MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation

Haiying Guan[1], Mark Kozak[2], Eric Robertson[2], Yooyoung Lee[1], Amy N. Yates[1], Andrew Delgado[1], Daniel Zhou[1], Timothee Kheyrkhah[1], Jeff Smith[3], Jonathan Fiscus[1]
[1]National Institute of Standards and Technology, [2]PAR Government, [3]University of Colorado Denver
haiying.guan@nist.gov, {mark_kozak, eric_robertson}@partech.com, {yooyoung.lee, amy.yates, andrew.delgado, daniel.zhou, timothee.kheyrkhah}@nist.gov, jeff.smith@ucdenver.edu, jonathan.fiscus@nist.gov

Abstract

We provide a benchmark for digital Media Forensics Challenge (MFC) evaluations. Our comprehensive data comprises over 176,000 high provenance (HP) images and 11,000 HP videos; more than 100,000 manipulated images and 4,000 manipulated videos; 35 million internet images and 300,000 video clips. We have designed and generated a series of development, evaluation, and challenge datasets, and used them to assess the progress and thoroughly analyze the performance of diverse systems on a variety of media forensics tasks in the past two years.

In this paper, we first introduce the objectives, challenges, and approaches to building media forensics evaluation datasets. We then discuss our approaches to forensic dataset collection, annotation, and manipulation, and present the design and infrastructure to effectively and efficiently build the evaluation datasets to support various evaluation tasks. Given a specified query, we build an infrastructure that selects the customized evaluation subsets for the targeted analysis report. Finally, we demonstrate the evaluation results in the past evaluations.

1. Introduction

The explosion of media storage, transmission, editing, and sharing tools has the potential to foster an increase in tampered data and challenge the traditional trust in visual media. The creation, modification, and distribution of digital media are extremely simple to use for even inexperienced users and require only minimal effort. In addition, with developments in the advent of new techniques such as Generative Adversarial Networks (GANs, "deepfakes") [1] and Computer Generated Imagery (CGI) [2], it becomes increasingly challenging for existing media forensic technologies[3]-[10]to identify the integrity, trustworthiness, and authenticity of visual content.

In order to facilitate the development of media forensics research, we started work on the Media Forensics Challenge (MFC) Evaluation for the DARPA MediFor program [11] in 2015. We design, collect, annotate, and assemble a series of comprehensive digital forensic databases for the evaluation of media forensic technologies. The benchmark data contains four major parts: (1) 35 million images and 300,000 video clips downloaded from the internet with their characteristics and labels; (2) up to 176,000 pristine, self-collected, high provenance (HP) images and 11,000 HP videos; (3) approximately 100,000 manipulated images covering over 100 manipulation types produced by professional manipulators from approximately 5,000 image manipulation journals with manipulation history graphs and annotation details; 4,000 manipulated videos and over 500 video manipulation journals; (4) a series of evaluation datasets with reference ground-truth to support several challenge tasks in media forensics challenge evaluations from the last two years.

In this paper, we first survey existing datasets and discuss the special characteristics and challenges of media forensic data collection. Then based on the diversity and complexity of media forensic applications, we propose the following tasks [12][13]: Manipulation Detection and Localization: to detect if an image/video has been manipulated (MD), and if so, where it is manipulated (MDL); Splice Detection and Localization: to detect if a region of a given potential donor image has been spliced into a probe image (SD); if so, where are the regions in both images (SDL); Provenance Filtering (PF) and Provenance Graph Building (PGB): to reconstruct an image's phylogeny graph given a 'world' image pool; Camera Verification (CV): to verify if an image/video probe is from a claimed camera sensor; and Event Verification (EV): to verify if an image is from a claimed event. Detection systems are measured by the Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC), localization systems are measured by the Matthews Correlation Coefficient (MCC), PF systems are measured by Recall, and PGB systems are measured by the Node Link Overlap Similarity Metric (SimNLO). Afterwards, we present the approach to building evaluation datasets for different evaluation tasks and demonstrate the results from the last two years' evaluations.

The major contributions are: (1) we propose and demonstrate a methodology for MFC evaluation data design; (2) we build large-scale media forensics evaluation benchmark datasets for quantitative media forensics system performance evaluations. The evaluation data can be directly used for existing task evaluations. The world data, HP data, and manipulation data can be easily customized to other evaluation tasks; (3) we provide a description about the data, metadata, training data, and ground-truth evaluation reference data of our datasets, which includes but is not limited to the capture camera/device's model and identity, the step-by-step manipulation operations with parameters, the manipulation history graph (i.e. journal), the localized manipulated region for every step, the semantic category, the model and parameters of the special filters (e.g. GAN, jpeg anti-forensic filter, social media laundering etc.); (4) we present the state-of-the-art media forensics system evaluation results and their progress over the past two years.

2. Related Work[1]

Many general benchmark media datasets are available in literature: UCID (2003) [14] and ImageCLEF (2004-2013) [15] for image retrieval; Caltech-256 (2007) [16], 80 million tiny images (2008) [17], Mammal Image (2008) [18], PASCAL (2008-2015) [19], ImageNet (2009) [20], SUN (2010) [21], Kitti (2013) [22], and Microsoft COCO (2014) [23] for object and scene detection, segmentation, and recognition; TRECVid (2000-2006) [24], Event recognition video dataset (2011) [25], YFCC100M (2015-2016) [26], MediaEval (2013-2017) [27] for multimedia research. Unfortunately, none of them could be directly used for media forensics evaluation purposes due to lack of manipulated media or sufficient annotation. Given a media from any of the above datasets, we would not know if it was manipulated. Furthermore, it is impractical to do post-annotation for manipulation metadata.

Recently, several datasets were collected specifically for media forensics research. The EU REWIND (REVerse engineering of audio-VIsual coNtent Data) (2011-2013) [28] digital forensics project focuses on digital watermarking, passive image authentication, and capture source identification (camera PRNU etc.). The 'Realistic' dataset contains 69 manually manipulated fake images and 69 original images. The 'Synthetic' dataset contains 4800 automatically manipulated images. 200 images from a Nikon D60 camera have been released. The dataset is small and only small portions of it are available to public.

The First Image Forensics Challenge (2013) [29], an international competition organized by the IEEE Information Forensics and Security Technical Committee (IFS-TC), collected thousands of images of various scenes, both indoors and outdoors with 25 digital cameras. We use this dataset as our first reference, and expand the design's breadth and depth in several aspects: scale, manipulation types and graph, annotations, PRNU training data etc.

Some forensic databases target particular manipulation operations. The Columbia automatically-spliced image database (2004) [30] has two parts: a grayscale image dataset with 933 authentic and 912 spliced 128×128 pixel-grayscale image blocks, extracted from images in CalPhotos [31], and a color image dataset with 183 authentic uncompressed color block images and 180 spliced uncompressed color block images. CASIA's Image Tampering Detection Evaluation Database (2013) [32] focuses on splicing and tampering. CASIA v1.0 has 800 authentic and 921 spliced 384×256 images. CASIA v2.0 contains 7,491 authentic and 5,123 tampered images. The CoMoFoD dataset (2013) [33] designed for copy-move forgery detection consists of 260 forged image sets. Each set includes a forged image, two masks and its original image. The manipulations include translation, rotation, scaling, distortion, and postprocessing such as JPEG compression, blurring, noise adding, color reduction etc. The MICC F220, MICC F2000 (2011) [34] and FAU-Erlangen Image Manipulation Datasets (2012) [35] are other copy-move datasets. The Rebroadcast dataset (2018) [36] contains 14,500 large diverse rebroadcast images captured by screen-grabs from 234 displays, scanning printed photos using 173 scanners, or re-photographing displayed or printed photos with 282 printers and 180 recapture cameras. UMDFaces (2016) [37] has 367,888 annotated faces of 8,277 subjects. About 115,000 images have their key point annotations verified by humans. The UMD face swap dataset contains tampered faces created by swapping one face with another using multiple face swapping apps. The VISION dataset (2017) [38] contains 11,732 native images, which are then shared through Facebook and WhatsApp resulting in a total of 34,427 images, and 648 native videos, which are shared through YouTube and WhatsApp, resulting in a total of 1,914 videos. The FaceForensics dataset (2018) [39] has about a half million edited images (from over 1000 videos at various quality levels) using a state-of-the-art face editing approach, and annotated with classification and segmentation references. Those datasets are limited to only single or several manipulation types like splicing, social app, copy-move, rebroadcast, or face manipulations etc.

Other datasets are collected for other purposes. Break Our Steganographic System (BOSS) (2011) [40] is for steganalysis challenge evaluation. Two datasets are created: BOSSBase and BOSSRank, BOSSBase was composed of 9,074 never-compressed cover images coming from 7 different cameras, and created from full-resolution color images in RAW format. The BOSSRank has 1,000 512×512 grayscale images. Several datasets are designed for source identification, that is, to verify the trust and authenticity of data and the devices that create it. Purdue Sensor and Printer Forensics (PSAPF) Dataset (2008) [41] provides an overview of current characterization techniques for 5 scanners and 21 printers. Goljan et al. (2009) [42] presented one million images collected from Flickr, spanning 6,896 individual cameras covering 150 commerce models. The Dresden image database (2010) [43] contains 14,000 high resolution images from 73 digital cameras covering 25 camera models. The images are collected from different scenes with two additional sets of auxiliary images for special use in camera Photo Response Non-Uniformity (PRNU) studies. RAISE (RAw ImageS datasEt) (2015) [44] is a collection of 8,156 high-resolution raw images using three Nikon devices. The images are taken at very high resolution and saved in an uncompressed format. The images cover a wide variety of semantic content, subjects, scenarios, and technical parameters, and are properly

[1] Any mention of commercial products or reference to commercial organizations in this report is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

annotated with 7 category labels. These datasets are valuable for our Camera Verification task only.

Despite the availability of the existing datasets, there is a need for a sufficiently large, representative benchmark dataset containing a wide range of realistic media manipulations with detailed phylogeny graphs and ground-truth references for media forensic evaluation to push the state-of-the-art forward.

3. Media Forensic Dataset Design

Our evaluation, besides answering the general question: "How well do the state-of-the-art forensic systems perform?", aims to answer deeper questions such as: (1) What major factors affect the performances? (2) Accounting for the diversity and specialization of media forensics technologies, which systems are suitable for which situations? To answer such questions, it is insufficient merely to have ground-truth about whether the media is manipulated. The data, metadata, manipulation data, and reference data are crucial for the evaluation. The row headers of Table 1 denote data/metadata availability; the row index labels show the tasks and evaluation report availability given the data/metadata references. It shows that more tasks could be evaluated if more reference resources are available. The quality and quantity of the analysis reports depends on the availability of metadata. With well-structured data, metadata, and reference data, we can provide detailed reports and comparisons using factor analysis, which can help us better understand system performance, promote system development, and provide directions for future development and evaluation.

Table 1: Evaluation capability vs. data availability.

Report contents	Task	Manipulated ?	Probe Ref. Mask	Donor Ref. Mask	Meta data	PRNU Train Data	Manip History Graph
Det. Scores (ROC, AUC) on full dataset	**MD**	✓	✗	✗	✗	✗	✗
Det. Scores; Localization score (MCC) on full dataset	**MDL**	✓	✓	✗	✗	✗	✗
For both probe and donor: Det. Scores; Loc. score on full dataset only	**MDL** **SDL**	✓	✓	✓	✗	✗	✗
For both probe & donor, Det. & Loc. score; Factor analysis on both full dataset and subsets by selective scoring using defined metadata queries	**MDL** **SDL**	✓	✓	✓	✓	✗	✗
For probe & donor, Det. & Loc. score; Factor analysis on both full dataset and subsets by selective scoring using defined metadata queries	**MDL** **SDL** **CID**	✓	✓	✓	✓	✓	✗
For probe & donor, Det. & Loc. score; Factor analysis on both full dataset and subsets by selective scoring using defined metadata queries; PF Recall; PGB SimNLO;	**MDL** **SDL** **CID** **PF,** **PGB**	✓	✓	✓	✓	✓	✓

support more tasks and more detailed evaluation report

3.1. Data Collection Challenges

Besides the general challenges of computer vision dataset collection, the major challenges for media forensics datasets are: (1) Highly diverse research topics involving multidisciplinary areas: multimedia security, computer forensics, image processing, computer vision, imaging, and signal processing. Technologies include but are not limited to JPEG artifact detection, crop/contrast/clone/splice detection, lighting/shadow/reflection consistency, physical/semantic consistency, Electrical Network Frequency (ENF), PRNU, and audio/video person ID consistency. The evaluation of different types of media forensics systems requires different types of meta, training, testing, and reference data. In addition, different systems may only work with particular constraints. For example, face systems work only on a face region, clone detectors only work well on cloning, which increases the complexity of the evaluation metadata and infrastructure. (2) Intrinsically high dimensionality: besides image/video dimensions and their metadata (EXIF, camera ID etc.), we noticed that the manipulation operations and history of a media are very important for media forensics research. The history generates additional factors which affect system performance. In addition, the manipulation was often done with a purpose and the semantic meaning of the manipulation presents yet more factors. (3) High complexity and cost for structured data and metadata collection, manipulation, and annotation. (4) "Curse of dimensionality" issues: to better understand the system performance, and to do factor analysis (apple-to-apple) comparisons, we need systematically structured orthogonal fractional factorial data. The dataset size increases exponentially as factors increase. (5) The post-annotation approach, which is used for biometric or video analytics evaluation ground-truth data generation, does not work well in the media forensics domain. Given a media, even forensics experts find it difficult to deduce whether it is manipulated, let alone compose a step-by-step description of the manipulation operations and their corresponding manipulated regions used for evaluation.

3.2. Dataset Design and Solutions

Given the challenges listed above, we propose the following approaches and solutions:

(i) A set of sufficiently large and publicly available datasets are collected, annotated, and manipulated. We hired professional manipulators using various media editing software and tools to produce manipulated media suited to real-world applications.

(ii) We proposed a manipulation history graph representation to capture the structured manipulation data, operations, metadata, reference mask etc.

(iii) We designed and developed the manipulation Journaling Tool (JT) [45], a software application that assists us in generating and annotating the data, metadata, reference data, and reference ground-truth mask collection effectively and efficiently. The JT integrates the functions such as semantic and metadata collection, automatic and human annotations, manipulation reference mask generation, manipulation operation data collection, and automatic/semi-automatic manipulation etc. The JT has both online and offline functions to support data collection, generation, annotation and verification based on the collection requirements.

(iv) We have developed an automatic journaling tool (AutoJT), which generates manipulated media and its accompanying journal from non-manipulated media.

(v) We have also developed an extended journaling tool (ExtendedJT), which extends and journals partial or complete manipulated media from existing journals, for factor analysis. It automatically generates a large number of manipulated images/videos given a manipulation operation filter and their parameters.
(vi) We designed the data collection and evaluation infrastructure which shares the data among different evaluation tasks to reduce the data collection cost.
(vii) Because different systems may work on different test sets, in addition to testing systems on all testing data, we also designed and developed a selective scoring evaluation infrastructure to dynamically extract a specified subset from the whole test set for selective evaluation.

4. MFC Data Collection

4.1. Images and Videos

World Data: Publicly available imagery acquired off the internet is referred to as World Data. To date the corpus contains 35 million images and 300,000 videos of World Data. It is anticipated that the program will have downloaded 45 million images and 450,000 video clips from the internet. The initial collection of World Data was random, to be used as clutter in evaluations. After the initial collection, we focused on maximizing the diversity of the camera models. Over 500 distinct camera models are represented.

HP Data: At the time this paper was written a total of 176,000 High Provenance (HP) images and 11,000 HP videos were collected by our team. An additional 35,000 photographs and 3,500 videos have been planned to be collected. HP data is always collected on physical devices which team members have physical access to, where all relevant device-intrinsic parameters are known and recorded. Using HP data ensures that all manipulations occur on images with no previous manipulations and avoids copyright infringement. All HP data collected is released under the Creative Commons 0 (CC0) license. CC0 effectively releases the images into the public domain.

PRNU Training Data: Several hundred cameras have been enrolled. These largely consist of moderate to high end Digital Single Lens Reflex (DSLR) and mirrorless cameras. To date the camera database has 574 distinct HP cameras enrolled in the database. Enrolling an HP camera in the database is a multistep process. It begins with the collection of Photo Response Non-Uniformity (PRNU) training data. Image PRNU data is collected with a perfectly diffuse practical light source to evenly illuminate the sensor. Ideally images with pixel saturation at 80% and 50%, and then with a lens cap on are collected at each resolution the camera is capable of. After PRNU data collection the user completes a camera enrollment form to record the owner of the camera, an inventory control ID, make, model, and several other metadata fields. Mobile devices with front and rear facing cameras are further identified by their primary camera (non-selfie) and the secondary (selfie) camera. Most cameras have several hundred, sometimes reaching over 2,000 images per camera.

To facilitate the discovery and management of the media, the “MediFor Browser”, a PostgreSQL database and web-based interface, has been developed.

4.2. Manipulation Data and Annotations

Manipulated media is accompanied by a journal of the sequential manipulations executed to produce the media. We describe each manipulation with a software agnostic description called an operation, generalizing the behavior of the manipulation. Along with the operation name, each manipulation description is accompanied by the software name and version used to execute the manipulation, generalized parameters providing details of the operation, and semantic annotations describing the purpose of the manipulation in the context of a group of manipulations.

4.3. Journaling Tool (JT) for Manipulation Recording

The JT [45] collects and administers a sequential record of manipulations applied to non-manipulated media to produce one or more complete manipulated products (Figure 1 (a)). Given the detailed record, the JT provides a set of operational checks for accuracy and unintended side effects. The checks ensure each manipulation is discrete. For example, a common mistake in cropping images is to realign pixels, thus polluting the crop operation with interpolation and resizing.

The journal includes non-manipulated base media, manipulated final media, media captured after each discrete manipulation and the data capturing relevant changes in the media for each manipulation. The last kind of data includes metadata and per-pixel indicators in the form of manipulation masks. Metadata is often a relevant indication of manipulation, such as GPS coordinates, time of day and camera-model adjustments. Retaining all relevant media serves journal extension, in which each step of the manipulation sequence becomes a launching point for another set of manipulations.

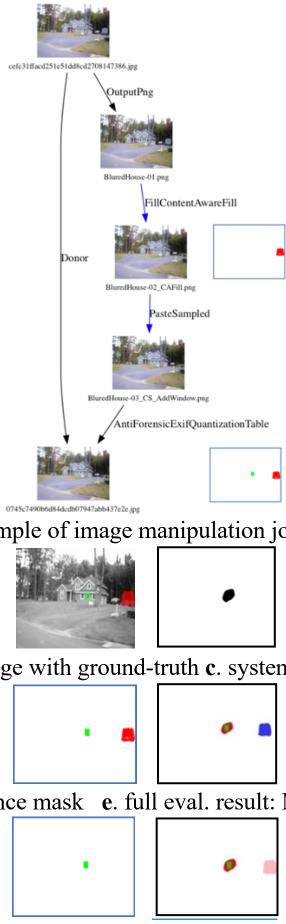
5. MFC Evaluation Tasks

Due to the nature and diversity of media forensics, one evaluation task cannot cover all applications. We designed 5 evaluation tasks (MDL image and video, SDL, EV, PF, and PGB), and 2 challenges (CV and GAN).

5.1. Image Manipulation Detection and Localization

The image Manipulation Detection and Localization (MDL) task is to detect if an image has been manipulated, and if so, to spatially localize the region determined to be manipulated. The reference ground-truth mask for localization is a JPEG 2000 image; each bit plane indicates the manipulated region(s) of a distinct manipulation operation step. Localizable manipulations (e.g., clone) have corresponding mask regions while global manipulations (e.g., blur) affecting the entire image do not. We also collect the following metadata: the camera information (camera ID, PRNU training data etc.); the image metadata (EXIF header information, captions, face(s) in image and their locations etc.); and the manipulation information (whether the image is manipulated, the major manipulation operations and their filters, parameters, and their orders, the corresponding masks for each operation step etc.). With rich metadata and manipulation information, we are able to provide a detailed evaluation report.

Figure 1 (a) shows a simple example of an image manipulation journal graph for localization evaluation. The top image in Figure 1 (a) is the nonmanipulated base image. There are two major manipulation steps: the manipulator removes the truck near the car using the content aware fill operation; and the manipulator clones a first-floor window.


cefc31ffacd251e51dd8cd2708147386.jpg
OutputPng
BluredHouse-01.png
FillContentAwareFill
Donor
BluredHouse-02_CAFill.png
PasteSampled
BluredHouse-03_CS_AddWindow.png
AntiForensicExifQuantizationTable
0745c7490b6d84dcdb07947abb437e2e.jpg

a. An example of image manipulation journal

b. test image with ground-truth **c**. system output results

d. full reference mask **e**. full eval. result: MCC = 0.196

f. selective scoring ref. **g**. selective eval.: MCC = 0.521

Figure 1: Image manipulation localization evaluation.

Two manipulated images could be used as testing images: the intermediate image after the truck is removed with only one major manipulation: content aware fill, and the final manipulated image with the truck removed (red region) and window cloned (green region). Given the final manipulated image as a test image, there are two types of evaluations. The first evaluation is on all manipulation operations regardless of manipulation type (Figure 1 (d) and (e)). The MCC of the system output mask shown in Figure 1 (c) is 0.196. The second evaluation is called selective localization (Figure 1 (f) and (g)); that is, we evaluate the system's performance on a single or a subset of manipulation operations only. If we evaluate the system performance only on the paste sample clone, then only the green region is used as the reference mask. Given the same system output as Figure 1 (c), the selective scoring MCC on paste sample operation is 0.521.

5.2. Video Detection and Temporal Localization

The video manipulation detection task is to detect if a video has been manipulated. Video temporal localization is to temporally localize the edit frame(s). The reference for temporal localization is the intervals where the frames were manipulated. Figure 2 shows the frames in a timeline, the blue part denotes the original frames, the green part denotes the manipulated frames. Given the ground-truth reference interval and the system detection result intervals, the evaluation scores can be reported given evaluation metrics.

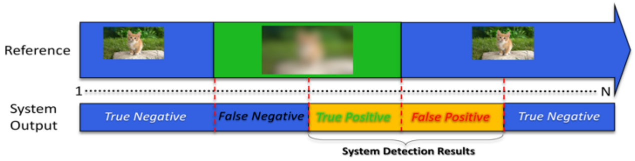


Figure 2: Video temporal localization.

5.3. Splice Manipulation Detection and Localization

Splice Detection (SD) and Localization (SDL) is to detect if a region of a given potential donor image has been spliced into a probe image and, if so, provide the mask regions for both images (Figure 3). Besides the data for the manipulated image described in MDL, the donor data such as the region in the donor image where the content was used for splice is also used for the evaluation.

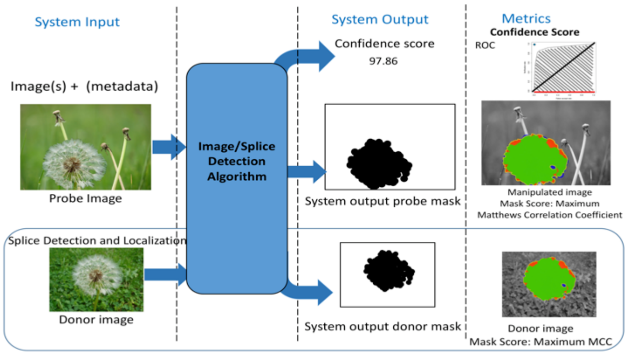


Figure 3: Splice detection and localization task.

5.4. Provenance Filtering and Graph Building

The Provenance Filtering (PF) task (Figure 4) is defined as follows: given a test image, find all images related to the test image from the given large-scale 'world' dataset. PF systems return up to n ancestors and descendants in the test image's phylogeny graph, including a base image and donor images. The 'world' dataset contains a large portion of images downloaded from internet, a portion of nonmanipulated HP images, and all images from the manipulation journals. The test images are HP images and the manipulated images.

The Provenance Graph Building (PGB) task is to first retrieve related images with respect to the given query image from the world dataset, then reconstruct a provenance phylogeny graph; that is, the relationships among the associated images with manipulation sequences. There are two input dataset options for the graph building task: one is to use all data in the 'world' collection, the other is to use a small set of images, an 'oracle' set that contains all relevant images for the given probe with distractors. This oracle set allows the performer to do graph building without working on filtering first, testing PGB systems without being affected by the PF system's performance.

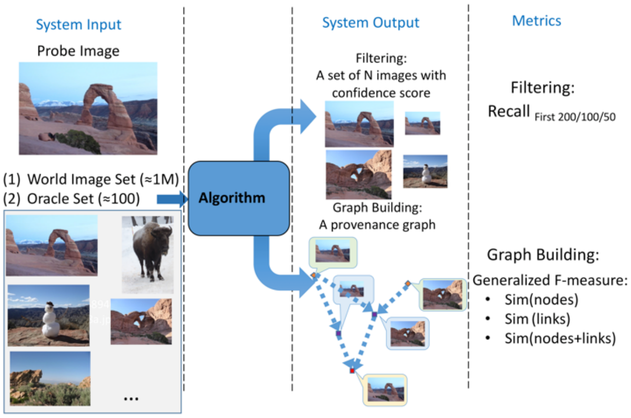


Figure 4: Provenance filtering and graph building task.

5.5. Event Verification

The MFC event verification task is defined as follows: given a collection of images (or videos) captured during an event (e.g. parade) and a probe image asserted to be captured during the event, verify if the probe image was taken during the event or if it was re-purposed. The data collection team collects images from several events such as air shows, hurricanes, marathons, blizzards etc. First, a small set of images is selected and released to the performer team as training images for each event, then another set of images is selected as testing images for each event. Each test image is then paired with each event name to generate the testing pairs serving as the performer's system input. The systems verify if the image belongs to the given event.

5.6. Camera Verification

The Camera Verification (CV) task is to verify if a media is captured by a claimed camera sensor. Distinct from the existing camera model detection task, which identifies the camera model given a media, CV focuses on sensor fingerprint verification. The traditional camera task trains and tests on the same modality (media type: image or video). We include cross-modality evaluations: e.g. the system trains on images, but tests on videos, which are captured from the same group of cameras. This gives the MFC18 CV task six data subsets, shown in Table 2. We also support the localization and selective scoring evaluation.

Table 2: MFC18 Camera Verification datasets.

Test	Train	Probe Pair	Target	Cam	Journal
Image (3761 img.)	Image	5275	2440	39	452
	Video	3383	1720	25	410
	Multimedia	3383	1720	25	410
Video (224 vid.)	Image	289	101	11	67
	Video	289	101	11	67
	Multimedia	289	101	11	67

5.7. GAN Challenge

GAN [1] are new technologies garnering a lot of recent attentions [46][47]. The GAN challenge is to evaluate if a system detects manipulated media produced by GAN-based manipulations. We created the MFC18 GAN full set (1340 images), GAN crop set (1000 images), and GAN video set (118 videos). The major image GAN operations include face swap, fill, erasure, camera model etc. The major video GAN operations include face swap, frame drop, erasure, and inpainting.

6. Evaluation Datasets

We have generated and released the following datasets: the Nimble Challenge 2016 (NC16) kickoff dataset, the NC17 development dataset, the NC17 evaluation dataset, the MFC18 development datasets 1 and 2, the MFC18 evaluation dataset, the MFC18 GAN image and video challenge datasets, the MFC18 camera ID verification dataset, and the MFC18 Event Verification development and evaluation datasets. All development datasets are publicly releasable. We partition the evaluation dataset into three subsets: Evaluation Part 1 (EP1) for public release and EP2 and EP3 for sequester evaluation.

Figure 5 shows the evolution of the evaluation datasets. The initial data collection, manipulation, and annotation started summer 2015. We designed, collected, manipulated, and released the NC16 kickoff dataset, which only contains single-step manipulation journals.

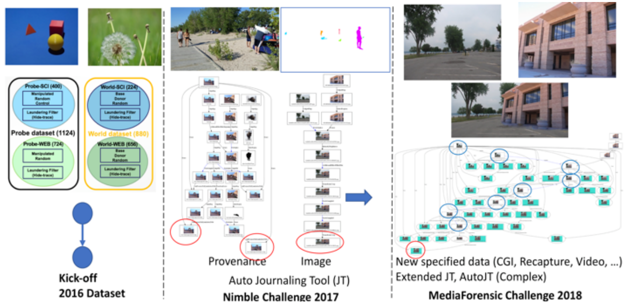

Probe-SCI (400)
Manipulated Random Control
Laundering Filter (Hide-trace)
Probe dataset (1124)
Probe-WEB (724)
Manipulated Random
Laundering Filter (Hide-trace)
World-SCI (224)
Base Donor Random
Laundering Filter (Hide-trace)
World dataset (880)
World-WEB (656)
Base Donor Random
Laundering Filter (Hide-trace)
Kick-off
2016 Dataset
Provenance
Image
Auto Journaling Tool (JT)
Nimble Challenge 2017
New specified data (CGI, Recapture, Video, ...)
Extended JT, AutoJT (Complex)
MediaForensic Challenge 2018

Figure 5: Media Forensic Challenge Dataset History.

In NC17, the data collection team joined the program; we designed and developed the JT and AutoJT, creating complex manipulations and collecting all data and metadata using manipulation journals. We built the NC17 development and evaluation datasets for the first year

MediFor evaluation. Later, the ExtendedJT was developed in 2018 to extend existing human journals with additional automatic manipulations to generate a large amount of manipulation testing images with little cost.

Table 3: MediFor Datasets summary.

Datasets	Image	Video	Size	Releasable
NC17 Dev	3.5 K	23	379	Y
NC17 EP1	4 K	45	3507	Y
NC17 Eval	6 K	1083	3600	N
MFC18 Dev 1	5.6 K	9	88.8	Y
MFC18 Dev 2	38 K	86	524	Y
MFC18 EP1	17 K	323	3200	Y
MFC18 Eval	80 K	2868	12300	N
MFC18 GAN	2.3 K	118	30	Y
MFC18 Cam	3.8 K	224	98	Y
MFC18 Event	2.4K	0	5.6	Y

Table 3 summarizes the number of images and videos, the total sizes, and the access permissions of our MFC development and evaluation datasets.

7. Generation of Evaluation Datasets

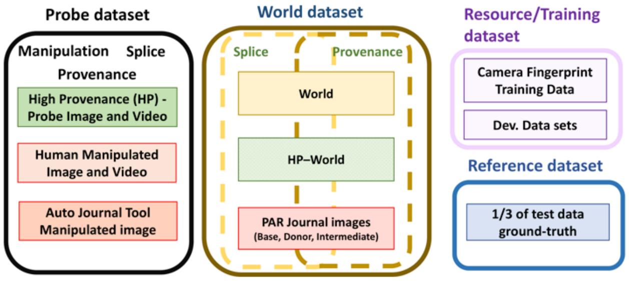

Probe dataset
Manipulation Splice
Provenance
High Provenance (HP) - Probe Image and Video
Human Manipulated Image and Video
Auto Journal Tool Manipulated image
World dataset
Splice
Provenance
World
HP–World
PAR Journal images (Base, Donor, Intermediate)
Resource/Training dataset
Camera Fingerprint Training Data
Dev. Data sets
Reference dataset
1/3 of test data ground-truth

Figure 6: Media forensic evaluation dataset components.

To generate the testing data and its evaluation references, we developed an evaluation dataset generation infrastructure to build the dataset given raw data described in Section 4. Figure 6 shows the general components in the dataset for all tasks. Figure 7 shows the dataset production infrastructure.

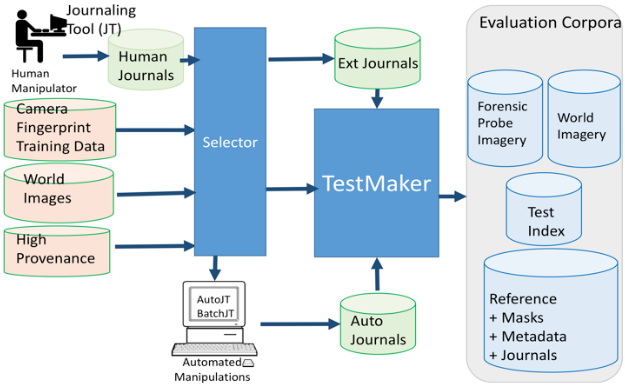


Figure 7: Evaluation datasets generation infrastructure.

We define the first camera set to control public releasable data. We release their PRNU images, development images, and the previous year's testing images to performers for PRNU training. We sequester another set of cameras and their images for future gradual release or sequester evaluation. To make sure the performer's system has enough data for training, we control the number of images released for each camera. Figure 8 shows how many images are released for each camera before the MFC18 evaluation.

Afterwards, we sample a set of manipulation journals for the dataset based on the manipulation operation distribution. Figure 9 shows a stacked histogram of the number of unique manipulation operations in different datasets. It shows that the NC17 datasets (yellow bar) have a limited number of manipulation operations: the distribution of the operations is uneven due to incorporating most of the journals available into the dataset. The MFC18 EP1 dataset's distribution covers most of the operations we wish to test. Note that some operation names were changed across datasets (e.g. "Intensity Normalization" in NC17 is covered by "Normalization" in MFC18). Finally, we select HP data and World data as nontarget for different tasks.

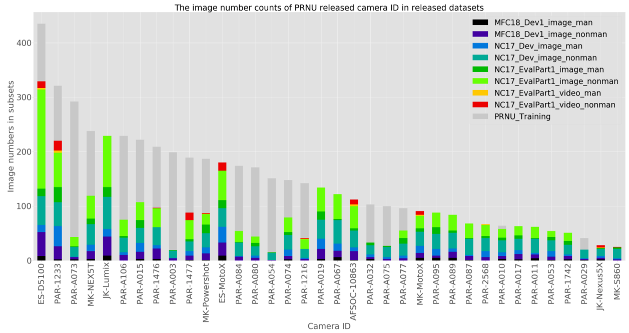


Figure 8: The stacked histogram of released image counts.

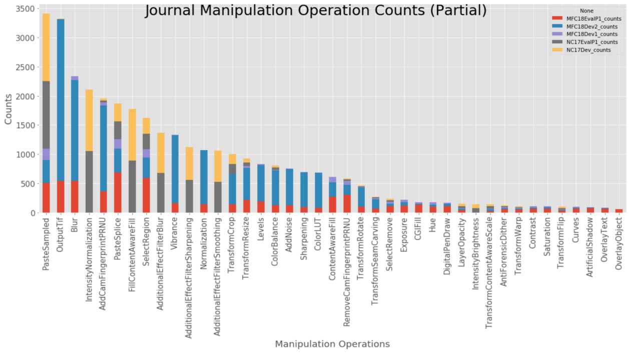


Figure 9: Stack histogram of manipulation operations.

Given the image/video manipulation journal, TestMaker extracts the data and metadata from the journal files, dynamically selects the testing images in the journals, and generates its JPEG2000 localization reference masks.

With the AutoJT and ExtendedJT, many testing images can be automatically generated with designated operations and parameters. Due to limited computational resources and evaluation time, we control the size of the test data by down sampling the manipulated images based on factor independence among data. Three approaches are used for data selection: (i) Random sampling, (ii) Specified sampling: e.g. traverse the longest journal path and select the middle and final node images. (iii) Sampling based on the distributions of the given factors.

8. MFC Evaluation Results

First, we demonstrate the selective scoring function to enable us to better understand the systems. Given the same set of systems, Figure 10 compares the evaluation results on all data with all manipulation types (left) with selective scoring results (right) on a particular manipulation (the target set is crop images extracted from the same test set). Two ROC curves (right) highlight the scoring system's ability to isolate the performance of components for particular manipulations. A separate document will discuss the MFC18 EP1 evaluation results in greater detail.

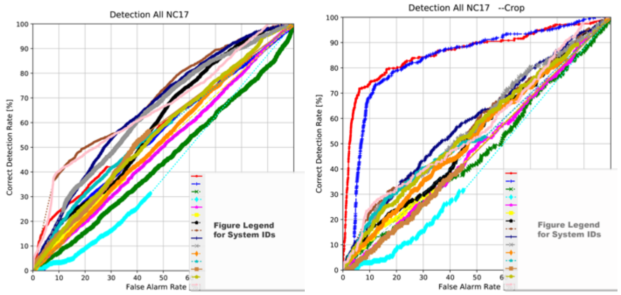


Figure 10: Full (left) vs. Selective Scoring on NC17 EP1.

Figure 11 shows an example graph building evaluation result. Green represents a correctly identified node/link, red represents an incorrect one, and gray represents a missing one. The graph node and link overlap similarity score (SimNLO, a generalized F-measurement) is 0.7.

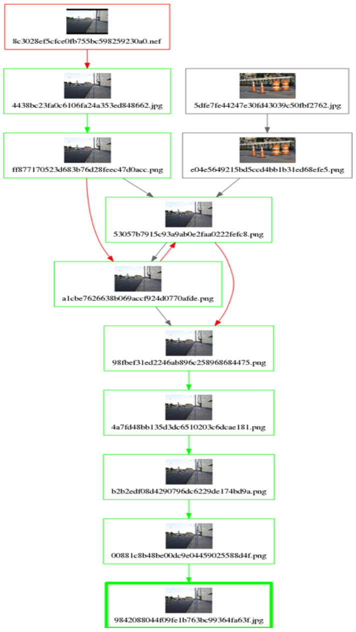

8c3028ef5cfce0fb755bc598259230a0.nef
4438bc23fa0c6106fa24a353ed848662.jpg
5dfe7fe44247e30fd43039c50fbf2762.jpg
ff877170523d683b76d28feec47d0acc.png
e04e5649215bd5ccd4bb1b31ed68efe5.png
53057b7915c93a9ab0e2faa0222fefc8.png
a1cbe7626638b069accf924d0770afde.png
98fbef31ed2246ab896c258968684475.png
4a7fd48bb135d3dc6510203c6dcae181.png
b2b2edf08d4290796dc6229de174bd9a.png
00881c8b48be00dc9e04459025588d4f.png
9842088044f09fe1b763bc99364fa63f.jpg

Figure 11: An example graph building evaluation.

Table 4 compares the performance of the best systems (the maximum score of all systems) of the NC17 and MFC18 evaluations. We compare the systems that evaluated all data of NC17 on NC17 EP1 with MFC18 on MFC18 EP1. The MFC18 EP1 and NC17 EP1 columns show: (i) the Max AUC for image detection systems improved from 0.69 to 0.84; (ii) the Max MCC for image localization systems improved from 0.08 to 0.26; (iii) the Max Recall for provenance filtering systems improved from 0.69 to 0.88; (iv) the Max SimNLO for provenance graph building systems on oracle systems on Full Graph metrics improved from 0.49 to 0.61. The system performance on NC17 Eval and NC17 EP1 is similar given the same conditions. After NC17 EP1 was released, the performances improved on both NC17 EP1 and Eval data.

Table 4: NC17 evaluation vs. MFC18 evaluation

Task and Metric Score	NC17 EP1	NC17 Eval	NC17 EP1 after NC17EP1 release	NC17 Eval after NC17EP1 release.	**MFC18** EP1
Image MD (AUC)	0.688	0.696	0.822	0.858	0.837
Image MDL (MCC)	0.082	0.009	0.290	-	0.258
PF (Recall)	0.689	0.649	0.782	0.688	0.882
PGB (SimNLO)	0.489	-	-	-	0.612
SD (AUC)	0.769	0.795	0.832	0.863	0.767
SDL (MCC)	-	-	0.295	-	0.361
Video MD (AUC)	-	0.580	-	0.580	0.594

9. Conclusion

We proposed an approach on designing and building an evaluation benchmark on a new research domain: media forensics. We present a series of datasets for different types of media forensic system evaluations and compare the evaluation results from the last two years' evaluations. We are continuing to collect and generate more data for MFC19 and MFC20 evaluations. The proposed methodology could be further generalized to other research domains.

The released datasets for NC17[2] and MFC18[3] are available upon request via email: mfc_poc@nist.gov. The evaluation scoring software package, MediScore, can be downloaded from https://github.com/usnistgov/MediScore.

10. Acknowledgement

The authors gratefully acknowledge the members of the DARPA Media Forensics (MediFor) Program and members of the Air Force Research Lab for managing the program and weekly data team meetings; special thanks goes to David Doermann and Rajiv Jain for their direct instructions and support. The authors would like to thank other PAR Government team members for media collection, manipulation, JT development; Rochester Institute of Technology, Drexel University, and University of Michigan members for their technical supports to PAR government; University of Colorado Denver for their professional media manipulations, antiforensic tools, social media laundering tools; the RankOne MediFor team for world data collection and the Medifor Browser; and Wendy Dinova-Wimmer for the NC16 Kickoff professional data manipulations and initial manipulation data collection design. PAR Government conducted this work under DARPA sponsorship via Air Force Research Laboratory (AFRL) contract FA8750-16-C-0168. NIST conducted this work under NIST Interagency Agreement Number 1505-774-08-000.

[2] https://www.nist.gov/itl/iad/mig/nimble-challenge-2017-evaluation

[3] https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018

References

[1] T. Karras, T. Aila, S. Laine, and J. Lehtinen, Progressive Growing of GANs for Improved Quality, Stability, and Variation, International Conference on Learning Representations, 2018.

[2] T. Wang, M. Liu, J. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro, Video-to-Video Synthesis, arXiv:1808.06601, 2018

[3] A. Piva, An Overview on Image Forensics, ISRN Signal Processing, 2013:1–22, 2013.

[4] M. C. Stamm, M. Wu, and K. J. R. Liu, Information Forensics: An Overview of the First Decade, IEEE Access, 1:167–200, 2013.

[5] H. Farid, Image forgery detection, IEEE Signal Process. Mag., 26(2):16–25, 2009.

[6] J. Fridrich, Digital image forensics using sensor noise, Signal Proc. Mag. IEEE, 26(2): 26–37, 2009.

[7] H. T. Sencar and N. Memon, "Overview of state-of-the-art in digital image forensics," Algorithms Archit. Inf. Syst. Secur., 3:325–348, 2008.

[8] R. Boehme, M. Kirchner, Counter-Forensics: Attacking Image Forensics, Digital Image Forensics, H. T. Sencar and N. D. Memon, eds., pp. 327-366, Springer, 2012.

[9] Siwei Lyu, Natural Image Statistics in Digital Image Forensics, Springer, Digital Image Forensics, pp 239-256, 2012.

[10] S. Tariq, S. Lee, H. Kim, Y. Shin, S. S. Woo, Detecting Both Machine and Human Created Fake Face Images In the Wild, the 2nd International Workshop on Multimedia Privacy and Security, pp. 81-87, 2018.

[11] DARPA MediFor, https://www.darpa.mil/program/media-forensics,https://www.fbo.gov/index?s=opportunity&mode=form&id=bfa29e5f04566fbb961cd773a8a8649f&tab=core&_cview=1.

[12] A. Yates; H. Guan; Y. Lee, A. Delgado, D. Zhou, J. Fiscus, Nimble Challenge 2017 Evaluation Plan, NIST, 2017

[13] A. Yates; H. Guan; Y. Lee, A. Delgado, D. Zhou, J. Fiscus, Media Forensics Challenge 2018 Evaluation Plan, NIST, 2018

[14] G. Schaefer and M. Stich, UCID: an uncompressed color image database, Proceedings of the SPIE, 5307:472-480, 2003.

[15] J. Kalpathy-Cramer, A. Herrera, D. Demner-Fushman, S. Antani, S. Bedrick and H. Müller, Evaluating performance of biomedical image retrieval systems - an overview of the medical image retrieval task at ImageCLEF 2004-2013, Computerized Medical Imaging and Graphics, 39:55-61, 2015.

[16] G. Griffin, A. Holub, and P. Perona, Caltech-256 Object Category Dataset, California Institute of Technology. http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001, 2007.

[17] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition, PAMI, 30(11):1958–1970, 2008.

[18] M. Fink and S. Ullman. From Aardvark to Zorro: A Benchmark for Mammal Image Classification. IJCV, 77(1-3):143–156, 2008.

[19] M. Everingham, S. M. Ali Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL Visual Object Classes Challenge: A Retrospective, International Journal of Computer Vision, 111(1): 98–136, 2015.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[21] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN Database: Large-scale Scene Recognition from Abbey to Zoo, IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR), 2013.

[23] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft COCO: Common Objects in Context, Computing Research Repository (CoRR), arXiv, abs/1405.0312, 2014.

[24] A. F. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and TRECVid. MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2006.

[25] S. Oh et al., A large-scale benchmark dataset for event recognition in surveillance video, CVPR, pp. 3153-3160, 2011.

[26] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L. Li, YFCC100M: The New Data in Multimedia Research, Communications of the ACM, 59(2):64-73, 2016.

[27] M. Larson, M. Soleymani, G. Gravier, B. Ionescu and G. J. F. Jones, The Benchmarking Initiative for Multimedia Evaluation: MediaEval 2016, IEEE MultiMedia, 24(1):93-96, 2017.

[28] EU project REWIND (REVerse engineering of audio-VIsual coNtent Data), https://sites.google.com/site/rewindpolimi/home .

[29] A. Rocha, A. Piva, and J. Huang, The First IFS-TC Image Forensics Challenge, http://ifc.recod.ic.unicamp.br/.

[30] T.-T. Ng, S.-F. Chang, and Q. Sun, A data set of authentic and spliced image blocks, Columbia Univ. ADVENT Tech. Rep., pp. 203–2004, 2004.

[31] CalPhotos, A database of plants, animals, habitats and other natural history subjects, http://calphotos.berkeley.edu/, 2015.

[32] J. Dong, W. Wang, and T. Tan, CASIA image tampering detection evaluation database, in Signal and Information Processing (ChinaSIP), IEEE China Summit & International Conference on, pp. 422–426 2013.

[33] D. Tralic, I. Zupancic, S. Grgic and M. Grgic, CoMoFoD - New database for copy-move forgery detection, IEEE Proceedings ELMAR-2013, pp. 49-54, Zadar, 2013.

[34] I. Amerini, L. Ballan, R. Caldelli, A. D. Bimbo, and G. Serra, A SIFT-based forensic method for copy-move attack detection and transformation recovery, IEEE Trans. Inf. Forensics Security, 6(3): 1099–1110, 2011.

[35] V. Christlein, C. Riess, J. Jordan, C. Riess, E. Angelopoulou, An Evaluation of Popular Copy-Move Forgery Detection Approaches, IEEE Transactions on Information Forensics and Security, 7(6):1841-1854, 2012.

[36] S. Agarwal, W. Fan, and H. Farid, A Diverse Large-Scale Dataset for Evaluating Rebroadcast Attacks, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1997-2001, 2018.

[37] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, R. Chellappa, UMDFaces: An Annotated Face Dataset for Training Deep Networks, arXiv:1611.01484v2,2016.

[38] D. Shullani, M. Fontani, M. Iuliani, O. Al Shaya, A. Piva, VISION: a video and image dataset for source identification, EURASIP Journal on Information Security, 2017:15, 2017.

[39] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces, arXiv:1803.09179, 2018.

[40] P. Bas, T. Filler, T. Pevny, "Break Our Steganographic System": The Ins and Outs of Organizing BOSS. INFORMATION HIDING, Czech Republic. 6958/2011:59-70, Lecture Notes in Computer Science, 2011.

[41] N. Khanna, A. K. Mikkilineni, G. T. Chiu, J. P. Allebach, E. J. Delp, Survey of Scanner and Printer Forensics at Purdue University. Computational Forensics. IWCF 2008. Lecture Notes in Computer Science, vol. 5158. Springer, 2008.

[42] M. Goljan, J. Fridrich, and T. Filler, Large scale test of sensor fingerprint camera identification, in Proc. SPIE Media Forensics and Security, Jan. 2009, vol. 7254.

[43] T. Gloe and R. Böhme, The Dresden Image Database for Benchmarking Digital Image Forensics, Proceedings of the 25th Symposium On Applied Computing (ACM SAC 2010), 2:1585–1591, 2010.

[44] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, G. Boato, RAISE – A Raw Images Dataset for Digital Image Forensics, ACM Multimedia Systems, 2015.

[45] E. Robertson, "Manual for MaskGen Journaling Tool", PAR Government Solution, 2018, online link: https://github.com/rwgdrummer/maskgen/blob/master/doc/MediForJournalingTool-public.pdf.

[46] Jennifer Finney Boylan, Will Deep-Fake Technology Destroy Democracy? The New York Times, Oct. 17, 2018. https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html.

[47] Hilke Schellmann, Deepfake Videos Are Getting Real and That's a Problem, The Wall Street Journals, Oct. 15, 2018. https://www.wsj.com/articles/deepfake-videos-are-ruining-lives-is-democracy-next-1539595787.