	Noiseprint	0.816	0.736	0.905	0.817	0.800	0.782	0.928	0.775	0.415	0.404	0.781	0.352
WBA	EXIF-SC	0.696	0.553	0.812	0.688	0.350	0.249	0.691	0.266	0.327	0.308	0.769	0.281
	SpliceRadar	0.583	0.365	0.691	0.498	0.362	0.274	0.687	0.273	0.305	0.289	0.756	0.246
	Noiseprint	0.748	0.636	0.859	0.769	0.573	0.515	0.812	0.506	0.332	0.311	0.730	0.273
TRA	EXIF-SC on SpliceRadar Ex.	0.896	0.860	0.973	0.937	0.557	0.507	0.845	0.516	0.378	0.360	0.795	0.341
	EXIF-SC on Noiseprint Ex.	0.887	0.849	0.970	0.933	0.556	0.502	0.840	0.513	0.375	0.360	0.797	0.338
	SpliceRadar on EXIF-SC Ex.	0.583	0.364	0.692	0.498	0.569	0.524	0.859	0.515	0.344	0.330	0.786	0.291
	SpliceRadar on Noiseprint Ex.	0.719	0.587	0.825	0.681	0.676	0.640	0.903	0.637	0.389	0.367	0.807	0.332
	Noiseprint on EXIF-SC Ex.	0.729	0.610	0.855	0.746	0.679	0.638	0.876	0.648	0.321	0.301	0.747	0.264
	Noiseprint on SpliceRadar Ex.	0.702	0.564	0.818	0.670	0.735	0.704	0.911	0.720	0.355	0.335	0.763	0.304
<i>All Authentic Attack</i>		0.570	0.276	N/A	N/A	0.279	0.136	N/A	N/A	0.174	0.125	N/A	N/A

Table 1: QUANTITATIVE RESULTS. This table shows performance measures for EXIF-SC, SpliceRadar, and Noiseprint on the manipulated images of the Columbia, DSO-1, and NC16 datasets, respectively. We show the baseline performances (BL) and how the manipulation localization performances of the targeted image-forensics methods change due to adversarial white-box attacks (WBA) and the transferability (TRA) of the formed adversarial perturbations. The last row (All Authentic Attack) presents the applicable measures for ideal attacks where each pixel is declared authentic by the targeted image-forensics tool.

the aforementioned modifications, we use the three image-forensics methods with their default settings.

Datasets. Regarding the publicly available datasets that are commonly used in the literature for evaluating image-forensics tools, we choose two on which the targeted image-forensics methods perform well. Therefore, forming adversarial perturbations for these spliced images is more challenging. These are the Columbia [14] and DSO-1 [9] datasets containing 180 and 100 spliced images, respectively, along with manipulation masks. Comparing the two datasets, we would like to highlight that the images of Columbia contain larger, very perceptible spliced regions. Note that we have excluded 5 of the 100 DSO-1 images from our experiments as the dimensions of the spliced images and their corresponding masks differ. In addition to Columbia and DSO-1, we also run experiments on the more challenging NC16¹ dataset which was collected for the Media Forensics Challenge². Compared to Columbia or DSO-1, NC16 is a relatively large dataset containing 564 images with a great variety regarding image sizes, sizes of manipulated regions, acquired camera models, etc.

Attacks. We utilize the algorithm that we described in Section 3. Based on small-scale experiments, we have chosen step-size of 5 pixels ($\alpha = 5$) and iterations to be limited to 50. Limiting the number of iterations naturally requires a larger step-size to be able to mimic the predefined target. Furthermore, with a larger step-size, it is less likely that the algorithm will get stuck in a local optima regarding the surface of the Euclidean loss. While these settings might yield sub-optimal results, we can limit the computational costs

LOCALIZATION METHOD (results on spliced images)		NC16 SUBSET			
		F1	MCC	AUC	MAP
BL	EXIF-SC	0.843	0.746	0.925	0.864
	SpliceRadar	0.823	0.734	0.912	0.827
	Noiseprint	0.837	0.754	0.868	0.799
WBA	EXIF-SC	0.644	0.542	0.851	0.631
	SpliceRadar	0.602	0.474	0.809	0.569
	Noiseprint	0.721	0.635	0.822	0.669
TRA	EXIF-SC on SpliceRadar Ex.	0.739	0.670	0.904	0.733
	EXIF-SC on Noiseprint Ex.	0.676	0.603	0.845	0.646
	SpliceRadar on EXIF-SC Ex.	0.803	0.719	0.917	0.821
	SpliceRadar on Noiseprint Ex.	0.729	0.646	0.859	0.705
	Noiseprint on EXIF-SC Ex.	0.793	0.714	0.917	0.810
	Noiseprint on SpliceRadar Ex.	0.764	0.678	0.898	0.760
<i>All Authentic Attack</i>		0.387	0.234	N/A	N/A

Table 2: QUANTITATIVE RESULTS ON NC16-SUBSET. This table shows performance measures on the subset of the NC16 dataset consisting of manipulated images for which EXIF-SC, SpliceRadar, and Noiseprint yield F1 scores greater than or equal to 0.6 (83 images). We show the baseline performances (BL), performances of the targeted image-forensics tools under adversarial white-box attacks (WBA), and the transferability (TRA) of the formed adversarial perturbations. The last row (All Authentic Attack) presents the applicable measures for ideal attacks where each pixel is declared authentic by the targeted image-forensics method.

and evaluate the effectiveness of a general attack. Note that due to our definition of an authentic patch and the default patch size of EXIF-SC (128×128 pixels), four images of the NC16 dataset do not have a single authentic patch and, therefore, cannot be attacked with our algorithm.

Evaluation metrics. To evaluate the adversarial robustness of the three targeted image-forensics methods, we compute and compare four performance measures that are commonly used in the literature [15, 11, 7] – F1 score, the Matthews Correlation Coefficient (MCC), the Mean Aver-

¹<https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation>

²<https://www.nist.gov/itl/iad/mig/media-forensics-challenge>

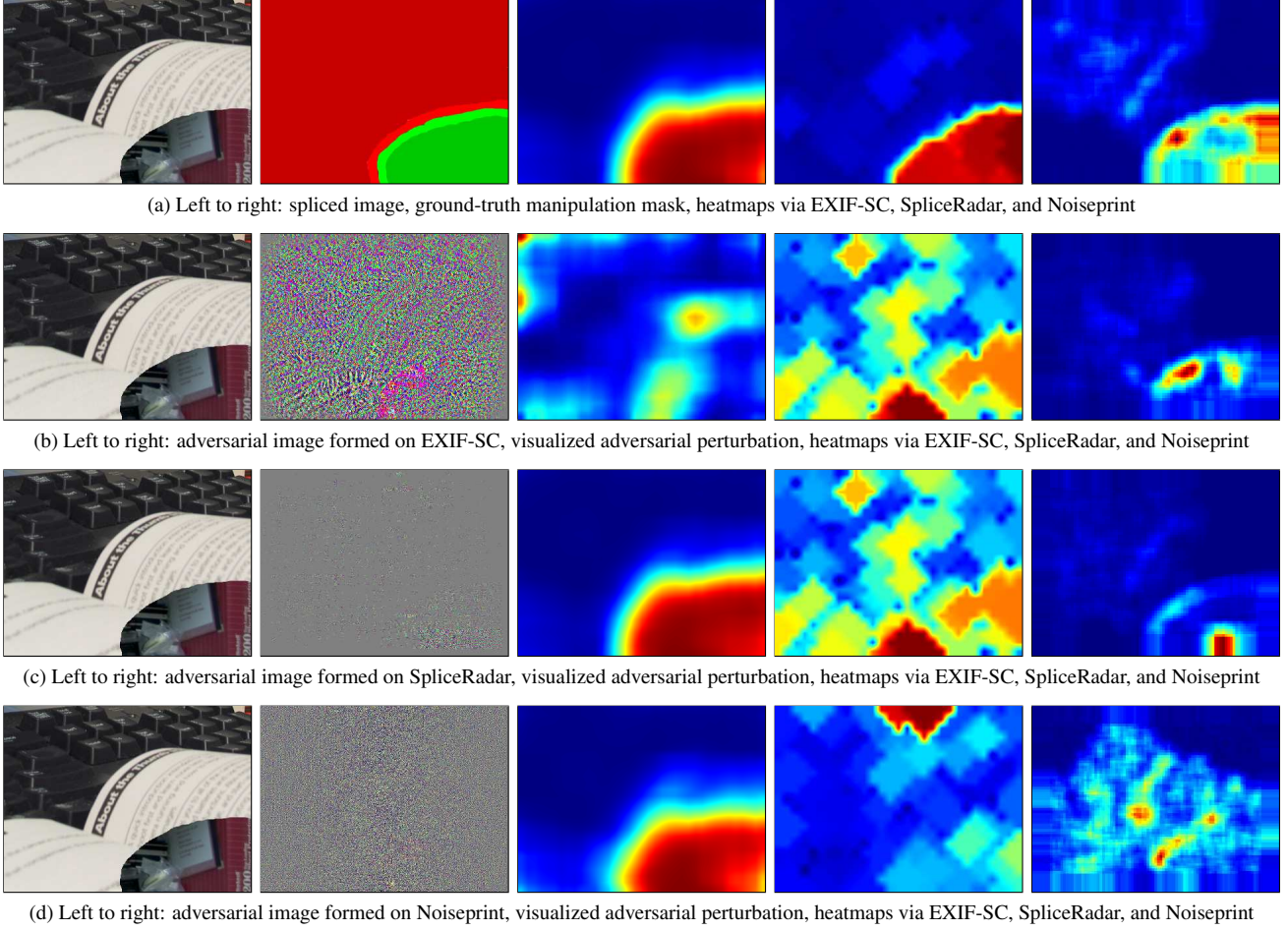


Figure 2: ADVERSARIAL EXAMPLES ON THE COLUMBIA DATASET. This figure presents the qualitative results of adversarial white-box and black-box attacks compared to the baseline performance of EXIF-SC, SpliceRadar, and Noiseprint on a spliced image. The visualized heatmaps are scaled; blue for lowest and red for highest values. Note that adversarial perturbations are depicted on a gray background and magnified by a factor of 100 for better visualization.

age Precision (MAP), and the area under the Receiving Operating Characteristic curve (ROC-AUC or AUC, in short) – before and after the adversarial attack takes place. Note that both F1 and MCC operate on binary decision maps, while forgery localization tools typically provide a probability map as output, which is then converted into a binary map using a threshold. Regarding the selection of the optimal threshold value, which is utilized to distinguish genuine and forged regions based upon the formed heatmap, we follow the same procedure as described in [15, 11].

4.2. Results

The quantitative results of our experiments are summarized in Table 1. To demonstrate the qualitative properties of our approach, we show examples for the Columbia, DSO-1, and NC16 datasets in Figures 2, 3, and 4, respectively. As we will see, on the more challenging NC16 dataset the tested image-forensics methods do poorly. Since at-

tacking an already struggling splice localization tool does not have much space for making further damage, we also present results on a subset containing the manipulated images for which EXIF-SC, SpliceRadar, and Noiseprint perform better regarding the localization of the manipulated areas ($F1 \geq 0.6$). The results for this subset consisting of 83 images of the NC16 dataset are presented in Table 2.

Baselines. As a starting point, we obtain the performance measures for the three targeted image-forensics algorithms on the forged images of the three datasets. We can see in Table 1 that EXIF-SC outperforms SpliceRadar and Noiseprint on the Columbia dataset but falls behind them on the more challenging DSO-1 and NC16 datasets. In Figures 2(a), 3(a), 4(a), we show the qualitative splice localization performances of EXIF-SC, SpliceRadar, and Noiseprint on spliced images of the Columbia, DSO-1, and NC16 datasets, respectively. As we can see, the three image-forensics methods can more or less successfully lo-

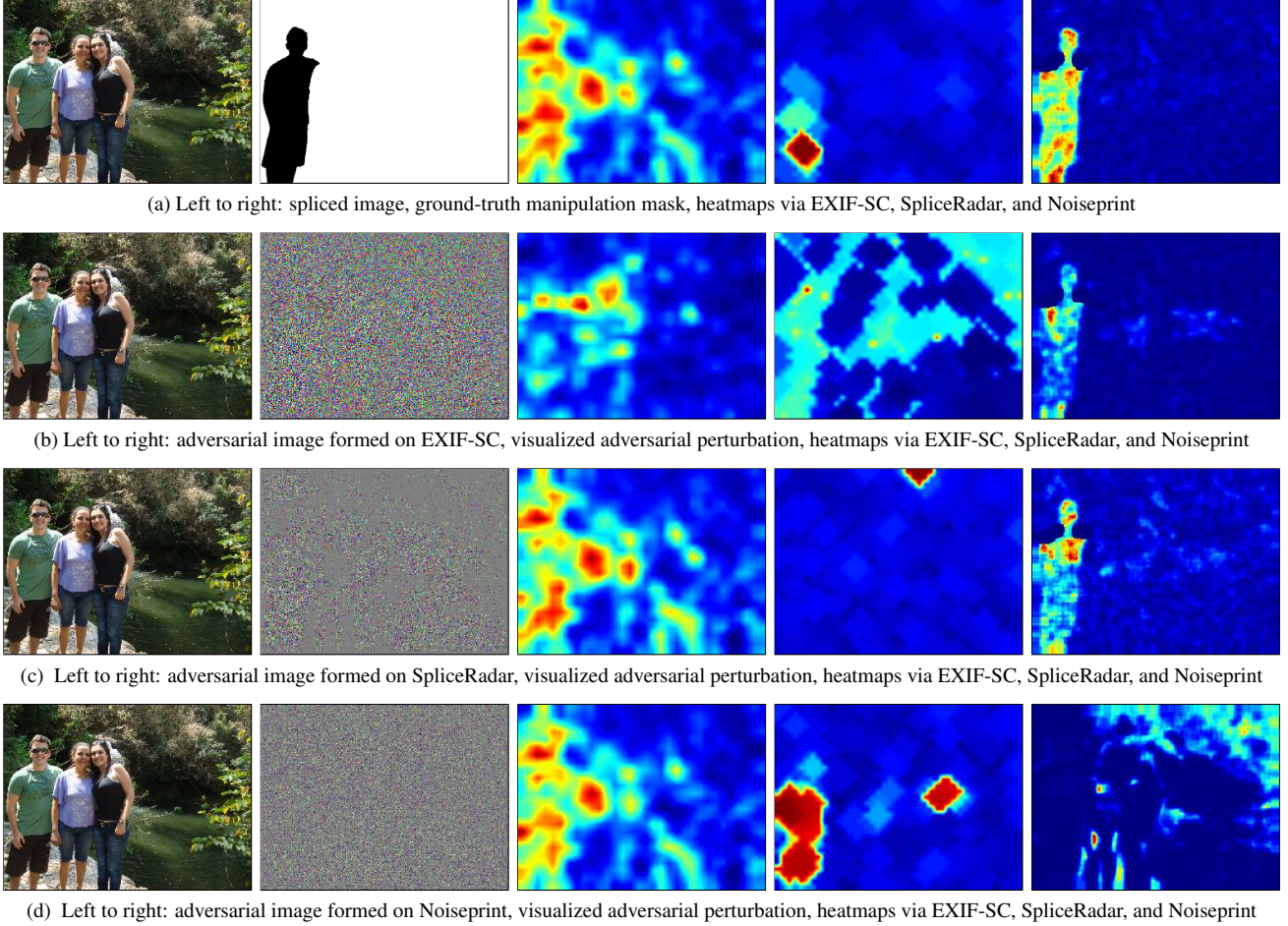


Figure 3: ADVERSARIAL EXAMPLES ON THE DSO-1 DATASET. This figure demonstrates the qualitative results of adversarial white-box and black-box attacks compared to the baseline performance of EXIF-SC, SpliceRadar, and Noiseprint on a spliced image. The visualized heatmaps are scaled; blue for lowest and red for highest values. Note that adversarial perturbations are depicted on a gray background and magnified by a factor of 100 for better visualization.

calize the manipulated regions denoted by the manipulation masks as the authentic and forged regions can be separated on the computed heatmaps by applicable threshold values. Note that EXIF-SC also produces some false positive splice localization on the DSO-1 example shown in Figure 3(a).

White-box attacks. To put the results of our adversarial white-box attacks on EXIF-SC, SpliceRadar, and Noiseprint into perspective, we show the applicable performance measures for an “ideal attack” in the last row of Table 1 labeled as *All Authentic Attack*. In short, this ideal attack occurs when each pixel – both authentic and manipulated – of the manipulated image is declared as authentic by the image-forensics tool. Note that this ideal attack does not yield lower bounds regarding the performance measures that we can achieve. Due to the fact that SpliceRadar and Noiseprint utilize a Gaussian Mixture Model (GMM) with two components, they always declare authentic and manipulated regions. Consequently, it is impossible to per-

form an ideal attack on these two image-forensics tools. Compared to baselines, we can see in Table 1 that white-box attacks greatly decrease the performance measures of both EXIF-SC and SpliceRadar, and we can notice that Noiseprint is more resilient, especially on the Columbia dataset. The qualitative results depicted in Figures 2(b)-(d), 3(b)-(d), and 4(b)-(d) (leftmost images) demonstrate that the formed adversarial examples contain imperceptibly small perturbations that are capable of impeding the targeted image-forensics methods. Considering the visualized perturbations, we notice that those formed on EXIF-SC are stronger than the others generated on SpliceRadar and Noiseprint. Note that due to our algorithm mimicking the target representation with each patch, the false positive splice localizations via EXIF-SC for the DSO-1 example are all declared as authentic regions for the adversarially perturbed image shown in Figure 3(b) (third image). Overall, the qualitative results visually demonstrate that the ad-

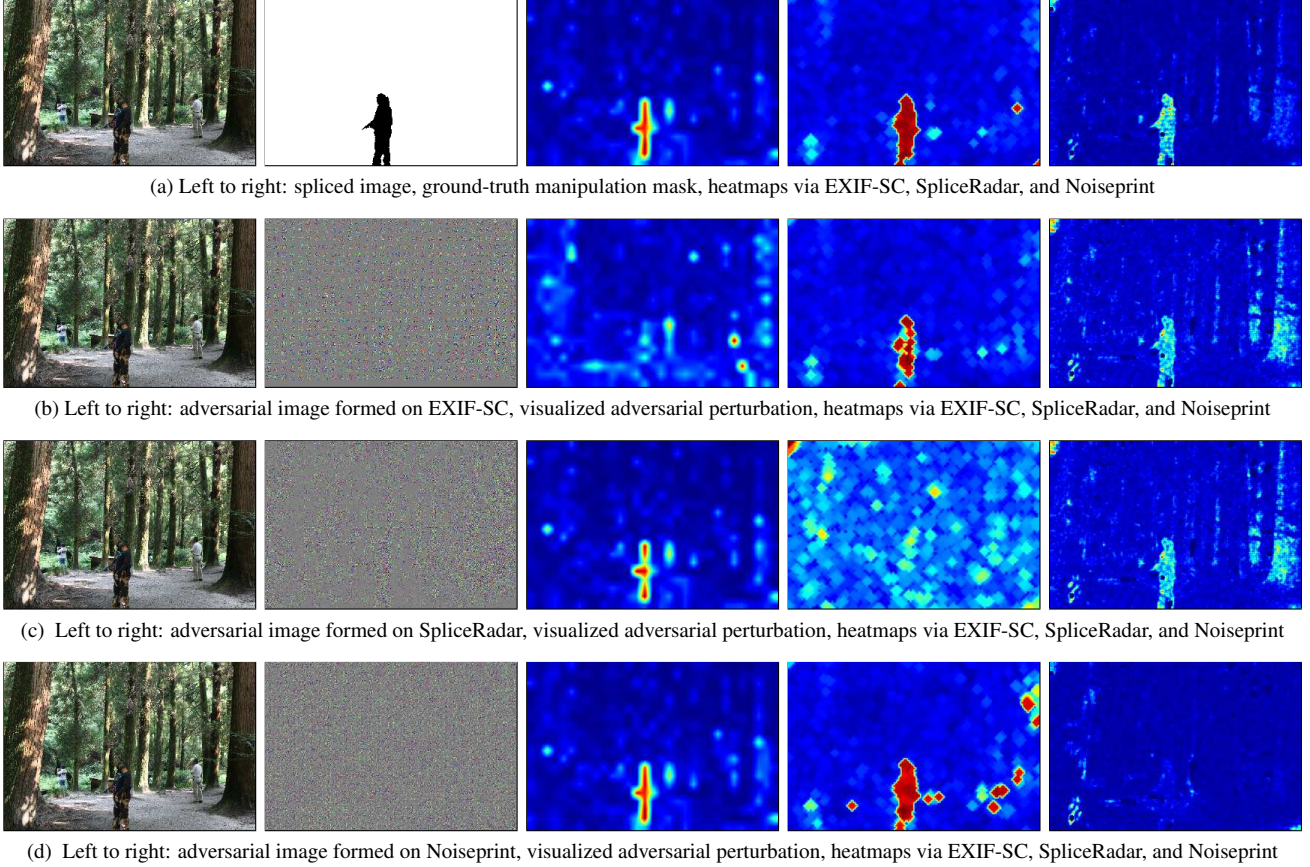


Figure 4: ADVERSARIAL EXAMPLES ON THE NC16 DATASET. This figure demonstrates the qualitative results of adversarial white-box and black-box attacks compared to the baseline performance of EXIF-SC, SpliceRadar, and Noiseprint on a spliced image. The visualized heatmaps are scaled; blue for lowest and red for highest values. Note that adversarial perturbations are depicted on a gray background and magnified by a factor of 100 for better visualization.

versarial white-box attacks have succeeded as authentic and manipulated regions cannot be separated anymore via the targeted splice localization tools.

Transferability of adversarial examples. Using the adversarial examples that we have generated via white-box attacks, we can also perform adversarial black-box attacks by simply evaluating those examples on the other image-forensics tools. This way we can assess the transferability of the formed adversarial perturbations and study the feasibility of such black-box attacks against EXIF-SC, SpliceRadar, and Noiseprint. Note that the vulnerability to black-box attacks is more concerning from a security perspective as adversaries can impede their targeted system without having direct access to it. The results in Table 1 highlight an asymmetric behavior; we can see that while adversarial perturbations formed on EXIF-SC cause a significant deterioration in the localization performance of SpliceRadar, perturbations generated on SpliceRadar barely make a difference on EXIF-SC compared to its baseline performance. While EXIF-SC appears to be less susceptible to examples

formed by adversarial white-box attacks on other image-forensics methods, the forgery localization performance of Noiseprint noticeably deteriorates under such black-box attacks. Interestingly, considering the results on the Columbia dataset we find that black-box attacks are slightly more successful on Noiseprint than the white-box attacks. Regarding qualitative results, we can see in Figures 2(b) and 3(b) that the adversarial perturbations formed on EXIF-SC limit the splice localization capabilities of SpliceRadar. While the other adversarially perturbed Columbia, DSO-1, and NC16 images shown in Figures 2-4 are not fully transferable to other image-forensics tools, some yield predictions of smaller manipulated regions, e.g., Noiseprint for the adversarial examples formed on EXIF-SC and SpliceRadar presented in Figure 2(b)-(c) (rightmost images).

5. Discussion

As we have seen, adversarial white-box attacks can generally be considered successful on both EXIF-SC and

LOCALIZATION METHOD (results on perturbed images)	COLUMBIA			DSO-1			NC16		
	l_0	l_2	l_∞	l_0	l_2	l_∞	l_0	l_2	l_∞
EXIF-SC	1109644.39	1877.47	19.86	5263130.16	3318.37	18.16	9012808.50	3592.89	15.05
SpliceRadar	568532.84	981.04	12.06	1828703.85	1625.76	9.98	3195024.54	1606.36	9.07
Noiseprint	681162.14	1338.88	22.40	1762103.27	1554.88	24.88	3464998.91	1599.90	15.50

Table 3: PERTURBATION METRICS. This table shows the means for various norms of the adversarial perturbations that we obtained while attacking EXIF-SC, SpliceRadar, and Noiseprint using the manipulated images of the Columbia, DSO-1, and NC16 datasets.

SpliceRadar. However, the results of the white-box attacks on Noiseprint and the asymmetric transferability of adversarial examples require further analysis.

5.1. On the robustness of Noiseprint

Compared to EXIF-SC and SpliceRadar, the results indicate that Noiseprint is more resilient to our adversarial white-box attack, especially on the Columbia dataset. To make sure that the higher resistance is not due to our choice of mimicking a target defined as the mean representation of authentic patches, we have conducted image-based adversarial white-box attacks on the Columbia dataset that also contains authentic counterparts of the spliced images. Namely, instead of manipulating patches in order to move their representations closer to the predefined target, for each spliced image we have identified their authentic counterpart, extracted the feature representation from that image, and used it as our target to mimic with the feature representation extracted from the whole corresponding spliced image. Overall, this image-based white-box attack has failed to outperform our patch-based approach.

We conjecture that the higher robustness of Noiseprint compared to EXIF-SC and SpliceRadar is due to the higher dimensional features extracted from manipulated images and the lack of bottlenecks in the network architecture used for feature extraction. Both bottlenecks and lower dimensional features yield possible collisions – patches or images having the same extracted feature representations – which can be exploited by adversaries. While both EXIF-SC and SpliceRadar work with feature representations of lower dimensionalities – with default settings, 72×72 -pixel patches projected into 100-dimensional and 128×128 -pixel patches mapped into 4096-dimensional feature representations, respectively – Noiseprint yields 2-D noise residual maps with height and width identical to its inputs.

5.2. Asymmetric transferability

To understand the asymmetric transferability properties of adversarial examples, especially between EXIF-SC and SpliceRadar, we have analyzed the perturbations that we formed on the three targeted image-forensics algorithms. The quantitative results are summarized in Table 3. We can see that the adversarial perturbations formed by our approach while attacking EXIF-SC are stronger than those obtained on SpliceRadar or Noiseprint. Adversarial attacks on

EXIF-SC affect the most pixels, as indicated by l_0 norms, and they lead to the highest Euclidean norms as well.

Assuming that stronger perturbations yield higher transferability rates is a reasonable hypothesis. To validate that, we have formed new examples by scaling the formed perturbations; we have doubled the perturbations of examples that we obtained on SpliceRadar and halved those we generated while attacking EXIF-SC. After evaluating these new examples, we have found that the transferability of adversarial examples is not strictly determined by the strength of perturbations in terms of l_0 , l_2 , or l_∞ norms. Considering the results on the subset of NC16 in Table 2, we can see on both SpliceRadar and Noiseprint that stronger perturbations formed on EXIF-SC cause lower deterioration on the splice localization performance than weaker perturbations generated on the other two image-forensics methods.

6. Conclusion

Image forensics can play an important role to prevent the spread of manipulated content on communication channels, such as digital and social media. The detection and/or localization of manipulated regions on forged images have already attracted the research community, and this research area will likely gain more focus in the future thanks to the increasing demand for such tools. Due to recent advances and trends in machine learning, researchers have proposed deep learning-based image-forensics methods that perform well regarding the localization of various manipulations, including splices. However, the applied deep learning models are inherently susceptible to adversarial perturbations making such tools vulnerable as well.

In this paper, we have demonstrated on three image-forensics methods, EXIF-SC, SpliceRadar, and Noiseprint, that splice localization via non end-to-end deep learning-based approaches can be impeded by directly manipulating the internal representations of image patches. Furthermore, we have found that adversarial perturbations can be transferable among the tested image-forensics methods to decrease their capabilities with different effectiveness.

In summary, deep learning-based image-forensics methods that were designed to reveal manipulations, ironically, can be impeded by the application of additional manipulations. To avoid this problem, we need deep learning models that are inherently robust to adversarial perturbations.

References

- [1] Jawadul H Bappy, Amit K Roy-Chowdhury, Jason Bunk, Lakshmanan Nataraj, and BS Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *International Conference on Computer Vision (ICCV)*. IEEE, 2017. 1
- [2] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. On the transferability of adversarial examples against cnn-based image forensics. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019. 2
- [3] Mauro Barni, Matthew C Stamm, and Benedetta Tondi. Adversarial multimedia forensics: Overview and challenges ahead. In *26th European Signal Processing Conference (EU-SIPCO)*. IEEE, 2018. 2
- [4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*. IEEE, 2017. 2
- [5] Yizong Cheng. Mean shift, mode seeking, and clustering. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1995. 2
- [6] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Splicebuster: A new blind image splicing detector. In *International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2015. 2
- [7] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a CNN-based camera model fingerprint. *Transactions on Information Forensics and Security (TIFS)*, 2019. 1, 2, 3, 4
- [8] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Fast geometrically-perturbed adversarial faces. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019. 2
- [9] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *Transactions on Information Forensics and Security (TIFS)*, 2013. 4
- [10] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *Transactions on Information Forensics and Security (TIFS)*, 2012. 2
- [11] Aurobrata Ghosh, Zheng Zhong, Terrance E Boult, and Maneesh Singh. SpliceRadar: A learned method for blind image forensics. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 1, 2, 3, 4, 5
- [12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representation (ICLR)*, 2015. 2
- [13] David Güera, Yu Wang, Luca Bondi, Paolo Bestagini, Stefano Tubaro, and Edward J Delp. A counter-forensic method for CNN-based camera model identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 2
- [14] Y.-F. Hsu and S.-F. Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *International Conference on Multimedia and Expo*, 2006. 4
- [15] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 4, 5
- [16] Nathan Inkawhich, Wei Wen, Yiran Chen, et al. Feature space perturbations yield more transferable adversarial examples. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [17] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representation (ICLR)*, 2017. 2
- [18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representation (ICLR)*, 2017. 2
- [19] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2016. 2
- [20] Alin C Popescu and Hany Farid. Statistical tools for digital forensics. In *International Workshop on Information Hiding*. Springer, 2004. 2
- [21] Andras Rozsa, Manuel Günther, and Terrance E. Boult. Adversarial robustness: Softmax versus Openmax. In *British Machine Vision Conference (BMVC)*, 2017. 3
- [22] Andras Rozsa, Manuel Günther, and Terrance E. Boult. LOTS about attacking deep features. In *International Joint Conference on Biometrics (IJCB)*, 2017. 3
- [23] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J. Fleet. Adversarial manipulation of deep representations. In *International Conference on Learning Representation (ICLR)*, 2016. 3
- [24] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (MFCN). *Journal of Visual Communication and Image Representation*, 2018. 1
- [25] Christian J. Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representation (ICLR)*, 2014. 1, 2
- [26] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1
- [27] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2
- [28] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1