PANCHROMATIC IMAGERY COPY-PASTE LOCALIZATION THROUGH DATA-DRIVEN SENSOR ATTRIBUTION

E. D. Cannas[*], *J. Horváth*[†], *S. Baireddy*[†], *P. Bestagini*[*], *E. J. Delp*[†], *S. Tubaro*[*]

[*]Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano - Milan, Italy
[†]Video and Image Processing Lab (VIPER) - Purdue University - West Lafayette, Indiana, USA

ABSTRACT

Overhead images can be obtained using different acquisition and processing techniques, and they are becoming more and more popular. As with common photographs, they can be forged and manipulated by malicious users. However, not all image forensics methods tailored to normal photos can be successfully applied out of the box to overhead images. In this paper we consider the problem of localizing copy-paste forgeries on panchromatic images acquired with different satellites. We leverage a set of Convolutional Neural Networks (CNNs) that extract traces of the acquisition satellite directly from image patches. We then determine whether an image region appears to have been acquired with a different satellite than the rest of the picture. Results show that the proposed technique outperforms more sophisticated image forensics tools tailoring common photographs.

***Index Terms*—** Overhead forensics, panchromatic images, satellite attribution, forgery localization

1. INTRODUCTION

Digital image manipulation is an easy task for everybody, thanks to the availability of many software suites for image editing together with automatic services and techniques executing this task. As digital images are now more widespread than ever, tools for verifying the integrity of such data have become of paramount importance. This is the objective of the multimedia forensics research community, which has put in a great effort to develop image and video forensic techniques [1, 2].

Thanks to the presence of specialized companies and websites, also overhead imagery (i.e., images of the ground acquired by satellites) is much more common and easy to obtain compared to a few years ago. As with normal photos, this kind of imagery can be easily manipulated with common image editing tools. Therefore, malicious editing of overhead imagery is worrying the forensics community as it may lead to serious consequences [3].

This concern is partly motivated by the fact that the vast majority of forensics methods look for specific editing traces to accomplish the task at hand. However, these traces are often specific to the generation pipeline of a multimedia object, and are different for photographs and overhead images [4]. Therefore, off-the-shelf forensics techniques developed for natural images are not guaranteed to offer optimal performances when applied to overhead imagery.

For this reason, in the last few years the forensics community has started to develop methods specifically tailored to satellite data. In [5], a method based on Generative Adversarial Networks (GANs) and one-class classifiers is proposed to detect image splicing on RGB imagery. The authors of [6] rely on GANs for detecting image forgeries, while the authors of [7] cast forgery detection on RGB images as an anomaly detection problem. Similarly, in [8], deviation from pristine distributions of image pixels are detected using generative autoregressive models. Always regarding RGB imagery forgeries localization with deep learning techniques, the authors of [9] rely on deep belief networks. In [10] and [11], the authors instead exploit nested attention U-Nets and vision transformers, respectively.

When it comes to panchromatic images, the satellite forensics literature has not reached the same level of maturity achieved for RGB overhead imagery forgery localization. To the best of our knowledge, no contributions on the topic have been proposed yet. Still, panchromatic imagery is a vulnerable asset and therefore must be a subject of research for forensics techniques too.

In this paper, we tackle this problem by focusing on copy-paste attacks, i.e., the composition of images coming from different sensors. Given an input probe, we aim to localize which region comes from an alien satellite with respect to the vast majority of the image. This problem has been broadly investigated in the forensics literature for natural images using a variety of forensics traces; for instance, co-occurrences of high-frequency residuals [12, 13], the fusion of PRNU, patch-match and phylogeny features [14], or more sophisticated deep learning feature maps [15].

In our work, we want to localize copy-paste attacks by verifying the consistency of sensor-related features extracted from different patches of the sample under analysis. To do so, we rely on CNNs and ensembling techniques, as these tools proved to be effective in attributing a panchromatic image to the satellite that acquired it [16].

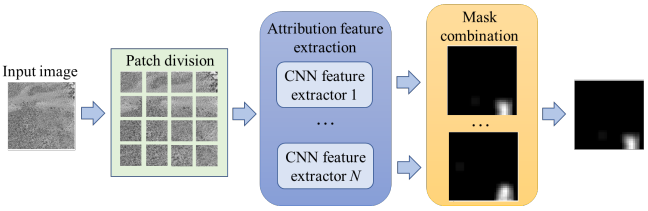


Fig. 1. Panchromatic copy-paste localization pipeline.

This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. 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 AFRL or the U.S. Government. This work was supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program.

Our pipeline, reported in Fig.1, is as simple as it is effective. We process an input image patch by patch, extracting features using an ensemble of CNNs. We then highlight possible copy-paste attacks computing a consistency heatmap that combines the information obtained from the networks. This pipeline is evaluated on a dataset of forged images coming from different satellites. Our performance is then compared to State-Of-the-Art (SOTA) image forensics tools.

2. PROBLEM FORMULATION

We focus on forged panchromatic images generated by copying a region from a source image $\mathbf{S}$ on top of a target region of an image $\mathbf{T}$. After pasting, usually some editing is applied in order to make the source region appear more coherent with the surrounding target area (e.g., applying some smoothing or filtering). In the forensics literature this attack is referred to as copy-paste, and many methods have been proposed to detect it and localize it by exploiting inconsistencies in the camera-related artifacts of the target image [17, 15].

Formally, let us define the coordinates of a pixel in a $U \times V$ resolution panchromatic image as (u, v), where $u \in [1, \dots, U]$ and $v \in [1, \dots, V]$. Let $\mathbf{T}$ and $\mathbf{S}$ be the target and source images, respectively. We define as $\mathcal{S}$ the donor region of $\mathbf{S}$, and as $\mathcal{T}$ the region of $\mathbf{T}$ under copy-paste attack. The resulting tampered image $\hat{\mathbf{T}}$ is defined as

$$\hat{\mathbf{T}}(u,v) = \begin{cases} \mathbf{f}\big(\mathbf{S}(u',v')\big) & \text{if} \quad (u,v) \in \mathcal{T} \\ \mathbf{T}(u,v), & \text{if} \quad (u,v) \notin \mathcal{T} \end{cases}, \tag{1}$$

where (u', v') represents a pixel coordinate in $\mathcal{S}$ and $\mathbf{f}$ is a suitable editing function (e.g., blurring, affine transform, etc.).

The integrity of the image $\hat{\mathbf{T}}$ can be represented by a mask $\mathbf{M}$ the same resolution of $\hat{\mathbf{T}}$, where each pixel takes a binary value of 0 or 1 depending on the pixel being pristine or forged. Formally, $\mathbf{M}$ is defined as

$$\mathbf{M}(u,v) = \begin{cases} 1, & \text{if} \quad (u,v) \in \mathcal{T} \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

The goal of this paper is the localization of the pasted region $\mathcal{T}$ by estimating a tampering mask $\hat{\mathbf{M}}$ as close as possible to $\mathbf{M}$ from the sole analysis of the suspect image $\hat{\mathbf{T}}$.

3. PROPOSED METHOD

A common assumption in the forensics field is that in copy-paste attacks the target region $\mathcal{T}$ presents different characteristic footprints with respect to the rest of the image. This is due to the fact that the source and target images may have been acquired with different devices, and may have undergone different processing operations. In this paper we consider the case of panchromatic copy-paste attacks operated between images coming from different satellites. For this reason, we propose a method that leverages characteristic traces left by different satellites on their acquisitions.

In [16], we have shown the possibility of using CNNs to extract characteristic satellite traces and attribute a panchromatic satellite image to the sensor that generated it. In this work, we leverage the finding of [16] for copy-paste localization. To do so, given an image under analysis $\mathbf{T}$, we follow the pipeline reported in Figure 1. In a nutshell, we split the image under analysis into patches that are fed to multiple CNNs. Each CNN extracts a series of features used to estimate a forgery mask. All masks are then aggregated to obtain the final results. In the following we provide additional details about each step of the method.

CNN-feature extractor. Given an image under analysis $\mathbf{T}$, we split it into a set of P patches $\mathcal{P}_p$, $p \in [1, \dots, P]$. A feature extractor CNN is applied to each patch $\mathcal{P}_p$ to obtain a feature vector $\mathbf{f}_p$.

The feature vectors $\mathbf{f}_p$ must capture sensor-related artifacts useful in discriminating the origin of the different patches. For this reason, we rely on the CNN presented in [16]. This is an EfficientNetB0 [18] trained as a satellite sensor classifier. Considering a training dataset of M known satellites, the output of the network is an M-element vector

$$\mathbf{f} = [f_1,\ f_2, \dots,\ f_M], \tag{3}$$

where the m-th element f_m expresses the likelihood of a panchromatic image patch to be generated from the m-th satellite in the training set. This vector provides a compact yet expressive representation of the information needed for our task.

Preliminary tampering mask generation. After computing the feature vectors $\mathbf{f}_p$ for all patches of the image under analysis, we use this information to estimate the tampering mask $\hat{\mathbf{M}}$. Our previous assumption of target and source images coming from different satellites translates in the assumption that the majority of patches present feature vectors with similar characteristics, whereas those extracted from patches belonging to the tampered region $\mathcal{T}$ show a different behavior.

To represent the global average sensor characteristic of the image, we compute the arithmetic mean of the feature vectors as

$$\hat{\mathbf{f}} = \frac{1}{P} \sum_{p=1}^{P} \mathbf{f}_p. \tag{4}$$

Then, we compute the deviation $\mathbf{m}_p$ of each feature vector from the mean as the $\mathcal{L}_2$ norm of the distance from $\hat{\mathbf{f}}$. Formally,

$$\mathbf{m}_p = \|\hat{\mathbf{f}} - \mathbf{f}_p\|_2, \quad p = 1, \dots, P. \tag{5}$$

These operations provide us with single deviation scores for each vector, which can then be assembled into a tampering mask $\hat{\mathbf{M}}$ by assigning each score to all the pixels of the region from which the corresponding patch $\mathcal{P}_p$ has been extracted.

Model ensembling. We repeat the feature extraction and tampering mask estimation steps multiple times. Specifically, we train N CNNs considering different subsets of the available satellites. In this way, we obtain N different estimates $\hat{\mathbf{M}}_n$ of the tampering mask. With respect to relying on a single CNN model, we observed superior performances and more robustness to changes in the training data distribution adopting this approach.

All masks are then fused together relying on the Variance-to-Entropy-Ratio (VER) metric we developed in [19]. This measure accounts for the fact that, in case a tampered area has been localized in the image sample, a good tampering mask should present high variance of the pixel values, due to the diversity of values for the presence of the tampered region, and low entropy in the pixel coordinates, since the tampered region should be well localized in the mask.

The VER metric merges these conditions into a single score as

$$\Lambda(\hat{\mathbf{M}}_n) = \frac{\mathrm{Var}(\hat{\mathbf{M}}_n)}{\mathrm{Entropy}(\hat{\mathbf{M}}_n)}, \quad n = 1, \dots, N, \tag{6}$$

where Var and Entropy respectively compute the variance of the mask values and the entropy of the pixel coordinates, considering only pixels whose values are above a threshold.

We can then compute the final tampering mask $\hat{\mathbf{M}}$ by averaging the candidate masks after weighting them element-wise by their corresponding VER score. Formally

$$\hat{\mathbf{M}} = \frac{\sum_{n=1}^{N} \Lambda(\hat{\mathbf{M}}_n) \cdot \hat{\mathbf{M}}_n}{\sum_{n=1}^{N} \Lambda(\hat{\mathbf{M}}_n)}, \tag{7}$$

where all products are operated element-wise.

4. EXPERIMENTAL SETUP

Dataset. Following the approach in [16], we collected 8-bit panchromatic imagery from the DigitalGlobe portal [20]. We have been careful to avoid any possible semantic bias in the collected data. To this end, the downloaded imagery comes from different geographical areas: barren, field, forest, snow and urban regions. To respect our basic assumption, the imagery comes from 5 different source satellites: GeoEye (GE01), QuickBird (QB02), WorldView 1, 2 and 3 (WV01-02-03). All the data has been provided in the GeoTIFF format, with no compression, as non-overlapping orthorectified tiles of 16384×16384 pixels.

Due to the pixel resolution of each tile, too large to be processed by our CNN feature extractors, we cropped them to obtain non-overlapping 1024×1024 pixel images. This operation led us to a dataset of more than 100'000 samples, 20'000 for each satellite class, balanced among all geographical regions. We have then split them into a train and test set with a 50%-50% policy. The first set is used exclusively to train the CNN feature extractors, whereas the images in the latter set have been further elaborated (see below) to create credible copy-paste attacks. To ensure that we avoid any training-test contamination, cropped regions from the same tile all belong to either the training or test set.

To test our copy-paste localization method, starting from the 50'000 images of the test set we have created a number of tampered images by taking random regions $\mathcal{S}$ from source images and pasting them on random positions of target images. Source and target images have been taken from different satellites, and all copy-paste attacks have been executed coherently from a semantic point of view (i.e., only barren on barren, forest on forest, field on field, etc., copy-paste attacks have been realized). Figure 2 reports some test images.

We have also taken into account the possibility of applying editing operations on the tampered area $\mathcal{T}$ to make the attacks more plausible. In particular, we have created two different datasets. The first one is a preliminary test set, where we considered just 3 editing operations: Gaussian blurring, a simple random rotation and resize, and finally a single resize operation. This dataset counts 50 images per satellite, geographic region and per editing operation, with the resolution of $\mathcal{T}$ fixed to 256×256 pixels, for a total of 3750 samples. We relied on this dataset for comparing against two methods of the SOTA in a more controlled scenario.

The second dataset instead was designed with the goal of gauging our method on a wider variety of editing operations and resolutions of $\mathcal{T}$. For this reason, it comprehends 4 different types of blurring (i.e., Gaussian blur, motion blur, median blur, and average blur), 3 types of affine transforms (i.e., random rotation and resize, piece-wise affine transform and a prospective transform) and two different contrast enhancements (i.e., logarithmic and sigmoid contrast). Moreover, we have accounted for different 5 resolutions of $\mathcal{T}$ in pixels: 128×128, 160×160, 192×192, 224×224 and 256×256, equally distributed among all samples in number. For each one of the satellites, each one of the geographical areas, and each one of the operation, we have realized 100 copy-paste attacks, ending up with a tampered dataset of 22'500 samples.

(a) Forged image 1. (b) Tampering mask 1. (c) VER mask $\hat{\mathbf{M}}$ 1.

(d) Forged image 2. (e) Tampering mask 2. (f) VER mask $\hat{\mathbf{M}}$ 2.

Fig. 2. Copy-paste samples from the test set with ground truth and estimated masks. No editing applied, tampering size is 256×256.

Training. Each CNN feature extractor acts as a M-class classifier. Having 5 satellites in the training set, we considered training each CNN with $M = 4$, therefore leaving one satellite out of the training set at a time. This allowed us to test our method in the more difficult scenario where either the source or target image sensors are not known by one of the extractors. We ended up with $N = 5$ different CNN extractors that were jointly used for mask aggregation.

From the 50'000 samples of the training set, we proceeded in further dividing them into training-validation splits following a 50%-50% policy. As pre-trained models normally used in computer vision tasks work on natural images, we trained our models from scratch in a two-step fashion. Indeed, we initially pre-trained them using the samples at full 1024×1024 resolution, and then finetuned them on 256×256 patches extracted from the 4 corners of full resolution 1024×1024 samples. This procedure has been devised following an ablation study on the performances of the EfficientNetB0 in classifying small resolution patches $\mathcal{P}_p$, where we have seen that using a network trained on full resolution images or directly on small resolution patches brought worse and less consistent performances with respect to a pre-training and finetuning pipeline.

For both the pre-training and finetuning stages, we optimized the networks for 200 epochs using Adam [21] and a simple cross-entropy loss, with batches of 5 images in pre-training and of 40 patches in finetuning. In both operations, we started from a learning rate of 0.001 decreased by a factor 10 in case of plateau of the loss function after 10 consecutive epochs, and early stopping the training in case the learning rate reached a minimum of 10^{-8} or if the validation loss did not improve after 50 consecutive epochs.

Evaluation metrics. To quantify the performances of the proposed solution, we compared the estimated masks $\hat{\mathbf{M}}$ against ground-truth tampering masks $\mathbf{M}$. Since estimated masks provide a soft score for the pixel values, to evaluate the discriminating capabilities of our method in separating pristine from forged pixels we computed Receiver-Operating-Characteristic (ROC) curves and their respective Area-Under-the-Curve (AUC) values setting variable thresholds. Thresholds are used to evaluate the entire dataset and are not optimized per-image. All experiments have been executed on a Intel Xeon Gold 6246 CPU and a single NVIDIA Titan RTX GPU.

Table 1. Preliminary test set against baselines. Best result in bold.

Method	AUC
Noiseprint	0.946
Splicebuster	0.969
VER mask	**0.971**

5. RESULTS

Comparison against baselines. As introduced in Section 4, we compared our proposed pipeline against two SOTA methods on the preliminary test set, extracting patches for our method with a 32×32 stride. As well-known forensics baselines, we chose Splicebuster [12], since it looks for high-frequency traces, and Noiseprint [15], which has been applied with success to overhead RGB images. We relied on the pre-trained models of Noiseprint, generating tampering masks for this method using the same blind splicing localization algorithm of Splicebuster.

As we can see from Table 1, our technique shows really close localization performances in terms of AUC to SOTA methods. These results nevertheless are encouraging considering that our CNN feature extractors have been trained only on pristine images and for a different forensic task. Without seeing any kind of editing operation, nor copy-paste images, they are still able to extract meaningful satellite attribution features from the edited tampered regions. Indeed, if we look at localization results divided by regions and satellites in Table 2, our pipeline shows better results 7 times out of 10.

Moreover, from a performance point of view our method has the advantage of being faster at inference time. Figure 3 shows execution times against sample resolution, We can see that our pipeline, while still analyzing each sample patch by patch, performs faster at any image resolution than the two SOTA techniques considered, with a performance gap increasing with the sample size for Splicebuster in particular. We argue that this performance gap is motivated by our tampering mask estimation method requiring far less computations. Note also that the numbers reported have been computed without parallelizing the feature extraction of the ensemble, operation that can be easily done in a multi-GPU system.

Robustness to processing. To provide additional evaluation details about our method, we tested it against the second test set described in Section 4. Figure 4a shows ROC curves obtained when different editing operations are applied on $\mathcal{T}$ in the test set. As we can see, our method reaches satisfactory AUC values above 0.84 for all operations with very little variations across the types of editing. These

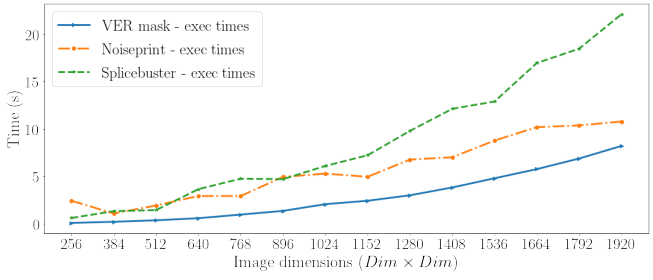
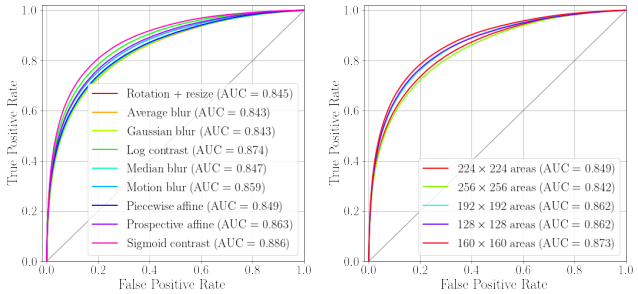


Fig. 3. Execution times comparison. Times account for the complete generation of $\hat{\mathbf{M}}$.

Table 2. Preliminary test set results divided per satellite and region. Best results per scenario in bold.

Region	Method	AUC	Satellite	Method	AUC
Barren	Noiseprint	0.949	*GE01*	Noiseprint	0,939
	Splicebuster	0.976		Splicebuster	0,969
	VER mask	**0.980**		VER mask	**0,991**
Field	Noiseprint	0.961	*QB02*	Noiseprint	0,952
	Splicebuster	0.976		Splicebuster	0,970
	VER mask	**0.977**		VER mask	**0,978**
Forest	Noiseprint	0.960	*WV01*	Noiseprint	0,948
	Splicebuster	0.974		Splicebuster	0,965
	VER mask	**0.978**		VER mask	**0,981**
Snow	Noiseprint	0.930	*WV02*	Noiseprint	0,944
	Splicebuster	0.950		Splicebuster	**0,964**
	VER mask	**0.952**		VER mask	0,927
Urban	Noiseprint	0.928	*WV03*	Noiseprint	0,946
	Splicebuster	**0.970**		Splicebuster	**0,978**
	VER mask	0.963		VER mask	0,974



(a) Results per editing, all tampered areas considered.

(b) Results per tampered area, all editing operations considered.

Fig. 4. Results for the final test set of 22'500 samples.

numbers therefore suggest a robustness of our solution to modifications in general, which again is surprising as our pipeline is agnostic with respect to editing operations.

Robustness to resolution. To further evaluate our method, we provide the results break-down considering different resolutions of $\mathcal{T}$ on the entire test set. As we can see in Figure 4b, with a minimum AUC of 0.84 our method performs fairly good and consistently also with different resolutions of the tampered areas.

6. CONCLUSIONS

In this work, we investigated the problem of copy-paste attacks localization in the context of panchromatic overhead imagery. We verified the hypothesis that copy-paste attacks can be localized by looking at satellite attribution features, exploiting CNNs, model ensembling and the VER metric for feature aggregation.

Results on the gathered dataset show how our technique is competitive with the SOTA while being considerably faster, robust to various editing operations and not requiring a training phase on copy-paste samples. These elements pave the road to studies on the generalization capabilities of our solution to scenarios where both the target and source image sensors are unknown, for instance integrating Open-Set-Recognition [22] and uncertainty analysis [23, 24] tools.

7. REFERENCES

[1] A. Piva, "An overview on image forensics," *ISRN Signal Processing*, vol. 2013, pp. 22, 2013. 1

[2] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, pp. e2, 2012. 1

[3] BBC News, *Conspiracy Files: Who shot down MH17?*, April 2016 (accessed January 12, 2020), http://www.bbc.com/news/magazine-35706048. 1

[4] PennState University, *Optical Sensors overview*, November 2020 (accessed November 20, 2020), *https://www.e-education.psu.edu/geog480/node/444*. 1

[5] S. K. Yarlagadda, D. Güera, P. Bestagini, F. M. Zhu, S. Tubaro, and E. J. Delp, "Satellite image forgery detection and localization using gan and one-class classifier," in *IS&T Electronic Imaging (EI)*, 2018. 1

[6] E. R. Bartusiak, S. K. Yarlagadda, D. Güera, P. Bestagini, S. Tubaro, F. M. Zhu, and E. J. Delp, "Splicing detection and localization in satellite imagery using conditional GANs," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019. 1

[7] J. Horvath, D. Güera, S. K. Yarlagadda, P. Bestagini, F. M. Zhu, S. Tubaro, and E. J. Delp, "Anomaly-based manipulation detection in satellite images," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 1

[8] D. Mas Montserrat, J. Horvath, S. K. Yarlagadda, F. M. Zhu, and E. J. Delp, "Generative autoregressive ensembles for satellite imagery manipulation detection," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020. 1

[9] Janos Horvath, Daniel Mas Montserrat, Hanxiang Hao, and Edward J. Delp, "Manipulation detection in satellite images using deep belief networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1

[10] Janos Horvath, Daniel Mas Montserrat, and Edward J. Delp, "Nested attention u-net: A splicing detection method for satellite images," in *Pattern Recognition. ICPR International Workshops and Challenges*, 2021. 1

[11] Janos Horvath, Sriram Baireddy, Hanxiang Hao, Daniel Mas Montserrat, and Edward J. Delp, "Manipulation detection in satellite images using vision transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 1

[12] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2015. 1, 4

[13] Davide Cozzolino and Luisa Verdoliva, "Single-image splicing localization through autoencoder-based anomaly detection," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2016. 1

[14] Lorenzo Gaborini, Paolo Bestagini, Simone Milani, Marco Tagliasacchi, and Stefano Tubaro, "Multi-clue image tampering localization," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014. 1

[15] D. Cozzolino and L. Verdoliva, "Noiseprint: A CNN-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2020. 1, 2, 4

[16] E. D. Cannas, S. Baireddy, E. R. Bartusiak, S. K. Yarlagadda, D. Mas Montserrat, P. Bestagini, S. Tubaro, and E. J. Delp, "Open-set source attribution for panchromatic satellite imagery," in *IEEE International Conference on Image Processing (ICIP)*, 2021. 1, 2, 3

[17] Luca Bondi, Silvia Lameri, David Güera, Paolo Bestagini, Edward J. Delp, and Stefano Tubaro, "Tampering detection and localization through clustering of camera-based cnn features," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 2

[18] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019. 2

[19] Sebastiano Verde, Edoardo Daniele Cannas, Paolo Bestagini, Simone Milani, Giancarlo Calvagno, and Stefano Tubaro, "Focal: A forgery localization framework based on video coding self-consistency," *IEEE Open Journal of Signal Processing*, 2021. 2

[20] Maxar Technologies, *DigitalGlobe discover portal*, accessed January 12, 2021, *https://discover.digitalglobe.com/*. 3

[21] D.P. Kingma and J. Ba, "Adam: a method for stochastic optimization. arxiv: 14126980," 2014. 3

[22] Abhijit Bendale and Terrance Boult, "Towards open set deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[23] Gal Y. and Ghahramani Z., "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML)*, 2016. 4

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *International Conference on Neural Information Processing Systems (NIPS)*, 2017. 4