
AFRL-RI-RS-TR-2021-035

# NEW GENERATION OF IMAGE PROCESSING HISTORY AND MANIPULATION DETECTION TECHNIQUES WITH VECTORIZED CONTEXT-AWARE DESCRIPTORS REFOCUSED TO MEDIA SOURCE IDENTIFICATION AND VERIFICATION, AND STEGANALYSIS OF JPEG IMAGES

THE RESEARCH FOUNDATION OF THE STATE UNIVERSITY OF NEW YORK AT BINGHAMTON

*FEBRUARY 2021*

FINAL TECHNICAL REPORT

***APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED***

STINFO COPY

**AIR FORCE RESEARCH LABORATORY**
**INFORMATION DIRECTORATE**

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

# NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-035 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

**/ S /**
JEFFREY T. CARLO
Work Unit Manager

**/ S /**
JAMES S. PERRETTA
Deputy Chief, Information
Exploitation and Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

# REPORT DOCUMENTATION PAGE

*Form Approved*
**OMB No. 0704-0188**



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| FEBRUARY 2021 | FINAL TECHNICAL REPORT | MAY 2016 – MAY 2020 |

**4. TITLE AND SUBTITLE**

NEW GENERATION OF IMAGE PROCESSING HISTORY AND MANIPULATION DETECTION TECHNIQUES WITH VECTORIZED CONTEXT-AWARE DESCRIPTORS REFOCUSED TO MEDIA SOURCE IDENTIFICATION AND VERIFICATION, AND STEGANALYSIS OF JPEG IMAGES

**5a. CONTRACT NUMBER**
FA8750-16-2-0192

**5b. GRANT NUMBER**
N/A

**5c. PROGRAM ELEMENT NUMBER**
62303E

**6. AUTHOR(S)**

Miroslav Goljan
Matthias Kirchner
Jessica Fridrich

**5d. PROJECT NUMBER**
MEDI

**5e. TASK NUMBER**
40

**5f. WORK UNIT NUMBER**
02

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

The Research Foundation of the State University of New York at Binghamton
PO Box 6000
Binghamton NY 13902-6000

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Air Force Research Laboratory/RIGC
525 Brooks Road
Rome NY 13441-4505

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AFRL/RI

**11. SPONSOR/MONITOR'S REPORT NUMBER**
AFRL-RI-RS-TR-2021-035

**12. DISTRIBUTION AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

This report describes the Binghamton University MediFor team's research results on the DARPA Media Forensics (MediFor) program. The results are divided into four focus areas aligned with the main thrust of the MediFor program: 1) detection of manipulated images and localization of the manipulated content, 2) detection of processing history of images, 3) media source identification and verification based on sensor fingerprint (PRNU), and 4) detection of a specific manipulation introduced due to data hiding (steganography) in JPEG images.

**15. SUBJECT TERMS**

Photo-Response Non-Uniformity (PRNU), steganography, media forensics, manipulation detection, processing history

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | **JEFFREY T. CARLO** |
| U | U | U | UU | 48 | 19b. TELEPHONE NUMBER (*Include area code*) **N/A** |

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## 1.0 EXECUTIVE SUMMARY

This report describes the results of research divided into four focus areas aligned with the main thrust of the MediFor effort: 1) detection of manipulated images and localization of the manipulated content, 2) detection of processing history of images, 3) media source identification and verification based on sensor fingerprint (PRNU), and 4) detection of a specific manipulation introduced due to data hiding (steganography) in JPEG images.

Progress has been made along all four focus areas that resulted in novel and improved algorithms that were implemented and evaluated via self-evaluation, through official program evaluation, and external independent evaluation.

Image manipulation localization has been advanced through development of techniques that make explicit use of contextual information and flexible unsupervised clustering approaches. As for the former, we have emphasized manipulation localization based on camera sensor noise fingerprints, which we have combined with Discriminative Random Field (DRF) inference to reason over local analysis neighborhoods jointly. Our unsupervised hierarchical clustering approach in the pixel descriptor space of residual co-occurrences relaxes the common constraint that probe pixels belong to one of exactly two classes, genuine or manipulated. Our greedy n-ary clustering approach produces non-binary label assignments instead, which helps mitigate common false alarms due to genuine but singular content. In both instances, program and self-evaluation point to favorable outcomes and measurable improvements over state-of-the-art.

Two different approaches were designed, implemented, and evaluated for detection of image processing history, which includes low-pass, high-pass filtering, denoising, and contrast enhancement possibly combined with gamma correction. The first approach is based on training binary linear classifiers in a feature space (features taken from steganalysis literature, such as the SRM feature vector) that distinguish unprocessed images versus images processed with a specific type (but a wide range of parameters) of processing. The weight vectors of each classifier were then used as a new basis within which a multi-class maximum-likelihood detector was implemented by modeling the projections of SRM features onto the basis vectors as multi-variate Gaussians. The second approach was based on building a multi-class classifier as a convolutional neural network designed specifically for the task. The second approach offered better classification accuracy. Both algorithms have been described in detail in two Electronic Imaging papers.

Photo-Response Non-Uniformity(PRNU) estimation for camera verification was expanded to HDR images produced by mobile devices. Such images had been previously avoiding camera identification. After identifying the culprit as being a combination of scaling, cropping, and localized translation of the sensor output, a series of synchronization steps were developed to deal with all involved geometric transforms. This solution was summarized in publication [HDR] at 2019 Information Hiding Workshop. A fast and high precision technique for image resampling factor estimation is another tool that was developed for PRNU matching. This novel method uses Fourier-Mellin transform for fast computation. PRNU estimation was further refined and sped up using convolutional neural networks (CNN). The first such method utilizes FFDNet developed outside this program for image denoising. Next, a custom designed CNN for direct PRNU estimation went through a series of development cycles in order to guarantee high performance and generalization ability.

The fourth focus area, steganalysis of JPEG images, aimed at detecting a diversified set of stego algorithms, variable payload, in JPEG images of variable size and a wide range of JPEG quality factors. All algorithms are based on the SRNet, a steganalysis network specifically designed for steganalysis in both spatial and JPEG domains. The detectors were designed as three-

channel YCrCb multi-class SRNets first trained on small tiles with the final detector built as a multi-layer perceptron on statistical moments extracted from the feature maps outputted by the last convolutional layer. A special extremely accurate and universal JPEG-compatibility attack for quality factors 99 and 100 was developed as part of the first independent evaluation, the steganalysis competition ALASKA I, which was won by Binghamton team. The second evaluation, ALASKA II competition run on Kaggle focused on detection of four modern steganography methods. The team developed a novel approach based on ImageNet-pre-trained computer vision models to claim a 2$^{nd}$ place in the competition.

## 2.0 INTRODUCTION

The focus of this effort was directed to four separate but closely related areas in the field of digital media forensics that can be summarized in one phrase as "detection of media manipulation and establishing media provenance and integrity." To this end, the effort was split into four different areas. Two of them, the detection of the processing history and the detection and localization of manipulation have been investigated during the first two years. After the PI, Matthias Kirchner, left Binghamton team, the focus of this effort for the next two years shifted to detection of a specific type of manipulation introduced during steganographic embedding (steganalysis) and forensics that leverages Photo-Response Non-Uniformity (PRNU). This text reports the achievements separately for all four research directions.

## 3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

This section describes the essence of all developed algorithms and briefly summarizes their performance as determined from experiments from published work and from evaluation.

### 3.1 Vectorized context-aware pixel descriptors for manipulation detection

We describe two major contributions in the domain of image manipulation localization below.

#### 3.1.1. PRNU-based image manipulation localization with discriminative random fields

Traditionally, many state-of-the-art forensics algorithms attempt to localize image manipulations by inspecting local image neighborhoods independent of their surroundings. However, local decisions are likely to depend on each other, in particular when a manipulation spans multiple analysis windows. More generally, local decisions may also be influenced by non-local image characteristics (e.g., global brightness or noise level). Forensic schemes that fail to incorporate such information explicitly risk ignoring valuable context that may help improve localization and detection accuracy otherwise.

It is well-known [SD] that minute manufacturing imperfections of individual sensor elements lead to a spatially varying multiplicative noise pattern, the photo-response non-uniformity (PRNU) that can be estimated and tested for in forensic applications. Localized image manipulations are likely to weaken or remove the noise pattern. The image content in an analysis window is likely not genuine, if a correlation-based similarity score with the source camera's fingerprint falls below a suitably chosen threshold [SD].

On MediFor, we have formulated image manipulation localization based on camera sensor noise as a probabilistic binary labeling task in a flexible discriminative random field (DRF) framework to incorporate contextual information [DRF]. We instantiate the association potentials with a novel local discriminator based on the deviation of the measured correlation from the expected local correlation as estimated by a correlation predictor. The interaction potentials introduce an explicit pairwise model for dependencies between local decisions (**Figure 1**). We use a grid search over suitable candidate settings to obtain favorable parameter settings. A GraphCut-based approach is used for efficient inference at test time. The algorithm output is a binary manipulation map that indicates for each pixel the most likely decision: genuine or manipulated. A crude image-level detection score is based on the relative size of inter-connected "manipulated" labels.

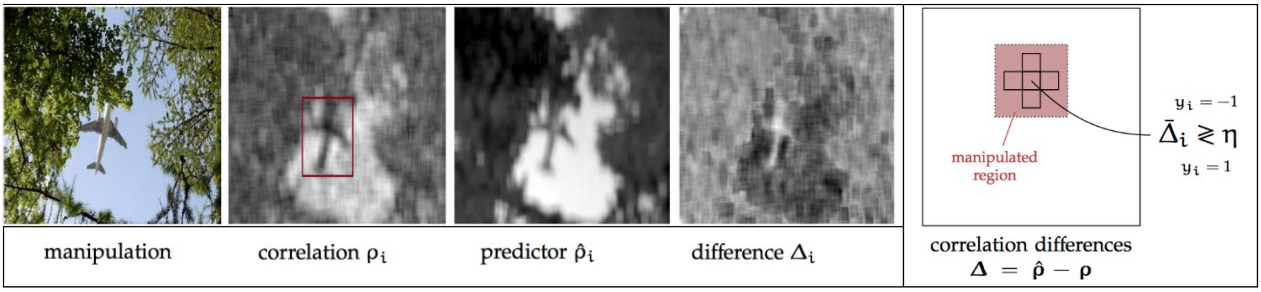



**Figure 1. From left to right: image manipulation with inserted plane; local correlation of image noise residual with camera fingerprint; expected correlation as obtained from a correlation predictor; difference between measured and predicted correlation; our algorithm reaches local decisions by taking neighborhood information into account.**

### 3.1.2 Unsupervised image manipulation localization with non-binary label attribution

Unsupervised forensic algorithms attempt to localize spurious pixel inconsistencies due to image manipulation solely based on a suitable parametrizable representation of a given probe, often framing the problem as clustering or outlier detection in the chosen descriptor space. This can be a desirable property in practice, yet the increased ability to generalize comes at the risk of impairing specificity: in the absence of labeled training examples, unsupervised techniques often tackle with false positives caused by singular but genuine content, in particular when it is assumed that probe pixels belong to one of exactly two classes, genuine or manipulated.

On MediFor, we have worked to relax such constraints via a greedy n-ary clustering approach, which we instantiate exemplarily in the popular pixel descriptor space of residual co-occurrences [SRM][SPLICE]. Once the pixel descriptors have been computed, the goal is to assign pixels to meaningful clusters in the descriptor space, where we allow pixel descriptors to fall into more than two categories. We adopt an *n*-ary hierarchical clustering approach with an agglomerative strategy. As the computational complexity of a full hierarchical clustering on pixel level is prohibitive, we employ the following greedy heuristic: starting from a simple k-means clustering with k>2, we merge clusters iteratively based on their distance in the descriptor space until n<k clusters remain. For the sake of brevity, we refer to this approach as HCl-n (i.e., HCl-3 indicates that we work with three clusters).

The result of the above clustering process can be represented in an *n*-ary discrete map that indicates the pixel attributions. This map by itself is of value to further forensic analyses: it provides a descriptor-based segmentation of the probe image. Examples in **Figure 2** indicate that the obtained maps (here for $n = 3$) will often outline non-genuine regions very accurately. Such segmentation can assist with directing further scrutiny to salient regions of interest. As a simple example, this work produces real-valued manipulation heat maps by assuming that the most populated cluster is most likely representing genuine content. This allows us to compute pixel-wise Mahalanbobis distances of the local descriptors from the ``background'' cluster. In the given context, larger distances indicate an increased likelihood of content being inconsistent.

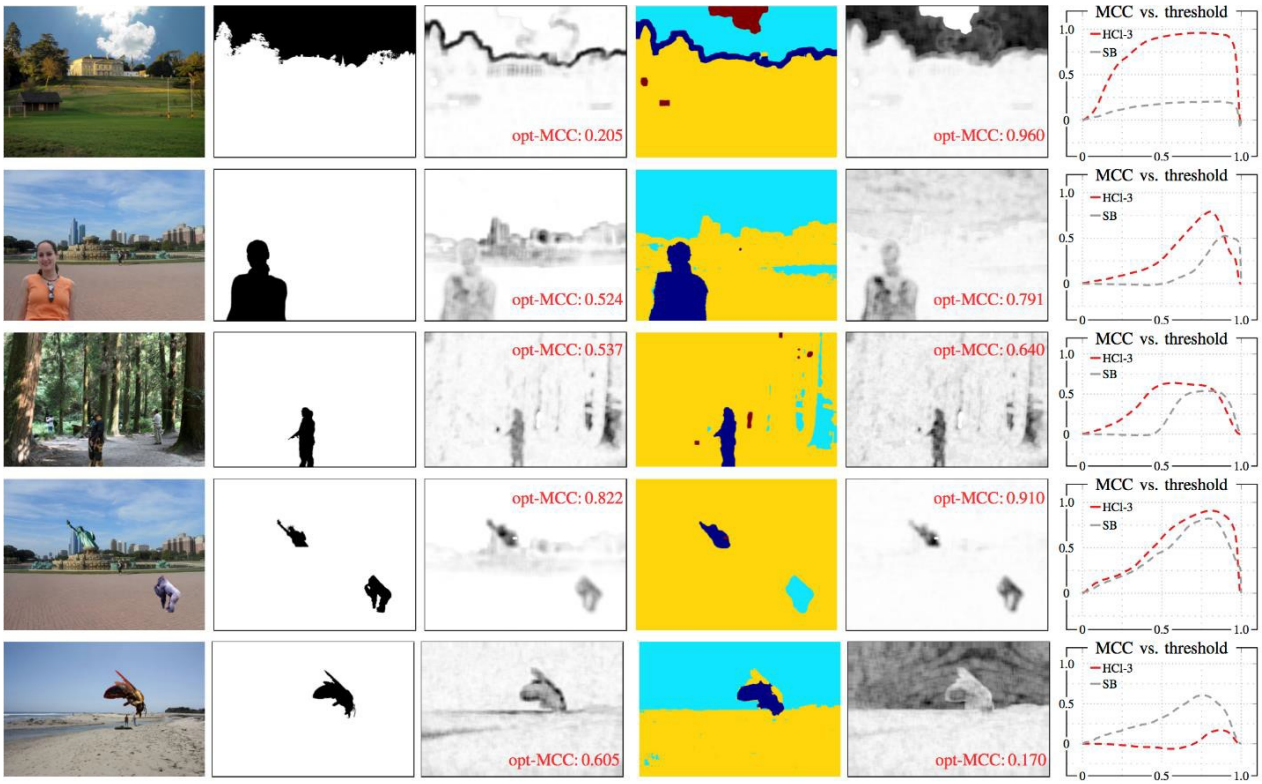



**Figure 2. From left to right: manipulated image, ground truth, SpliceBuster heat map with opt-MCC, HCl-3 ternary cluster attribution and heat map with opt-MCC, and MCC scores for all binarization thresholds. Red color in the ternary maps indicates pixel saturation (excluded from analysis in both algorithms).**

### 3.2 Detection of processing history

Establishing the pedigree of a digital image, such as the type of processing applied to it, is important for forensic analysts because processing generally affects the accuracy and applicability of other forensic tools used for, e.g., identifying the camera (brand) and/or inspecting the image integrity (detecting regions that were manipulated). Global edits have been proposed in the past for "laundering" manipulated content because they can negatively affect the reliability of many forensic techniques. This research focuses on the more difficult and less addressed case when the processed image is JPEG compressed. Two approaches were investigated and evaluated:

#### 3.2.1. Maximum-likelihood detector

A bank of binary linear classifiers with rich media models is built to distinguish between unprocessed images and images subjected to a specific processing class. For better scalability, the detector is not built in the rich feature space but in the space of projections of features on the weight vectors of the linear classifiers. This decreases the computational complexity of the detector and, most importantly, allows estimation of the distribution of the projections by fitting a multi-variate Gaussian model to each processing class to construct the final classifier as a maximum-likelihood detector. Because the ML is constructed in the space of projections, whose dimensionality is the number of processing classes (or processing chains for detection of an entire chain of operations), it is feasible to use parametric models. Well-fitting analytic models

permit a more rigorous construction of the detector unachievable in the original high-dimensional rich feature space.

The merit of the approach is demonstrated on grayscale and color images for a range of quality factors. Four processing classes are investigated – low-pass filtering, high-pass filtering, denoising, and tonal adjustment. The detector for grayscale images was implemented using the SRM [SRM] ($q = 1$) model while for color images, the SCRM model [SCRM] was used.

There are some natural as well as fundamental limitations of this approach. For example, it may be rather challenging to distinguish an out-of-focus image of a flat scene from a low-pass filtered (blurred) image. Strong operations may overpower (neutralize the impact of) others, e.g., sharpening followed by aggressive low-pass filtering. Also, the order of operations that commute cannot be established, e.g., contrast adjustment and linear filtering.

### 3.2.2. Convolutional neural network detector

Given the superiority of automatized tools called deep convolutional neural networks to learn complex yet compact image representations for numerous problems in steganalysis as well as in forensic, the co-PI explored this alternative, modern tool for detecting the processing history of images. This time, the processing pipeline included situations when an image acquired by a camera and processed was aggressively downscaled with a wide variety of scaling factors, and again JPEG compressed with a low quality factor since such processing pipeline is commonly applied for example when uploading images to social networks, such as Facebook. To allow the network to perform accurately on a wide range of image sizes, a novel CNN architecture with an IP layer accepting statistical moments of feature maps was incorporated.

To assure that the detector classifies images of arbitrary size with the best possible accuracy (accuracy similar to a detector that could be trained on large images), the training was performed in three phases. The first phase involved training a "moment extractor" module on small images (512×512 tiles) which was then in the second phase used to extract moments from all (arbitrarily sized) training images. In the final third phase, just the classification part, the inner-product layers, were trained to map the extracted moments to the same processing classes as in the previous section.

The detector was trained for three final JPEG quality factors separately. It was able to generalize to previous unseen median filtering and correctly classify it as either low-pass filtering or denoising.

## 3.3 Media source identification and verification

PRNU of an imaging sensor is caused by a small deviation from the sensor ideal response to incoming light. It can be estimated and detected from images the camera that is equipped with this sensor takes. Estimated PRNU can serve as a unique fingerprint of the camera. The PRNU-based method of camera identification from a single image the camera took was first introduced in 2005 [DIori]. Low error rates were demonstrated for an updated method in 2009 [LST]. This large scale test was conducted with an assumption that the image processing was the same for test images as for the reference images from which the PRNU was estimated. Detection statistic for matching a query image to camera fingerprints of equal dimensions is in a form of correlation

coefficient. Its computation requires that pixels from the query image come from the same physical location on the camera imaging sensor as the pixels from reference images. Therefore, any geometric transformation applied to the image hampers positive identification.

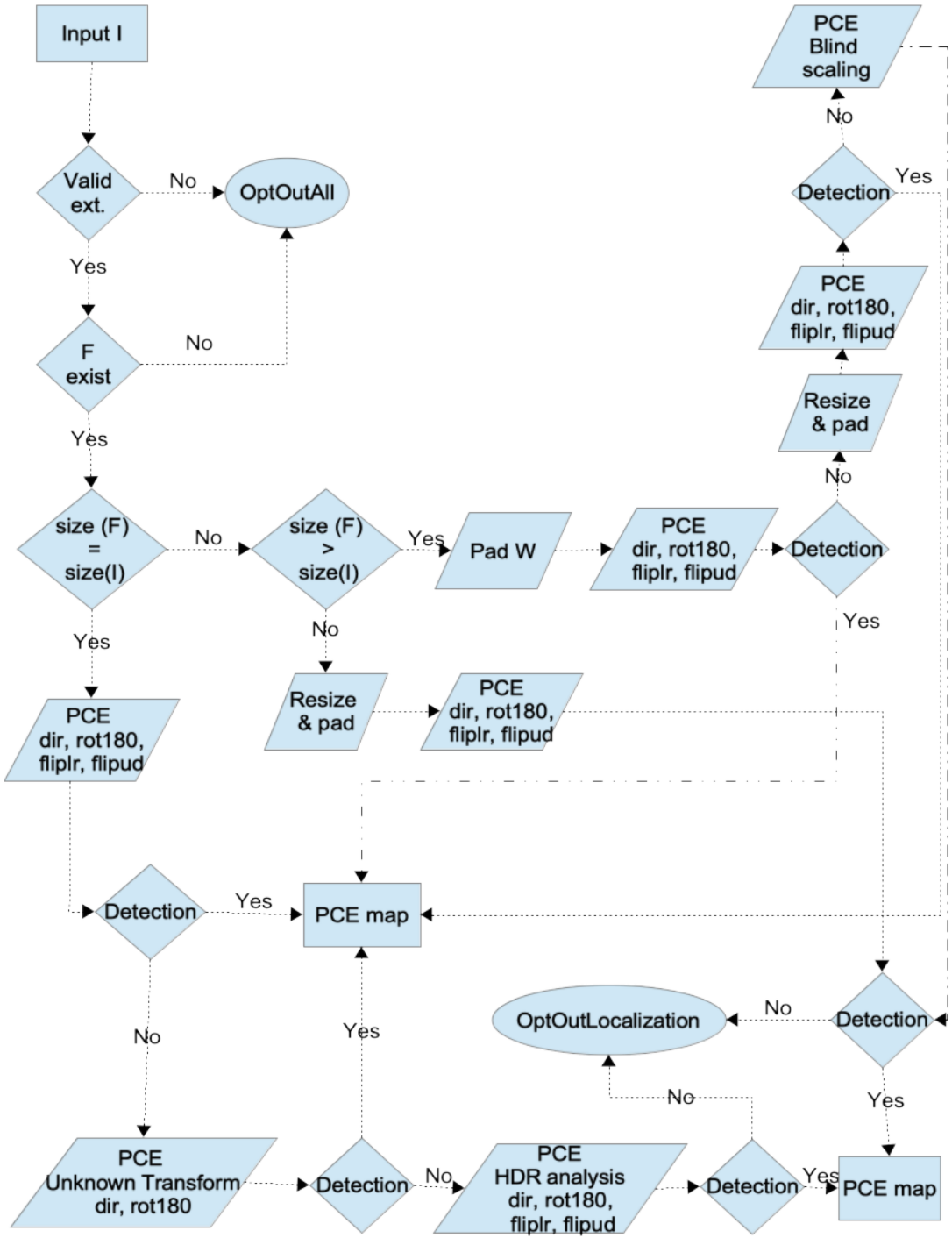



**Figure 3. Flow chart of the implemented Camera Verification System.**

Within this project, a number of cases of geometric transformations and their combinations have been addressed and implemented in the Camera Verification System (listed at the DARPA MediFor Analytic page as p-bingcamfinghdr20), such as high dynamic range (HDR) images, image resizing (including up-sampling), and digital zoom.

The flow chart of the containerized system is shown in **Figure 3**. The main branching depends on the relationship between the probe image size and the camera fingerprint size. For example, digitally zoomed image are assumed to have dimensions equal to that of the fingerprint.

### 3.3.1. Camera verification for HDR images

Producing HDR images became a standard option for cameras in consumer mobile devices. In most devices, such images are produced as a combination of a few exposures. If a camera or object movement occurs during shooting, image registration has to be applied to the three source images before the new pixel values are computed. The pixel-to-pixel correspondence between the image and the camera sensor is broken. The standard matching procedure for camera identification based on PRNU often fails for such images [SIDHDR].

The developed solution for camera verification from HDR images is based on inverting underlying geometrical transformations the image was subject to. The proposed reversal of spatial transformations consists of two main stages. In the first stage, upsampling factor and major shift parameters that maximize the peak-to-correlation energy ratio (PCE) are determined. The second stage called Patchwork (PW) deals with identifying groups of blocks (clusters) that were displaced in the same direction pertaining to one of the three input image. The flowchart in **Figure 4. Resynchronization process for HDR images.** depicts the main components of the entire system. While the first stage provides the main gain in terms of PCE and positive detection rate the second stage was proposed to prepare the grounds for tamper detection applications that utilize PRNU as the camera fingerprint. The fast block-wise searches BW1 and BW2 are responsible for reducing the average processing time.

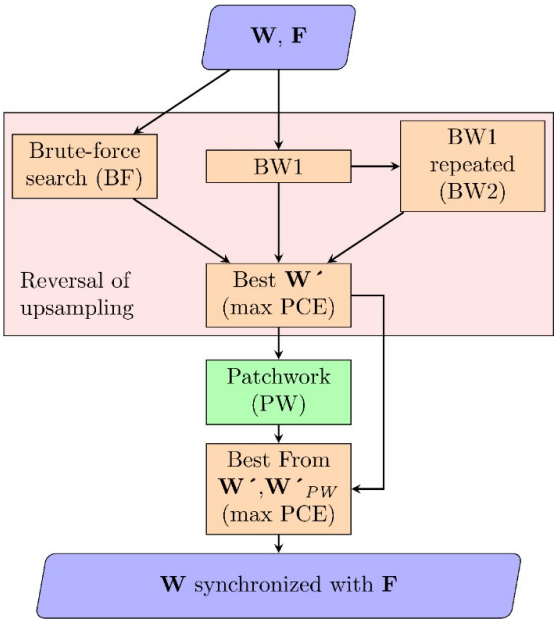



**Figure 4. Resynchronization process for HDR images.**

Details of this work including examples and experiments on a dataset of HDR images were published at the 7th Information Hiding Workshop IH&MMSEC in 2019 [HDR]. All components of the camera ID method for HDR images have been implemented in the camera verification tool listed on the DARPA MediFor Analytic page as p-bingcamfinghdr20.

### 3.3.2. Camera verification for digitally zoomed images and image resampling factor estimation

Digitally zoomed images are up-sampled images with the geometric center located within a small region around the original center. Addressing fingerprint matching for this type of images leads to the problem of upsizing detection and estimation. As for most images that do not maintain one-to-one correspondence between the image pixels and the camera fingerprint computed from other (reference) images, digitally zoomed images require reversing the spatial transform.

The most reliable method for matching a camera fingerprint to a digitally zoomed probe image deploys an exhaustive search for the upscaling factor [CISC]. Computation complexity of this method increases with increasing sensor resolution. In the effort of finding an efficient method for matching digitally zoomed images to camera fingerprints, the method originally proposed by Kirchner and Gloe [RD] was analyzed and improved to find scaling factors of in-device up-sampled images, including digitally zoomed images. This is a blind method (i.e. not using a reference image or sensor fingerprint) that analyses peaks in the Fourier domain of the probe image. The original method comes with limitations, one of which is an ambiguity problem. Even if the peak in Fourier domain is found correctly, the scaling factor cannot be uniquely determined unless a particular range into which the correct factor falls is provided at the input.

The ambiguity problem is now resolved by identifying the correct factor from candidate scaling factors larger than 0.5 through evaluation and thresholding of PCE using the provided camera fingerprint. Furthermore, filtering phase was modified to 4th order differencing from the 2nd order differencing, thus reducing cases when a false peak lead to an incorrect result. Correctness of the estimate is verified through PRNU matching with cameras fingerprints. This solution was also implemented in the Camera Verification System and was a part of the at home evaluation.

### 3.3.3. Semi-blind resampling estimation method

When performing camera verification task, the camera fingerprint estimated from a set of reference images is matched to the traces of PRNU detected in the probe image. Whenever the probe image has been resized with an unknown scaling factor, the pixel-wise resynchronization such as reversal of resizing needs to be performed. Blind methods of geometric transform parameters estimation refer to the case when the original image is not available for image content registration. In the camera verification scenario, the camera sensor pattern noise (SPN) and JPEG artifacts can serve as a side information to ease the parameters recovery. This is the idea of the semi-blind method developed under this project. The method even relaxes the requirement for the camera fingerprint assuming availability of a single reference image.

JPEG dimples found in about 60% of images in JPEG format in combination with the more universal linear pattern can be detected as peaks in frequency domain. The new developed semi-blind method executes the image registration without detecting the peaks. Image resizing becomes cropping in frequency domain, allowing for scaling estimation via pattern registration in Fourier domain. Two such patterns, each obtained from a different image are transformed using FFT and the relative scaling factor is found through cross-correlation. This method

demonstrated its power in PRNU estimation from scaling inconsistent training sets of reference images. This topic has been published at IS&T 2020 conference [SB].

### 3.3.4. Fingerprint quantization

Scaling up camera verification systems requires handling and storing large sets of camera fingerprints. Efficient representation of fingerprint data becomes important. Single or even double precision data for large fingerprints that meet the image/video resolution of the camera output is excessive and unnecessary. In another extreme, single bit per sample quantization as proposed by Bayram *et al.* [FB] comes with a significant loss in the correlation detector performance because the SNR of quantized uniformly distributed data is proportional to the number of bits $b$ per sample according to the formula, SNR = 6.02$b$ (dB).

For Gaussian pdf – a good model for camera fingerprints – a nonuniform quantizer can be found numerically using Lloyd's algorithm that outperforms uniform quantizers for small $b$ = 2, 3, or 4. Its disadvantage is the need for communicating reconstruction levels for proper decoding. For a large b and a continuous pdf, the optimal uniform quantizer achieves practically the same SNR as nonuniform quantizers.

The problem of optimal quantization is usually considered w.r.t. mean-square error (MSE) between the quantized data and the original non-quantized data. In the context of camera fingerprints, we consider data represented with double precision as non-quantized data since the quantization in double precision is extremely fine-grained ($b$ = 64; SNR > 300). Rather than in MSE, we are interested in how fingerprint quantization affects correlation-based detection of the fingerprint in images, evaluated in terms of receiver operating characteristic curve (ROC). We verified experimentally that a uniform quantizer that minimizes MSE also maximizes (up to a negligible difference) correlation between the quantized data and the nonquantized Gaussian data.

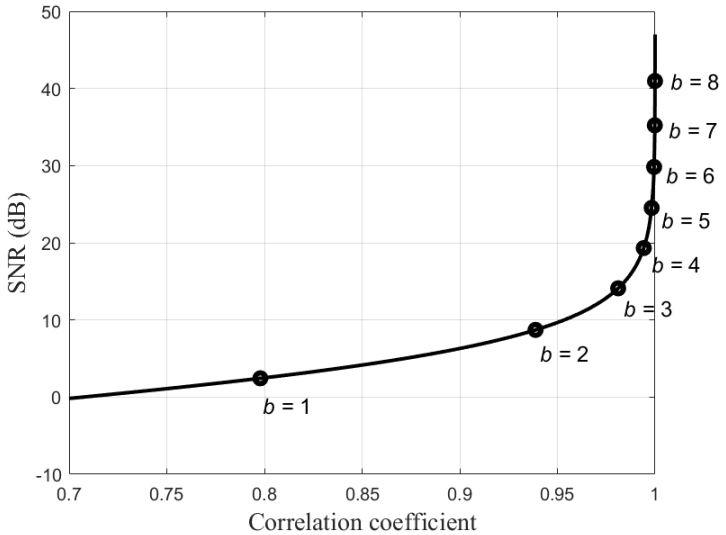



**Figure 5. Performance of optimal uniform quantizers for Gaussian data to be represented by *b* bits per sample.**

The proposed solution has two simple steps: normalize the fingerprint F to zero sample mean and unit sample variance. Then map it to 8-bit integers within 0 and 255 by rounding and clipping of ($a$F+127.5) using uint8 function. Solving the simple optimization task, the optimal parameter

$a$ was determined as $a \approx 32.5$, yielding correlation coefficient between the quantized and no-quantized fingerprint equal to 0.99996 and SNR above 40. These values are independent of the fingerprint size and guarantee that correlation properties of the fingerprint are preserved by this quantization. Whichever value for parameter $a$ is used, no decoding needs to be performed to run the camera verification. The integer data can be cast to single or double precision just before computing correlation or PCE statistics. In case of $b = 8$, the quantized fingerprints can be stored as 8-bit raster images, which allows for an easy cross-platform exchange of fingerprint data. SNR and corresponding correlation coefficients of this uniform quantizer for the number of bits between 1 and 8 are denoted in **Figure 5**.

The fingerprint quantization plays its role in PRNU estimation using custom designed CNN. This area of research is elaborated upon in the next subsection.

### 3.3.5. PRNU estimation using CNN

Having compared existing methods of PRNU estimation from single images in terms of quality and computation speed, an image denoising method using CNN called FFDNet was adopted for PRNU estimation. Its performance is comparable to the best of the previously existing methods. It offers the speed of CNN namely on GPUs and opens the new possibilities for PRNU estimation and its applications. This was the first step towards a specialized CNN-based methods of PRNU estimation that has been introduced during this project. Ongoing work is being wrapped in a journal publication to be submitted to IEEE TIFS.

## 3.4 Steganalysis of JPEG images

The JPEG image format is the most ubiquitous format because of its convenience; it offers high visual quality with compact file sizes. Most steganography applications can embed in this file format, and they do so by slightly modifying the Discrete Cosine Transform (DCT) coefficients. Detecting this specific type of image alteration amounts to detecting very weak signals that are unknown to the steganalyst because the signal is modulated by the message itself and a secret stego key.

Recently, detectors built as deep convolutional neural networks have firmly established themselves as superior to the previous detection paradigm – classifiers based on rich media models [SRM,JRM,SCRM]. Existing network architectures [XuNet,JXuNet,YeNet,PhANet], however, still contain elements designed by hand, such as fixed or constrained convolutional kernels, heuristic initialization of kernels, the thresholded linear unit that mimics truncation in rich models, quantization of feature maps, and awareness of JPEG phase. This makes them suboptimal and overly specialized to a specific domain.

### 3.4.1. SRNet: steganalysis with a deep residual convolutional network

During this project, a novel deep residual architecture called SRNet (Steganalysis Residual Network) has been designed to minimize the use of heuristics and externally enforced elements that is universal in the sense that it provides state-of-the-art detection accuracy for both spatial-domain and JPEG steganography. The key part of the proposed architecture is a significantly expanded front part of the detector that "computes noise residuals" in which pooling has been disabled to prevent suppression of the stego signal. The most significant improvement was observed in the JPEG domain. For a known embedding algorithm (stego scheme), further performance boost can be achieved by supplying the selection channel (the embedding change

probabilities) as a second channel to the SRNet. The details of this architecture, together with a comprehensive evaluation on standard image datasets appear in [SRNet].

SRNet is the first steganalysis network that is free of externally introduced design elements previously proposed specifically for steganalysis and forensics, such as constrained kernels, initialization with heuristic kernels, thresholding, quantization, and awareness of JPEG phase. Consequently, SRNet can be trained in an end-to-end fashion from randomly initialized convolutional kernels and in the same fashion independently of the embedding domain. The front part of SRNet contains seven residual layers in which pooling has been disabled to allow the network to learn relevant "noise residuals" for different types of embedding changes in both spatial and JPEG domain. The design of SRNet has been validated experimentally on standard datasets and both modern and older steganographic algorithms. State-of-the-art detection is observed in both domains with rather significant improvements in the JPEG domain. Receiver operating characteristics for selected combinations of embedding algorithms and payloads reveal especially favorable detection performance for very low false-alarm rates, which is expected to be significant for practitioners.

While SRNet was intentionally designed to minimize the use of heuristic design elements specific to steganalysis, it still benefits from being informed about the probabilistic impact of embedding in the form of the selection channel. It is also the first steganalysis network that makes use of the selection channel for JPEG domain steganalysis, a task that was achieved by adding a bound on embedding distortion to the feature maps outputted by the first layer to reinforce the output of neurons that are most affected by embedding.

Since steganalysis detectors by definition detect inconsistencies in the noise patterns of images, they often find applications in forensics. In particular, it is speculated that it may be especially useful for detecting inconsistencies within a single image to identify locally manipulated regions and for detecting camera models.

### 3.4.2. Detection of diversified stego sources

An important aspect of practical detectors is to be general (capable of detecting a variety of embedding algorithms) and possibly identify the steganographic method. For this task, the PI has explored binary classifiers trained as cover versus all stego, multi-class detectors, and bucket detectors in a feature space obtained as a concatenation of features extracted by CNNs trained on individual stego algorithms. Both content-adaptive and non-adaptive steganographic algorithms were included in the investigation.

The accuracy of the detector to identify steganography was compared with detectors trained for a specific embedding algorithm. The best detector was a multi-class SRNet. Its loss function had to be adjusted to control the false alarms or missed detection of selected stego schemes. When trained on seven embedding algorithms, this multi-class detector was able to reliably classify the stego algorithm, while its ability to detect steganographic content decreased only marginally w.r.t. binary CNN detectors dedicated (and tested) on a specific embedding algorithm.

While the multi-class SRNet was able to "contain" the complexity of the diversified stego source in the sense that it provided detection of steganography comparable to that of dedicated detectors, it appeared to struggle to recognize previously unseen steganographic methods. The PI conjectures that this could be remedied by significantly increasing the number of stego algorithm in the training set and by using models pre-trained on large datasets, such as the ImageNet [ImageNet].

### 3.4.3. Reference channels for steganalysis

When available, reference signals may dramatically improve the accuracy of steganalysis. Particularly powerful reference signals are embedding invariants that exist when the steganographic algorithm swaps values from small disjoint subsets of the cover elements' dynamic range, such as, but not limited to, embedding schemes utilizing Least Significant Bit (LSB) replacement. Such embedding paradigm is more typical of older steganographic schemes and is by far the most prevalent in commercially or freely available stego applications. Steganography that uses such embedding operations allows construction of reference signals that are close to the input image but are not sensitive to embedding modifications.

The PI considered more general type of embedding operations than replacement of LSBs by showing how the reference signal should be prepared for generic LSB flipper in the JPEG domain (LSBF), example of which is Jsteg [Jsteg], OutGuess [OG], and Model-Based Steganography [MBS] (MBS). Second, the PI introduced a method how this reference should be used within the novel detection paradigm – deep convolutional neural networks. Since the network learns the image representation as well as the steganalysis classification jointly, the reference image in inputted as a second input channel to the network. The SRNet is especially suitable for this extension because all its filters are randomly initialized, which allows the network to discover a way to incorporate the second channel via data driven end-to-end training.

The PI developed a general method how to prepare such reference signals for a certain type of embedding operations, and incorporate them in detectors built as convolutional neural networks to improve their detection accuracy. The beneficial effect of reference signals is especially apparent in the JPEG domain, on Model-Based Steganography (MBS) and a generic LSB flipper (LSBF) with and without stochastic restoration of the histogram (OutGuess).

### 3.4.4. ALASKA I evaluation

This section describes the architecture and training of detectors developed for the ALASKA steganalysis challenge, an independent evaluation of the detectors developed during reported effort.

The purpose of the ALASKA evaluation was to abandon typical sandboxed laboratory setting and force researchers to face more realistic conditions that are closer to what a steganalyst might have to deal with in real life. The reference [A2] contains the full details of the evaluation setup and interpretation of the final results across all competing teams. The teams were given a set of 5,000 JPEG images, some of which were cover images and some embedded with secrets. This set is called 'ALASKArank' because the detection results achieved on this set determined the ranking of the competing teams. Four JPEG steganographic schemes were used to produce the stego images: J-UNIWARD [UNI], UED-JC [UED], EBS [EBS], and nsF5 [nsF5] with priors 0.4, 0.3, 0.15, and 0.15, respectively, according to the embedding script shared by the organizers. All four embedding methods were adjusted to hide in color JPEG files by embedding in chrominance channels a fraction of the payload determined by the JPEG quality factor. The size of the embedded payload was determined by the cover image development history (starting with a RAW sensor capture), which was again randomized. It involved four different choices for demosaicking, resizing by a randomly selected factor in the range [0.6,1.3], a version ofsource-preserving cropping to A×B pixels with A, B ∈ {512,640,720,1024}, sharpening, denoising, and micro-contrast enhancement whose parameters were again randomized, and final JPEG compression with quality factor between 60 and 100 selected at random according to a prior that the organizers computed by analyzing a large number of JPEG images uploaded to the image sharing portal Flickr. The payload w.r.t. image size was scaled according to the square root law

[SRL] to obtain an approximately constant statistical detectability across different sizes. The smallest and largest sizes were 512×512 and 1024×1024, respectively.

The embedding code for all four steganographic schemes was given to the participants as was the script for developing a RAW image to a JPEG cover. This allowed the participants to generate their training sets.

The information that was *not* revealed to the competitors included:

1) The percentage of stego images in ALASKArank and per each quality factor.
2) The priors for all four stego schemes per each quality factor, possibly thus introducing an unknown stego-source mismatch.
3) The source of RAW images for ALASKArank, possibly thus creating a cover-source mismatch.

The competitors were permitted one submission per four hours per team. The submission was a text file with file names from ALASKArank ordered from the most likely stego to the least likely stego. This allowed the organizers to draw an ROC curve and report three quantities on ALASKA leaderboard: the missed detection rate at 5% false alarm, MD5, the minimum average total error under equal priors, $P_E$, and the false-alarm rate for 50% detection, FA50.

The approach chosen by the Binghamton team was a natural progression of the research on deep learning architectures for steganalysis funded by this project. When facing a multitude of embedding schemes, it is better to train a multi-class detector than a binary one-against-all classifier. The detectors had to be trained via payload curriculum learning on double payload first since training directly on the base payload would not always produce good detectors or convergent networks.

The SRNet was also modified to accept multiple-channel input and several versions were trained by separating the color channels, which turned out the key ingredient that allowed further improvement of detectors beyond their initial "off-the-shelf" form. This provided a significant boost over simply training a three-channel SRNet. This may be due to the way the channels were merged in the network.

While a separate detector was trained for each quality factor, this is not a scalable approach to cover, for example, non-standard quantization tables. Scalability was successfully resolved by a work that followed the competition [QFscale, QFsec]. For each quality factor in the range 60–98, several multi-class tile detectors implemented as SRNets were trained on various combinations of three input channels: luminance Y and two chrominance channels, Cr and Cb. To accept images of arbitrary size, the detector for each quality factor was a multi-class multi-layered perceptron trained on features (statistical moments) extracted by the tile detectors [ArbS]. For quality 99 and 100, a new "reverse JPEG compatibility attack" was developed and also implemented using the SRNet via the tile detector.

### 3.4.5. The reverse JPEG compatibility attack

This attack was discovered during the ALASKA I evaluation. It is a novel steganalysis method for JPEG images that is universal in the sense that it reliably detects any type of steganography as well as small payloads. It is limited to quality factors 99 and 100. The detection statistic is formed from the rounding errors in the spatial domain after decompressing the JPEG image. The attack works whenever, during compression, the discrete cosine transform is applied to integer-valued signal. Reminiscent of the well-established JPEG compatibility steganalysis, it is called the "reverse JPEG compatibility attack." While the attack can be analyzed under simplifying assumptions using reasoning based on statistical signal detection, the best detection in practice

is obtained with machine learning tools.

The main idea stems from the observation that, when decompressing a JPEG image, the rounding errors in the spatial domain exhibit a Gaussian distribution with variance 1/12 folded to $[-1/2,1/2)$ . Steganographic embedding changes made to quantized DCT coefficients increase the variance of the Gaussian distribution, allowing thus an extremely accurate detection. The attack is fundamentally possible due to the fact that the DCT is applied to integers.

The attack has been tested on five different embedding schemes, grayscale and color images, and diverse stego sources (the ALASKA I dataset). It gave the Binghamton team a substantial advantage. The attack is universal in the sense that a detector trained on one embedding algorithm generalizes to unseen embedding methods. It is also robust to various JPEG compressors. Moreover, detection of a specific embedding algorithm can be improved, especially for quality factor 99, by providing rounding errors together with decompressed image as a second channel to the network detector. Extension of this attack to double-compressed images appears in [DC].

To circumvent the attack, one needs to avoid applying the DCT to integer-valued images, which, however, none of the JPEG compressors known to the authors do. The second possibility to reduce the detectability is to use side-informed embedding schemes that minimize the combined distortion due to quantization and embedding. They, indeed, are less detectable than non-side-informed schemes. Our experiments showed that SI-UNIWARD on payload of 0.05 bpnzac essentially eluded detection. Thus, besides drastically reducing the payload, it currently appears that quality 100 and 99 JPEGs should be avoided for steganography by the same token as decompressed JPEGs should not be used for spatial-domain embedding.

### 3.4.6. ALASKA II evaluation: pre-trained CNN models for steganalysis

Work in progress based on [A2].

## 4.0 RESULTS AND DISCUSSION

### 4.1 Vectorized context-aware pixel descriptors for manipulation detection

#### 4.1.1 PRNU-based image manipulation localization with discriminative random fields

We have participated in the NC2017 and MFC2018 evaluations with overall positive outcomes. In addition, we have performed a self-evaluation on the "Realistic Tampering Image Dataset" [MSA], where we were able to demonstrate considerable improvements over prior work.

In the MediFor program evaluations, we have pursued an aggressive opt-out strategy, ignoring all images for which an initial camera identification failed (peak-to-correlation energy, PCE<60), as well as all images that did not have the same size as the candidate camera fingerprint (TRR 0.13 for MFC2018). Strong post-processing of the manipulation map helped to reduce false alarms. Our localization performance averaged to a trMCC of 0.202 (compared to 0.185 for NC2017). The example in **Figure 6** shows one positive outcome. While image-level detection was not our primary focus, we were still able to achieve a satisfactory trAUC of 0.68 on MFC2018 (compared to 0.85 for NC2017).

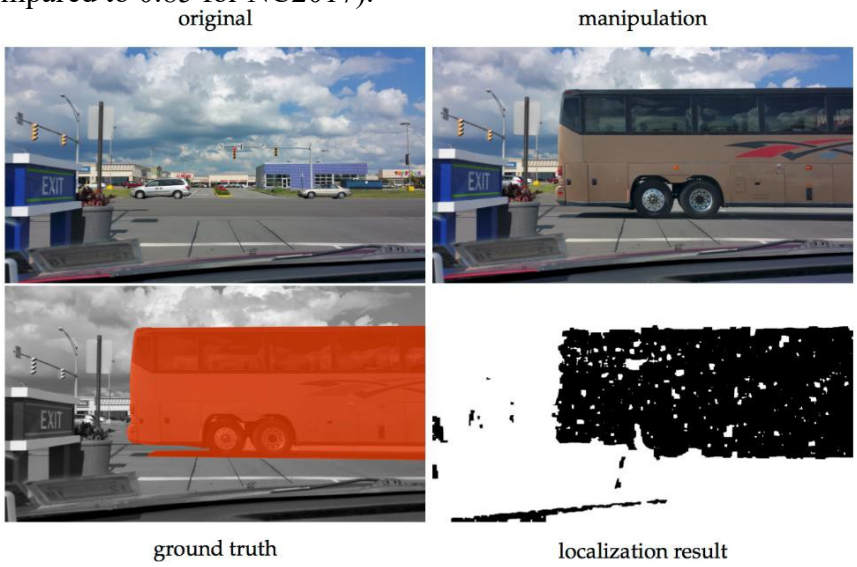



**Figure 6. Example output from the MFC2018 evaluation of our DRF-based manipulation localization algorithm (probe image: 34d5b4943b72f97eec56601cf07ef9f2).**

In comparison to the manipulations in the Realistic Tampering Dataset [MSA], the MediFor evaluation data was generally much more challenging. Among other factors, on main driver may be strong JPEG compression in the MediFor data. On the uncompressed Realistic Tampering Dataset, we were able to achieve an average MCC of around 0.6.

#### 4.1.2 Unsupervised image manipulation localization with non-binary label attribution

Our participation in the MFC2018 evaluation had positive outcomes. Additionally, we have performed a self-evaluation on the NIST Nimble dataset (64 high-quality JPEG images with splice manipulations) and 180 uncompressed images from the Columbia database [DIS].

On MFC2018, our main focus was on the localization of image manipulations. No particular

effort went into the design of a meaningful image-level score. We have pursued a fairly inclusive opt-out strategy, ignoring only unreadable file formats and images violating compute memory constraint. The resulting TRR was 0.9. The resulting trMCC score averaged to a (in relative terms) favorable 0.211.

**Table 1. Average opt-MCC and opt-F1 scores for SpliceBuster (SB) and n-ary hierarchical clustering (HCl-n) for different window sizes w with stride s.**

| dataset | | SB | | HCl-2 | | HCl-3 | | HCl-4 | | HCl-5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $w = 65$ | $w = 129$ | $w = 65$ | $w = 129$ | $w = 65$ | $w = 129$ | $w = 65$ | $w = 129$ | $w = 65$ | $w = 129$ |
| Columbia $(s = 1)$ | MCC | 0.683 | 0.769 | 0.638 | 0.480 | 0.776 | 0.605 | 0.813 | 0.687 | **0.814** | 0.745 |
| Nimble $(s = 4)$ | MCC | 0.588 | 0.625 | 0.630 | 0.590 | 0.635 | **0.661** | 0.612 | 0.637 | 0.555 | 0.580 |
| Columbia $(s = 1)$ | $F_1$ | 0.797 | 0.851 | 0.750 | 0.633 | 0.839 | 0.668 | **0.859** | 0.761 | 0.858 | 0.801 |
| Nimble $(s = 4)$ | $F_1$ | 0.618 | 0.646 | 0.641 | 0.594 | 0.634 | **0.662** | 0.608 | 0.636 | 0.552 | 0.580 |

**Table 1** reports quantitative results on the two other databases. Notably, HCl-n outperforms SpliceBuster with $w = 65$ on the Columbia images, while $w = 129$ is more favorable for the Nimble data. We surmise that this is linked to the ability to form meaningful clusters in relation to the image size. Taking also opt-F1 into account, we conclude that a ternary attribution provides a good overall setup, although HCl-5 ultimately performs best on the Columbia data.

**Figure 2** (in Section 3) showcases qualitative results for some of the Nimble images. It is worth pointing out that the HCl-3 ternary maps will very frequently reveal a non-genuine region as one of their final clusters, even in the last example of **Figure 2** where the heat map fails to reveal the manipulation as clearly as SpliceBuster. We support this claim by computing MCC scores directly from the ternary maps with the help of side information: iterating over all three possibilities, we consider one of the clusters as non-genuine at a time to obtain the binary map that gives the highest score. The observed MCC scores average, well above the numbers reported in **Table 1**, to 0.69 for the Nimble dataset and underscore the benefits of a shift of perspective away from bimodal to more versatile models.

### 4.2 Detection of processing history

The accuracy of the multi-class detector (average of the diagonal terms of the confusion matrix) went from 89.3% for the maximum-likelihood detector to 95.2% for the CNN detector. Both detectors are described in detail, including their self-evaluation, in [PH1] and [PH2]. The architecture is shown in **Figure 7**.

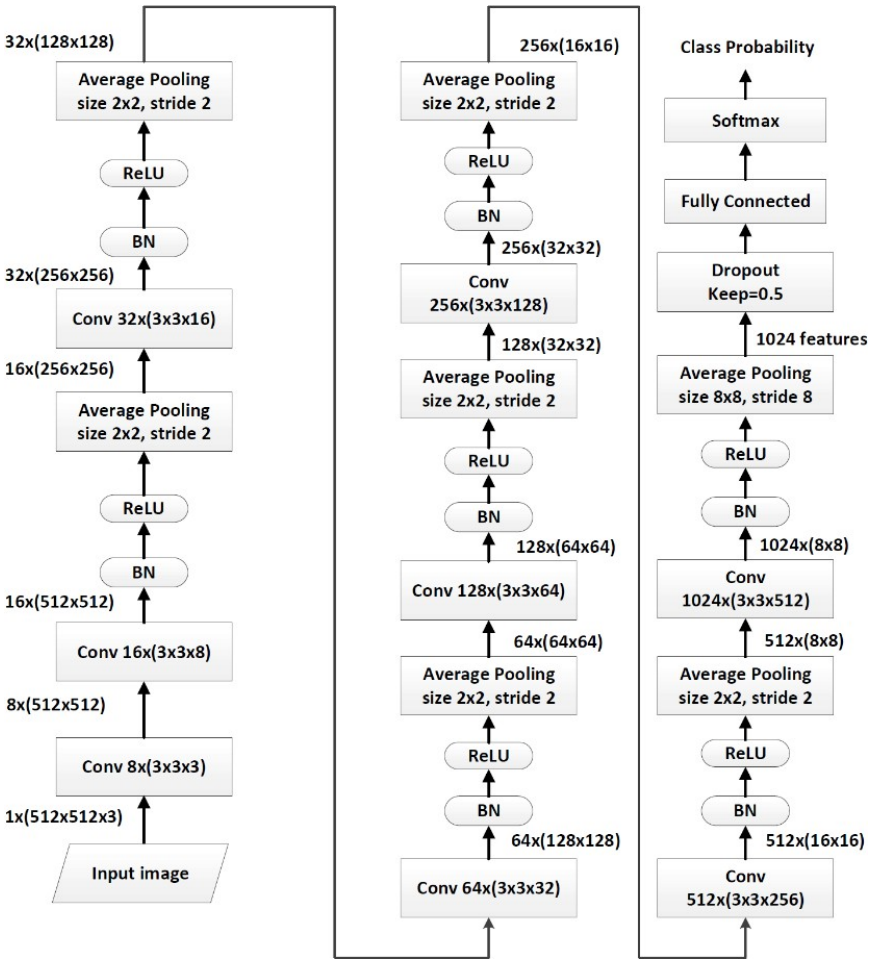



**Figure 7. Convolutional neural network architecture for detecting processing history.**

## 4.3 Media source identification and verification

### 4.3.1. Camera verification for HDR images

Evaluation of the camera verification for HDR images was performed on UNIFI dataset of HDR images provided by the research group at University of Florence, Italy [HDRset], expanded by three more cameras, a second Xiaomi 3 and two Xiaomi Note 4 devices. The full list of devices with the image resolution and acronyms that match those in **Figure 8** can be found in the resulting publication [HDR]. In **Figure 8**, the dashed line represents a detection threshold. Circles below this line mean that most images from that device could not be correctly matched to the source camera without the special handling developed under this MediFor project. While images from some cameras, namely A01, A02, A03, A12, A13, A17, and A19, tend to slip through the standard matching process unidentified, i.e. missed detected, camera verification for images from all tested Apple devices (I01-I06) does not suffer from HDR processing. The revealed reason is that these images are not cropped and up-sampled, like images from other cameras in the test, except Samsung Galaxy models A07-A10. Samsung apparently uses a different technology and possibly hardware for HDR than other manufactures included in this test set. No scaling and no local shift was detected for any of the Samsung devices.

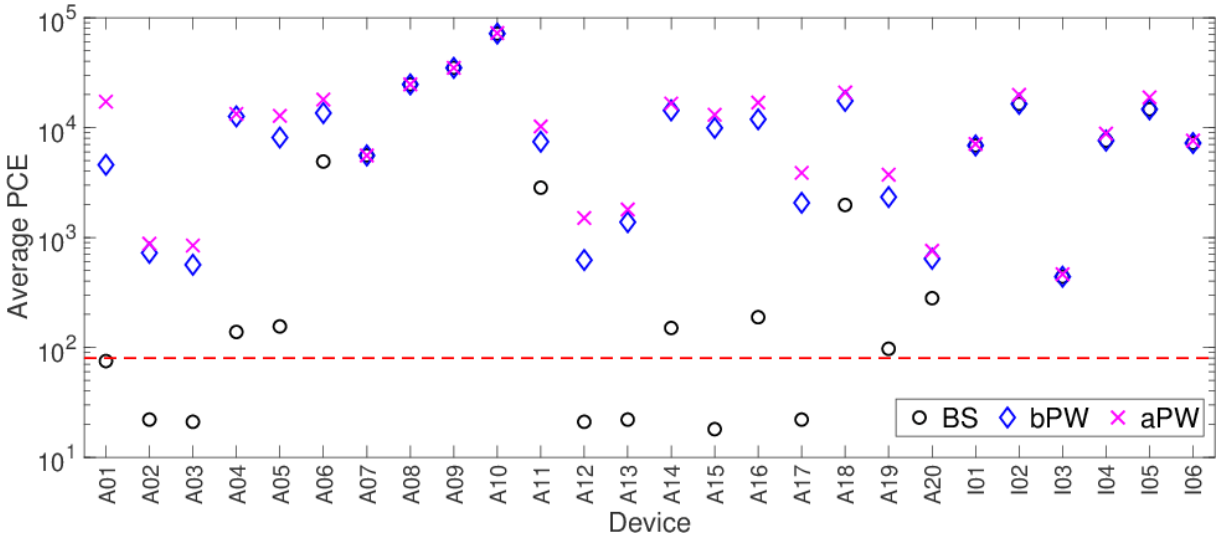



**Figure 8. Average PCE for HDR test images from UNIFI dataset before scaling inverted (BS), before patchwork (bPW) and after patchwork (aPW).**

**Table 2** brings yet more insight into the device dependent results. The subsets of Samsung Galaxy and Apple devices are separated in this table by horizontal lines. While correction for scaling and cropping was almost always sufficient for correct camera verification, the PW step is what reveals the "signature" of HDR processing and prevent misinterpretation of a lack of PRNU match in parts of the image as image tampering in many image manipulation detection applications.

**Table 2. Positive detection rates and average PCE detection statistic before and after reversal of geometric transformations for HDR images from the expanded UNIFI dataset. (PCE is averaged over the tests on *N* probe images.)**

| | | Positive Det Rate | | | PCE | | |
|---|---|---|---|---|---|---|---|
| Device | *N* | BS | bPW | aPW | BS | bPW | aPW |
| A01 | 20 | 0.25 | 0.95 | 1 | 75 | 4581 | 17196 |
| A02 | 24 | 0 | 0.96 | 0.96 | 22 | 724 | 877 |
| A03 | 20 | 0.05 | 0.45 | 0.45 | 21 | 564 | 844 |
| A04 | 15 | 0.40 | 1 | 1 | 138 | 12628 | 13304 |
| A05 | 24 | 0.58 | 1 | 1 | 155 | 8114 | 12769 |
| A06 | 24 | 0.25 | 0.96 | 1 | 4906 | 13545 | 17979 |
| A07 | 21 | 1 | 1 | 1 | 5574 | 5574 | 5575 |
| A08 | 24 | 1 | 1 | 1 | 24691 | 24691 | 24689 |
| A09 | 24 | 1 | 1 | 1 | 34927 | 34927 | 34966 |
| A10 | 24 | 1 | 1 | 1 | 71608 | 71608 | 72345 |
| A11 | 21 | 0.62 | 1 | 1 | 2836 | 7427 | 10227 |
| A12 | 20 | 0 | 0.70 | 0.90 | 21 | 623 | 1503 |
| A13 | 20 | 0 | 0.95 | 1 | 22 | 1379 | 1796 |
| A14 | 24 | 0.21 | 1 | 1 | 150 | 14343 | 16555 |
| A15 | 24 | 0 | 1 | 1 | 18 | 9913 | 13065 |
| A16 | 24 | 0.29 | 1 | 1 | 188 | 11883 | 16879 |
| A17 | 24 | 0 | 1 | 1 | 22 | 2067 | 3870 |
| A18 | 24 | 0.17 | 1 | 1 | 1975 | 17494 | 20844 |
| A19 | 24 | 0.42 | 0.96 | 0.96 | 97 | 2332 | 3724 |
| A20 | 24 | 0.38 | 0.92 | 0.92 | 280 | 640 | 755 |
| I01 | 15 | 1 | 1 | 1 | 6868 | 6868 | 7088 |
| I02 | 19 | 1 | 1 | 1 | 16420 | 16420 | 19853 |
| I03 | 24 | 1 | 1 | 1 | 439 | 439 | 463 |
| I04 | 24 | 1 | 1 | 1 | 7608 | 7608 | 8814 |
| I05 | 21 | 1 | 1 | 1 | 14665 | 14665 | 18825 |
| I06 | 24 | 1 | 1 | 1 | 7225 | 7225 | 7601 |

The overall improvement over the "HDR-unaware" camera identification on 16 cameras in the UNIFI HDR dataset is highlighted in **Figure 9**. Samsung Galaxy and Apple devices are not included here because their HDR images do not impose the same challenge and therefore the presented method does not need to be applied to them.

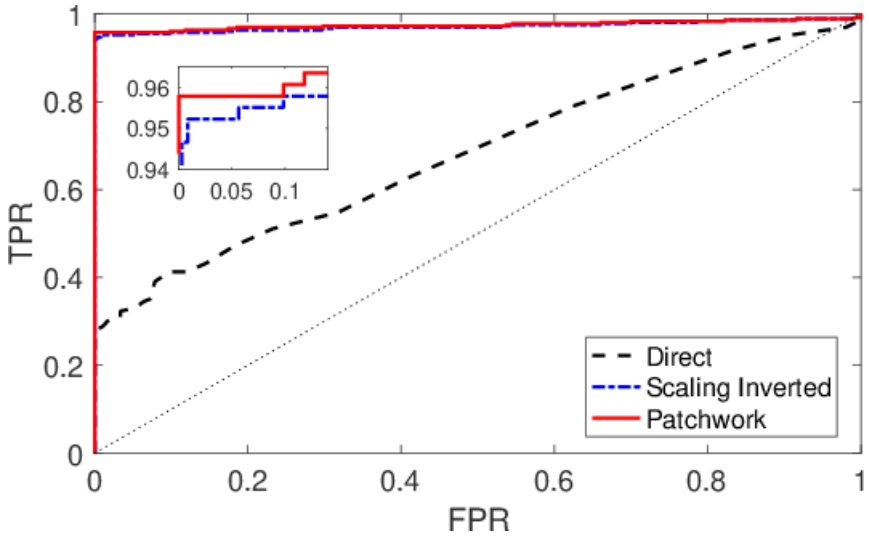



**Figure 9. ROC from tests on extended UNIFI dataset.**

### 4.3.2. Camera verification for digitally zoomed images

This special case of in-camera processed images was getting attention for two reasons. First, it has become much more frequent with increasing resolution of CMOS sensors in consumer cameras including mobile devices. Second, the brute force search based methods proposed in the past become slower as the digital zoom range increases. In the annual evaluation setting, a fast method was needed.

The improved peak analysis method is fast thanks to the Fast Fourier Transform (FFT). However, it cannot replace the brute force search when one of seven special scaling factors is involved, particularly one of $8/n$, $n = 1,2,\ldots,7$. As a consequence, once the scaling parameter is estimated, the probe image must be down-sampled with the estimated factor and correctness verified via PRNU matching process.

Experiments on images from three Xiaomi Redmi phone cameras in **Table 3** demonstrate both the precision of scaling factor estimation and the limitation. Whenever the parameter is correctly determined (which is vast majority of cases) this method reduces processing time, otherwise the slower brute force search can be deployed. This solution was implemented in the Camera Verification System evaluated on MFC20, from which the ROC is depicted in **Figure** 10.

**Table 3. Estimated upscaling factors in in-camera digitally zoomed images. Incorrect estimates are in red and blue italic font. Correct estimates are in higher precision than the ground truth recorded in the image's EXIF. This method cannot work for factors 8/*n*, *n* = 1, 2, …, 7, including 8/7 ≈ 1.15 and 8/3 ≈ 2.66 (highlighted in italic font).**

| | Redmi_C1 | | Redmi_C2 | | Redmi_C3 | |
|---|---|---|---|---|---|---|
| Image no. | EXIF | Estimated | EXIF | Estimated | EXIF | Estimated |
| 1 | 2 | 1.996 | 1.51 | 1.5088 | *2.66* | *2.004* |
| 2 | 2 | 1.996 | 1.51 | 1.0952 | *2.66* | *2.004* |
| 3 | 2.88 | 2.8746 | 1.51 | 1.5088 | *2.66* | *2.004* |
| 4 | 2.88 | 2.8746 | 1.55 | 1.5477 | 3.31 | 3.3084 |
| 5 | 1.64 | 1.6388 | 1.55 | 1.5477 | 3.31 | 3.3084 |
| 6 | 1.64 | 1.6388 | 1.55 | 1.5477 | 3.31 | 3.3084 |
| 7 | 1.64 | 1.6388 | 2.51 | 2.5054 | 4.32 | 4.3223 |
| 8 | 1.64 | 1.6388 | 2.51 | 2.5054 | 4.32 | 4.3223 |
| 9 | 2.6 | *4.3663* | 2.51 | 2.5054 | 4.32 | 4.3223 |
| 10 | 2.6 | *4.3663* | 3.47 | *12.8376* | 5.75 | 5.7548 |
| 11 | 2.6 | *4.3663* | 3.47 | *26.5841* | 5.75 | 5.7548 |
| 12 | 3.67 | 3.6724 | 3.47 | *12.8376* | 5.75 | 5.7548 |
| 13 | 3.67 | 3.6724 | 4.52 | 4.5241 | 5.75 | 5.7548 |
| 14 | 3.67 | 3.6724 | 4.52 | 4.5241 | 5.75 | 5.7548 |
| 15 | 5.14 | 5.135 | 4.52 | 4.5241 | *1.15* | *1.7733* |
| 16 | 5.14 | 5.135 | 5.62 | 5.6255 | *1.15* | *1.7733* |
| 17 | 5.14 | 5.135 | 5.62 | 5.6255 | | |
| 18 | 2.24 | 2.235 | 6.55 | 6.5447 | | |
| 19 | 2.24 | 2.235 | 6.55 | 6.5447 | | |

Overall system evaluation on the large MFC20 image dataset had its specifics that potential forensic investigator has to be aware of. Important facts that influenced the evaluation include availability of large numbers of reference images from which each camera fingerprint was computed. In many real life scenarios there may be only a handful of such images available. This is what made the camera verification task possibly easier than one could expect in most real forensic settings. On the other hand, much higher risk of missed detection than due to weak PRNU in the camera fingerprint comes from geometric transformations applied to the probe images. For example, significant downsampling with image cropping, namely when its parameters fall outside the expected range, can easily go undetected.

While the resulted ROC pictured in **Figure 10** was the best among all participated teams, there is a lot to be desired in terms of detection rate. The main weakness of the implemented system is incompleteness in terms of inclusion of brute force search for scaling factors in downsized images accompanied by cropping. Uncertainty in what would be the time limit for processing one probe image was the reason of not including a slow but powerful brute force search. Until now, all alternative methods suffer from higher missed detection. Thus, in real life setting, running the exhaustive search when other attempts for fingerprint matching fail, would yield significantly better detection rate at very low false detection than the ROC in **Figure 10** suggests.

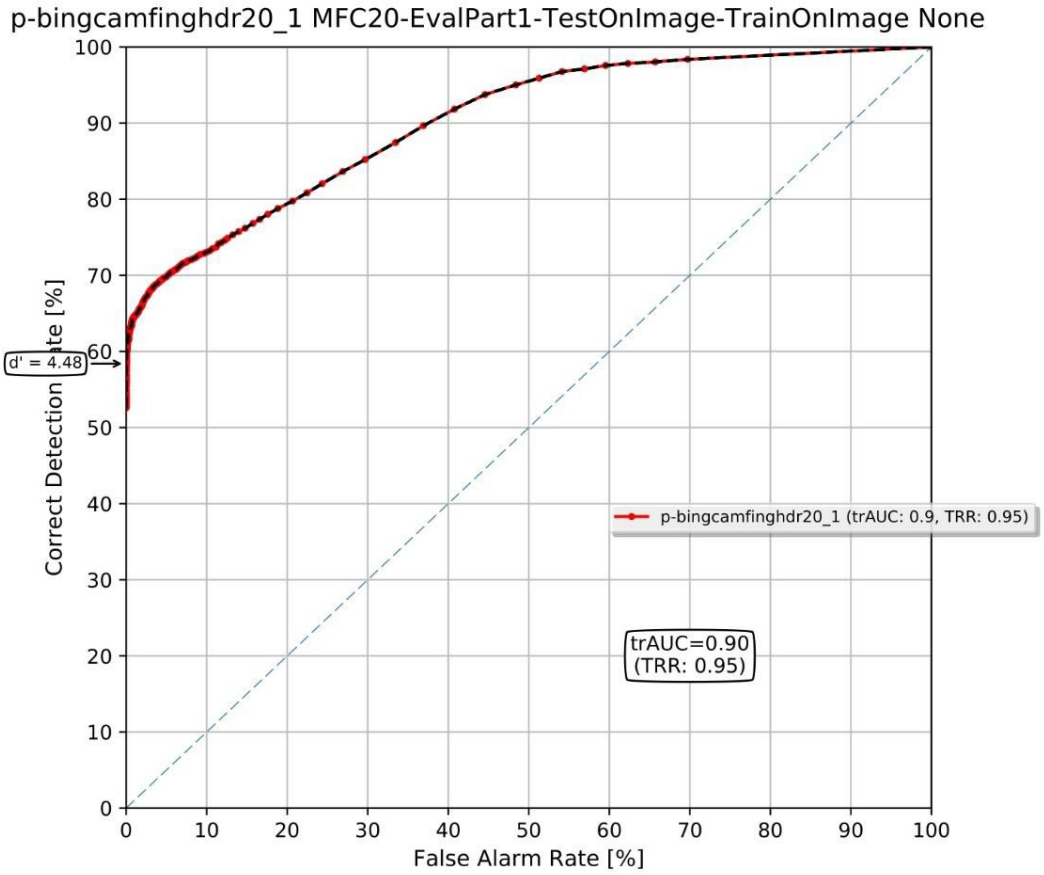



**Figure 10. Evaluation of camera verification on MFC20. Trained on images, tested on images.**

### 4.3.3. Semi-blind resampling estimation method

The new dimples-aware method (LPD method) was extensively tested on the extended UNIFI dataset of 26 cameras. Robustness to additional JPEG compression was evaluated for quality factor QF2 ranging from 100 to 60 and scaling factors between 0.7 and 1.3. [SB] shows resulting success rates for all cameras w.r.t. the scaling factor and **Table 5** w.r.t. the additional JPEG compression. The success is counted when the estimation error is below 0.002. Blue shading in tables highlights the varying success rates.

The final JPEG compression (down to QF2=65) does not negatively affect performance. There are other yet not determined factors that strongly influence this estimation method. Almost ideal performance was recorded for all images from Gionee S55 camera. At the other end of the performance spectrum lies iPhone 8.

**Table 4. Success rates (%) of the LPD method averaged over the list of JPEG quality factors QF2 = [100,98,96,94,92,90,85,80,75,70,65,60].**

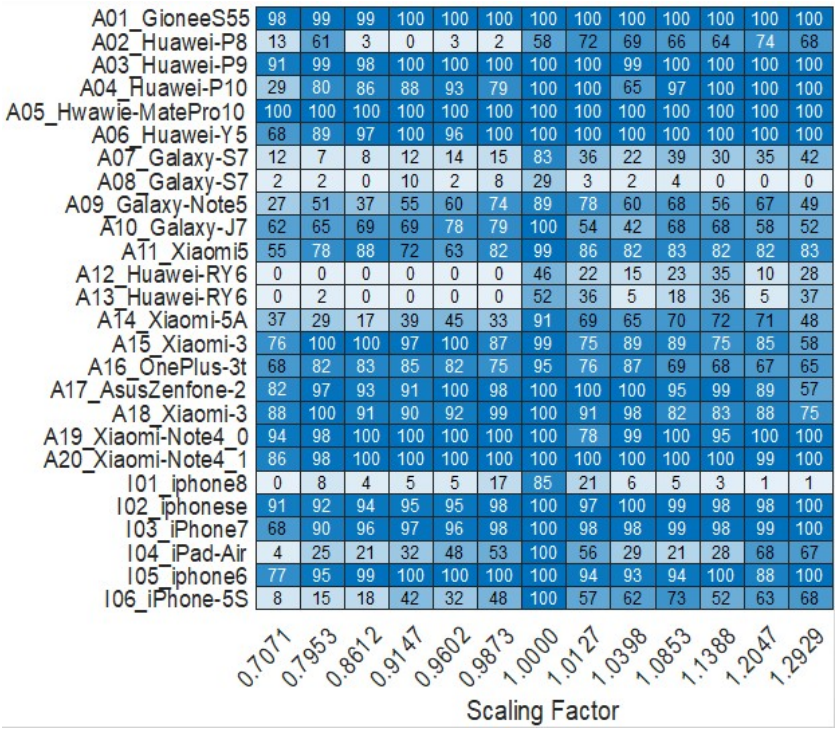
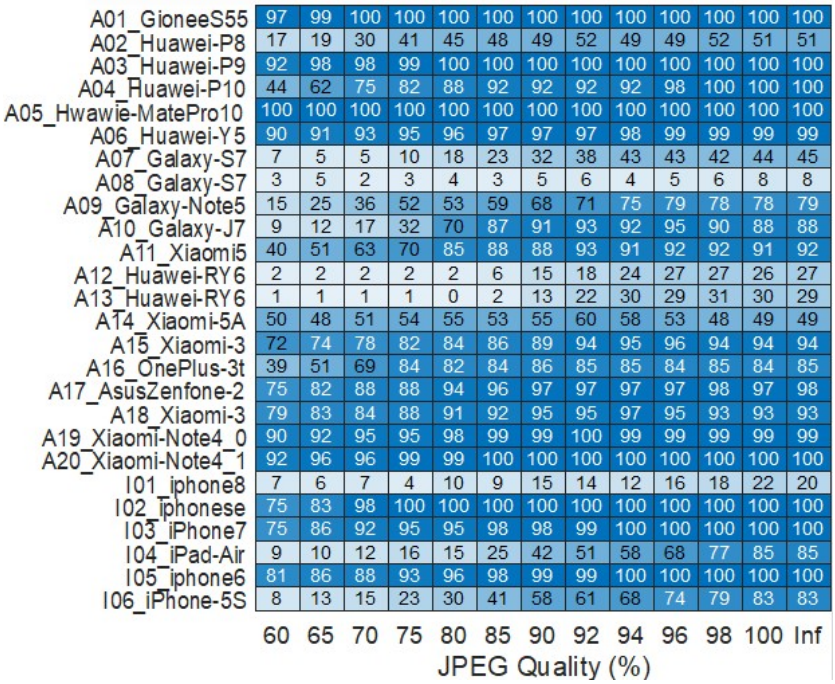



**Table 5. Success rates (%) of the LPD method vs. JPEG quality factor, averaged over all scaling factors.**




Further tests aim at detection of scaling applied to uncompressed images that were initially JPEG compressed with quality factor QF1 before resizing. For this purpose, 100 never compressed

images from the BossBase image dataset were used. Each image was resize with randomly chosen scaling factor. The plots in **Figure 11** are averaged over the range of QF2 as before. The success rate is at its maximum for initial compression with quality factor QF1 = 94 and slowly decreases with decreasing QF1. Because this compression occurs before images resizing the compression helps reduce noise while keeping the LP for proper registration.

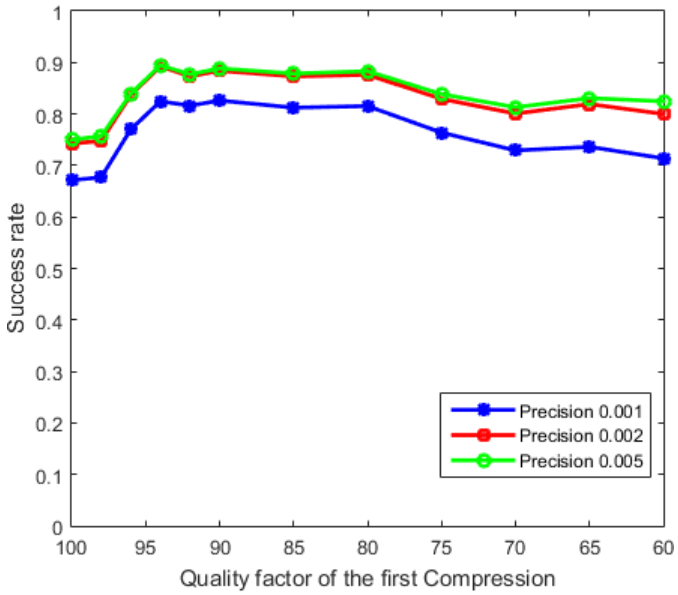



**Figure 11. Success rate of LPD method of scaling factor estimation for 100 images, compressed before resizing with quality factors QF1 between 60 and 100.**

This semi-blind method has not been implemented in the Camera Verification System for the simple reason that not a single reference image from the same camera was provided in evaluation. Only blind methods or fingerprint-assisted method could be used, while the fingerprint was or was not the correct one.

### 4.4 Steganalysis of JPEG images

The architecture of a single-channel SRNet (for grayscale images) is described in a modular fashion in **Figure 12**. Its performance on J-UNIWARD and UED for two quality factors 75 and 95 is contrasted with previous art in **Figure 13** and **Figure 14**. Both figures clearly demonstrate the superior performance of SRNet across all tested payloads, quality factors, and stego algorithms.

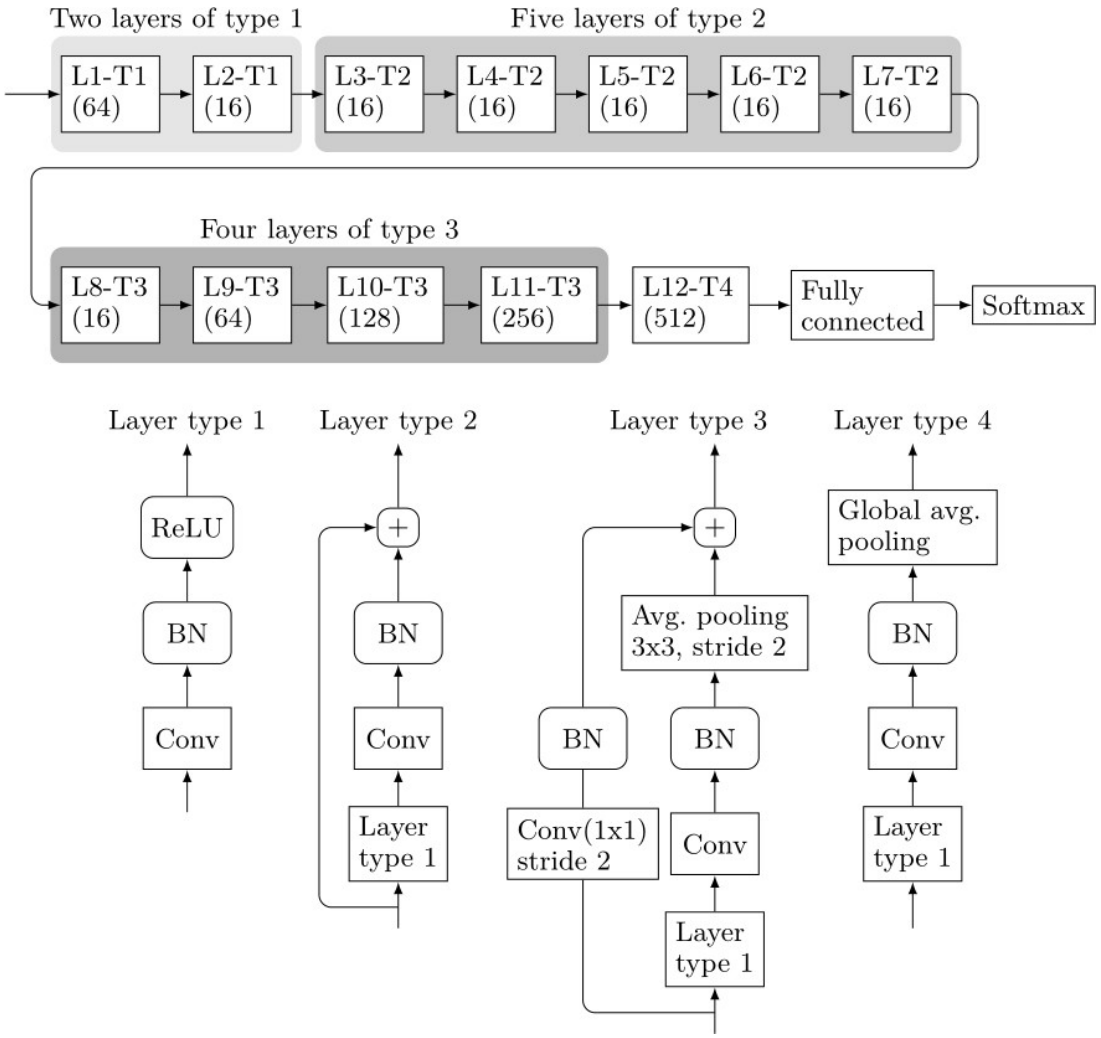



**Figure 12. Architecture of the proposed SRNet for steganalysis. The first two shaded boxes correspond to the segment extracting noise residuals, the dark shaded segment and Layer 12 compactify the feature maps, while the last fully connected layer is a linear classifier. The number in the brackets is the number of 3×3 kernels in convolutional layers in each layer. BN stands for batch normalization.**

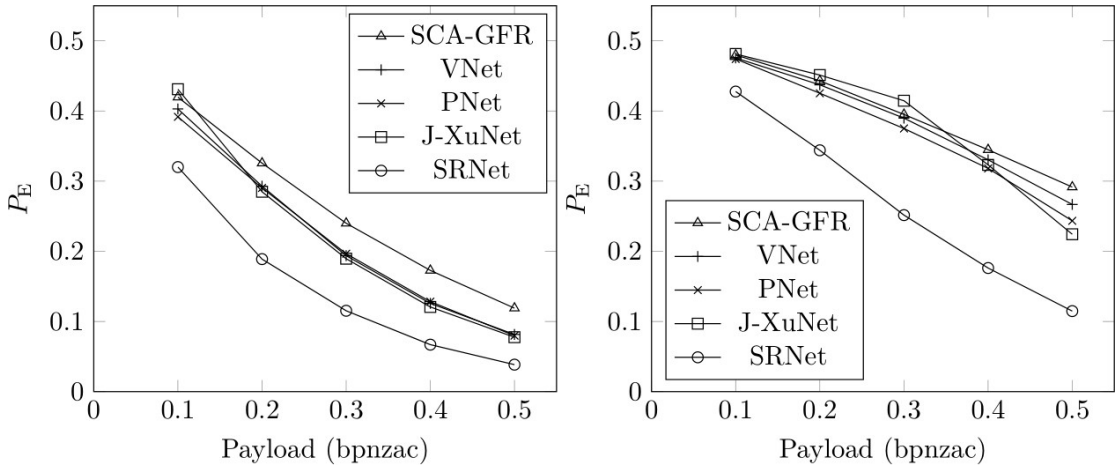



**Figure 13. SRNet total detection error $P_E$ for J-UNIWARD for JPEG quality 75 (left) and 95 (right) contrasted with previous art (PNet and VNet appear in [PhANet]).**

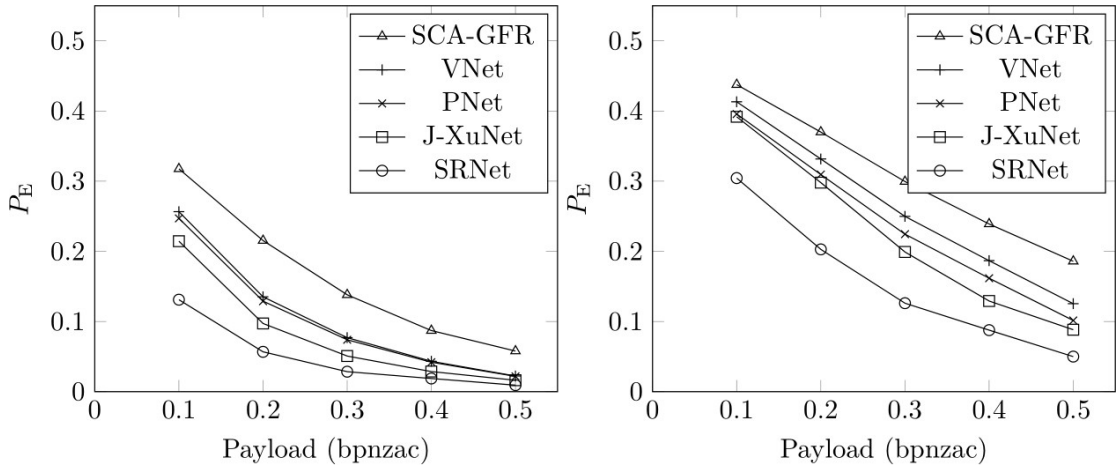



**Figure 14. SRNet total detection error $P_E$ for UED for JPEG quality 75 (left) and 95 (right) contrasted with previous art.**

Detection of diversified stego sources (multiple stego algorithms) in the spatial domain was evaluated on seven algorithms shown in **Table 6**. Multi-class SRNet performed the best out of all tested detectors, including a binary cover-vs-all-stego detector, and a bucket detector. The detector was tuned to have false alarm rate of 3.36% (the probability of detecting a cover as one of the stego algorithms). Notice that the most highly detectable algorithms, the simple LSBR and EA, are also the most reliably identified. On the other hand, the far more advanced HILL and MiPOD are the most difficult to identify. In general, content-adaptive algorithms in some sense follow similar patterns, so mutual confusion among them is to be expected.

**Table 6. Confusion table showing the performance of multi-class SRNet trained to recognize seven embedding algorithms in the spatial domain.**

| True\Det | Cover | HILL | WOW | S-UNI | MiPOD | LSBM | EA | HUGO |
|---|---|---|---|---|---|---|---|---|
| Cover | 0.9664 | 0.0078 | 0.0034 | 0.0024 | 0.0084 | 0.0010 | 0.0030 | 0.0066 |
| HILL | 0.3304 | 0.5962 | 0.0116 | 0.0102 | 0.0346 | 0.0018 | 0.0056 | 0.0096 |
| WOW | 0.1962 | 0.0192 | 0.7004 | 0.0514 | 0.0210 | 0.0012 | 0.0072 | 0.0034 |
| S-UNI | 0.2182 | 0.0306 | 0.0656 | 0.6150 | 0.0396 | 0.0082 | 0.0072 | 0.0156 |
| MiPOD | 0.3332 | 0.0766 | 0.0194 | 0.0326 | 0.5190 | 0.0026 | 0.0068 | 0.0098 |
| LSBM | 0.0596 | 0.0010 | 0.0010 | 0.0088 | 0.0008 | 0.9248 | 0.0020 | 0.0020 |
| EA | 0.1134 | 0.0030 | 0.0026 | 0.0020 | 0.0020 | 0.0002 | 0.8726 | 0.0042 |
| HUGO | 0.3102 | 0.0142 | 0.0048 | 0.0116 | 0.0094 | 0.0020 | 0.0162 | 0.6316 |

**Table 7** and **Table 8** show the detection accuracy for the SRNet trained on a two-channel input: the original decompressed image and its reference version (R-SRNet) for generic LSB flipper (LSBF), OutGuess [OG], and Model-Based Steganography (MBS) [MBS]. Note that even extremely small payloads can be detected very reliably. Also, note the gain brought by the reference channel (the difference between SRNet and R-SRNet), which ranges from 13% for OutGuess at quality 95 to 2% for MBS at quality 75. Reference [RefC] contains a more detailed exposition of the reference channel and extended experimental results.

**Table 7. Detection accuracy with SRNet with reference channel (R-SRNet) for LSBR in JPEG domain (LSBF) and OutGuess at different embedding change rates β (probability of modifying a DCT coefficient).**

| | LSBF QF 75 | | LSBF QF 95 | | OutGuess QF 75 | | OutGuess QF 95 | |
|---|---|---|---|---|---|---|---|---|
| $\beta$ | R-SRNet | SRNet | R-SRNet | SRNet | R-SRNet | SRNet | R-SRNet | SRNet |
| 0.03 | 99.96 | 99.13 | 99.87 | 99.50 | 99.11 | 96.08 | 99.50 | 97.21 |
| 0.02 | 99.65 | 98.42 | 99.64 | 98.54 | 97.18 | 90.36 | 98.71 | 94.79 |
| 0.01 | 96.93 | 92.98 | 97.95 | 94.07 | 90.14 | 81.44 | 94.75 | 84.94 |
| 0.005 | 89.10 | 83.16 | 91.90 | 85.74 | 79.39 | 71.00 | 84.83 | 74.20 |
| 0.003 | 80.84 | 74.42 | 84.75 | 77.37 | 72.86 | 64.50 | 79.37 | 66.95 |

The ALASKA I evaluation (competition) was won by the Binghamton team by a very large margin. A detailed exposition of the detectors and their performance appears in [A1]. The second team fell behind 25% in terms of the evaluation metric, MD5, the missed detection at 5% false alarm rate. The performance, evaluated in terms of three detection measures, the MD5, the total detection error under equal class priors $P_E$, and the false alarm rate at 50% correct stego detection, are displayed in **Table 9**. The performance ($P_E$ left, MD5 right) of the three-channel YCrCb SRNet per each quality factor is shown in **Figure 15**. It shows the performance of the tile detector trained on 256×256 tiles embedded with double payload, a detector trained for the same tile size with the original payload (the double payload was used for curriculum training to

**Table 8. Detection accuracy with SRNet with reference channel (R-SRNet) for LSBR in JPEG domain (LSBF) for MBS for a range of relative payloads R in bpnzac.**

| | MBS QF 75 | | MBS QF 95 | |
|---|---|---|---|---|
| *R* | R-SRNet | SRNet | R-SRNet | SRNet |
| 0.1 | 99.76 | 98.81 | 99.86 | 99.52 |
| 0.05 | 96.90 | 93.84 | 98.99 | 96.70 |
| 0.03 | 91.09 | 85.76 | 95.85 | 90.93 |
| 0.02 | 84.79 | 80.06 | 91.44 | 85.74 |
| 0.01 | 73.48 | 68.88 | 81.25 | 75.59 |
| 0.005 | 64.29 | 60.65 | 69.37 | 65.77 |
| 0.003 | 59.08 | 57.37 | 62.61 | 59.99 |

improve the detection accuracy), and the performance on arbitrarily sized images. The payload scaling w.r.t. the quality factor made the mid 90's qualities the hardest to detect. The detection performance on tiles (top) and arbitrary sizes (bottom) per each stego scheme is shown in **Figure 16** in terms of MD5 vs. JPEG quality. Notice the increased detection error for nsF5, which was caused by the failure of the SRNet to detect this embedding algorithm better. This problem was later addressed by the Binghamton team with the OneHotConvNet [OH] with one-hot encoding of DCTs to allow the detector to detect artifacts in the distribution of DCT coefficients.

**Table 9 Final scores obtained on the local test set and on ALASKArank.**

| Dataset | MD5 | $P_E$ | FA50 |
|---|---|---|---|
| TST | 18.55 | 11.50 | 0.09 |
| ALASKArank | 25.20 | 14.63 | 0.77 |

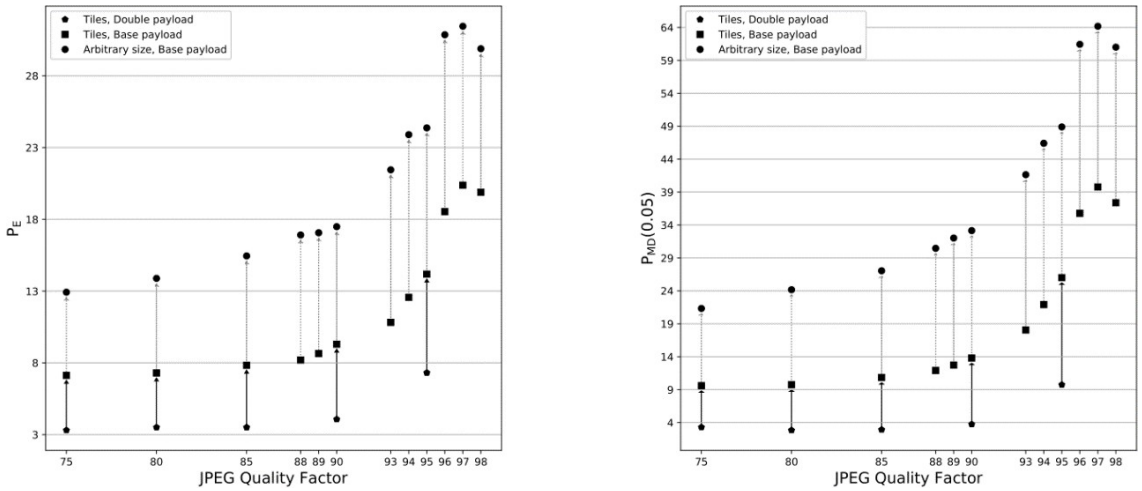



**Figure 15. Total detection error $P_E$ and MD5 for YCrCb-SRNet on color images of arbitrary size for each stego scheme.**

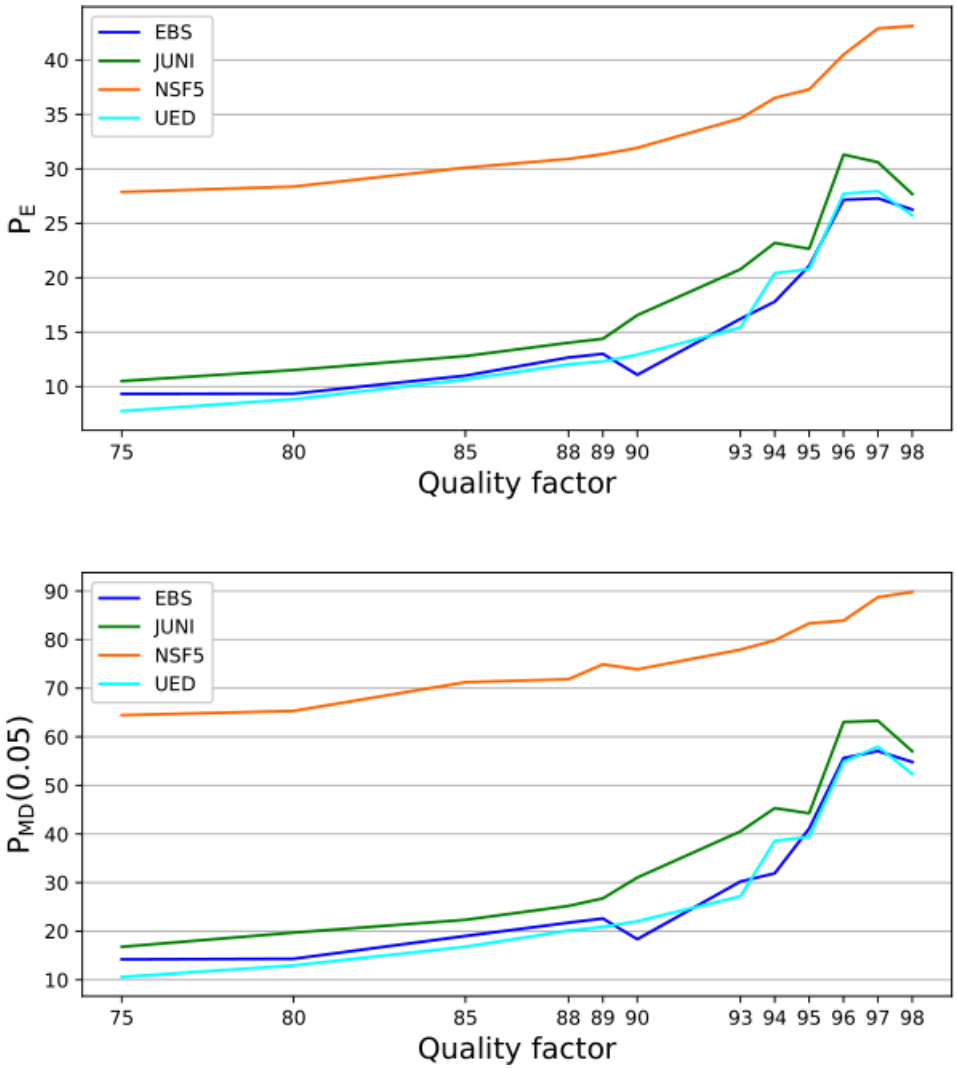



**Figure 16. MD5 for the tile detector (YCrCb-SRNet) trained on tiles (top) and on images of arbitrary size.**

The accuracy of the reverse JPEG compatibility attack [RJCA] is shown in **Table 10**. Three versions of the SRNet, the original single-layer network, the same network trained on rounding errors, e-SRNet, and a two-channel SRNet with both rounding errors and the decompressed luminance eY-SRNet. The table shows the detection accuracy for J-UNIWARD for quality factors 99 and 100 and a range of payloads. Note the improvement of the two-channel SRNet and the accuracy close to 100% for quality 100 even for very small payloads. Similar results can be observed for other embedding schemes as this attack is universal.

**Table 10. Detection accuracy of three different versions of SRNet when training on decompressed images (SRNet), rounding errors (e-SRNet), and both (eY-SRNet). Dataset: BOSSbase + BOWS2. J-UNIWARD, payload in bits per non-zero DCT coefficient.**

| | QF 100 | | | QF 99 | | |
|---|---|---|---|---|---|---|
| Payload | SRNet | e-SRNet | eY-SRNet | SRNet | e-SRNet | eY-SRNet |
| 0.4 | 0.8829 | 0.9998 | 0.9995 | 0.8592 | 0.9980 | 0.9994 |
| 0.3 | 0.8331 | 0.9998 | 0.9998 | 0.8054 | 0.9960 | 0.9990 |
| 0.2 | 0.7548 | 0.9998 | 0.9993 | 0.7257 | 0.9832 | 0.9981 |
| 0.1 | 0.6488 | 0.9998 | 0.9984 | 0.6015 | 0.9316 | 0.9780 |
| 0.05 | 0.5682 | 0.9946 | 0.9992 | 0.5437 | 0.7989 | 0.9287 |

## 5.0 CONCLUSIONS

### 5.1 Vectorized context-aware pixel descriptors for manipulation detection

We have developed two competitive image manipulation localization techniques that advance prior work through the explicit use of contextual information and unsupervised non-binary local decisions. Program evaluations demonstrate that our techniques produce favorable results, achieving trMCC's of 0.202 (TRR 0.13) and 0.211 (TRR 0.9) in the MFC2018 cycle, respectively.

### 5.2 Detection of processing history

Establishing the pedigree of a digital image, such as the type of processing applied to it, is important for forensic analysts because processing generally affects the accuracy and applicability of other forensic tools used for, e.g., identifying the camera (brand) and/or inspecting the image integrity (detecting regions that were manipulated). Global edits have been proposed in the past for "laundering" manipulated content because they can negatively affect the reliability of many forensic techniques. The research focused on the more difficult and less addressed case when the processed image is JPEG compressed.

Two approaches were investigated and evaluated a maximum-likelihood detector built from a parametric multivariate Gaussian model in the space of projections on weight vectors of binary classifiers trained between the class of unprocessed images and images processed with a given (range) of processing type. The second approach was a multi-class CNN designed specifically for this purpose. Both detectors have been evaluated experimentally with the CNN detector showing the better performance. The code has been delivered to DARPA.

### 5.3 Media source identification and verification

Media forensic techniques that are based on properties of digital camera hardware and build-in software, such as the propagation of PRNU, undergo evolution along with the evolution of the hardware and software themselves. In order to keep up with the fast increasing camera capabilities and software options offered to consumers the established forensic techniques have to be repeatedly scrutinized, tested, and adjusted to the new additions from camera manufactures.

One example of modern image enhancement supported by cameras hardware and built-in software that can severely jeopardize PRNU-based methods is HDR. The mechanism of producing HDR images was analyzed and a special treatment in the context of camera verification was developed and successfully tested. The camera PRNU is estimated from the probe image and matched to the camera fingerprint locally. Up to three source images that made the HDR image can be identified and their parts of PRNU properly composed.

Image resizing occurs more often, even without a camera user being notified, for example when HDR switch is on or during involuntary zooming. Determining the scaling factor is crucial for PRNU matching procedure for camera verification and for PRNU-based manipulation localization. Two qualitatively different methods have been introduced for fast estimation of the scaling factor. The first is the semi-blind method that utilizes one reference image from the same camera for matching contained linear patterns and JPEG dimples (if present) in Fourier domain. The second is the blind method, i.e. not relying on any side information, which is an improved version of previously existing method that takes advantage of resampling artifacts in frequency

domain.

Entire Camera Verification System was built, implemented and evaluated on independent set of probe images during annual evaluation. Both the AUC and the detection rate at false alarm of $10^{-3}$ in the 2020 evaluation was the best among teams in "train on images, test on images" setting.

### 5.4 Steganalysis of JPEG images

Besides improving the accuracy of detectors of steganography in JPEG images, this effort addressed several important practical aspects that have so far been largely ignored by the research community:

1. Ability to analyze color images
2. Images of arbitrary size
3. Diversified stego sources consisting of multiple stego algorithms
4. Unknown and variable payload
5. A wide range of quality factors
6. Diverse processing history of cover images

A new CNN architecture (the SRNet) has been designed that is universal (it can detect steganography in both spatial and JPEG domain) and free of fixed or pre-initialized filters and can be trained in an end-to-end fashion for state-of-the-art performance.

SRNet was expanded to accept three-dimensional inputs (a three-channel YCrCb version) for steganalysis of color images. When available, reference signals can be inputted as an additional channel to improve performance of older steganographic paradigms based on flipping bits (Jsteg, OutGuess, JP Hide&Seek) or exchanging values from small disjoint sets (model-based steganography). The reference channel provides a significant boost especially for detection of small payloads.

Furthermore, SRNet has been extended to work accurately for images of arbitrary size, which was achieved by first training a "tile detector" and then building the detector (multi-layer perceptron) on statistical moments of feature maps extracted from the output of the last convolutional layer. A multi-class SRNet was designed and shown to be effective when the detector needs to detect stego images from potentially many different stego schemes.

For better scalability w.r.t. the JPEG quality factor, and to give the detector the ability to generalize to custom quantization tables, the SRNet was trained on specific ranges of quality factors, which slightly improved performance and significantly decreased the computational burden associated with having to train a detector per each quality factor. Conventional CNN models pre-trained on ImageNet, such as the EfficientNet and the MixNet, were shown to be excellent detectors of diversified stego sources with variable payload. Additionally, such models can be trained on a much wider range of JPEG qualities, which further drastically simplifies the process of building the detector.

A novel special attack on JPEG steganography that is universal and extremely accurate was developed for qualities 99 and 100. It is called the reverse JPEG compatibility attack, and it works whenever the DCT is applied to integer-valued signal. Finally, a novel CNN architecture has been designed that uses OneHot encoding of DCT coefficients for steganalysis directly in the DCT domain. This detector gives the best performance for stego schemes that introduce artifacts in the DCT domain, such as F5 (nsF5), OutGuess, and for J-UNIWARD in JPEG images compressed with the 'trunc' quantizer instead of the 'round' quantizer.

## 6.0 REFERENCES

**[CISC]** M. Goljan, J. Fridrich, "Camera Identification from Scaled and Cropped Images," in E. J. Delp et al. editors, *Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, vol. 6819, pp. 0E–0F, 2008.

**[DIori]** J. Lukáš, J. Fridrich, and M. Goljan, "Determining digital image origin using sensor imperfections," *Proc. SPIE, Image and Video Communications and Processing*, SPIE, volume 5685, pp. 249–260, San Jose, CA, Jan 2005.

**[DIS]** Y.-F. Hsu and S.-F. Chang, "Detecting image splicing using geometry invariants and camera characteristics consistency," in IEEE International Conference on Multimedia and Expo (ICME), 2006.

**[DRF]** S. Kumar and M. Hebert, "Discriminative random fields," International Journal of Computer Vision, vol. 68, no. 2, pp. 179–201, 2006.

**[EA]** W. Luo, F. Huang, and J. Huang, "Edge adaptive image steganography based on LSB matching revisited," *IEEE TIFS*, 5(2):201–214, June 2010.

**[EBS]** C. Wang and J. Ni, "An efficient JPEG steganographic scheme based on the block–entropy of DCT coefficients." *IEEE ICASSP*, Kyoto, Japan, March 25–30, 2012.

**[FB]** S. Bayram, H. T. Sencar, and N. Memon, "Efficient sensor fingerprint matching through fingerprint binarization," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1404–1413, 2012.

**[HDRset]** O. Al Shaya, P. Yang, R. Ni, Y. Zhao, and A. Piva, "A new dataset for source identification of high dynamic range images," Sensors, 18(11), 2018.

**[HILL]** B. Li, M. Wang, and J. Huang, "A new cost function for spatial image steganography," *IEEE ICIP*, Paris, France, October 27–30, 2014.

**[HUGO]** T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in R. Böhme and R. Safavi-Naini, editors, *Information Hiding, 12th International Conference*, volume 6387 of Lecture Notes in Computer Science, pages 161–177, Calgary, Canada, June 28–30, 2010. Springer-Verlag, New York.

**[JRM]** J. Kodovský and J. Fridrich, "Steganalysis of JPEG images using rich models," in A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media*

*Watermarking, Security, and Forensics 2012*, volume 8303, pp. 0A 1–13, San Francisco, CA, January 23–26, 2012.

**[Jsteg]** D. Upham. Steganographic algorithm JSteg. Software available at http://zooid.org/paul/crypto/jsteg.

**[JXuNet]** G. Xu, "Deep convolutional neural network to detect J-UNIWARD," in M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017.

**[LST]** M. Goljan, J. Fridrich, and T. Filler,"Large scale test of sensor fingerprint camera identification," in N.D. Memon, E.J. Delp, P.W. Wong, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Media Forensics and Security XI*, volume 7254, pp. 0I 1–12, San Jose, CA, January 2009.

**[MBS]** P. Sallee, "Model-based methods for steganography and steganalysis," *International Journal of Image Graphics*, 5(1):167–190, 2005.

**[MiPOD]** V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," IEEE TIFS, 11(2):221–234, 2016.

**[MSA]** P. Korus and J. Huang, "Multi-scale analysis strategies in PRNU-based tampering localization," IEEE Transactions on Information Forensics and Security.

**[nsF5]** A. Westfeld, "High capacity despite better steganalysis (F5 – a steganographic algorithm)," in I. S. Moskowitz, editor, *Information Hiding, 4th International Workshop*, volume 2137 of Lecture Notes in Computer Science, pages 289–302, Pittsburgh, PA, April 25–27, 2001. Springer-Verlag, New York.

**[OG]** N. Provos, "Defending against statistical steganalysis," *10th USENIX Security Symposium*, pages 323–335, Washington, DC, August 13–17, 2001.

**[PhANet]** M. Chen, V. Sedighi, M. Boroumand, and J. Fridrich, "JPEG-phase-aware convolutional neural network for steganalysis of JPEG images," in M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017.

**[RD]** M. Kirchner and T. Gloe, "On resampling detection in re-compressed images," in *First IEEE International Workshop on Information Forensics and Security*, Dec 2009, pp. 21−5.

**[SCRM]** M. Goljan, R. Cogranne, and J. Fridrich, "Rich model for steganalysis of color

images," *Sixth IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

**[SD]** J. Fridrich, "Sensor defects in digital image forensics," in Digital Image Forensics: There is More to a Picture Than Meets the Eye, H. T. Sencar and N. Memon, Eds. Springer, 2013, pp. 179–218.

**[SIDHDR]** O. Al Shaya, P. Yang, R. Ni, Y. Zhao, and A. Piva, "A new dataset for source identification of high dynamic range images," Sensors, 18(11), 2018.

**[SPLICE]** 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.

**[SRL]** A. D. Ker, "The square root law of steganography," in M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017. ACM Press.

**[SRM]** J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE TIFS*, 7(3):868–882, June 2012.

**[UED]** L. Guo, J. Ni, and Y. Q. Shi, "Uniform embedding for efficient JPEG steganography," *IEEE TIFS*, 9(5):814–825, May 2014.

**[UNI]** V. Holub, J. Fridrich, and T. Denemark, "Universal distortion design for steganography in an arbitrary domain," *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop*, 2014:1, 2014.

**[WOW]** V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

**[Xu]** G. Xu, H. Z. Wu, and Y. Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Processing Letters*, 23(5):708–712, May 2016.

**[YeNet]** J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE TIFS*, 12(11):2545–2557, November 2017.

## APPENDIX A – Publications and Presentations

All listed conference papers have been either presented in person at the respective event or remotely for on-line events.

[PH1] M. Boroumand and J. Fridrich, “Scalable Processing History Detector for JPEG Images,” *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2017*, San Francisco, CA, January 29–February 2, 2017.

[IML] S. Chakraborty and M. Kirchner, "PRNU-based Image Manipulation Localization with Discriminative Random Fields," *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2017*, San Francisco, CA, January 29–February 2, 2017.

[PH2] M. Boroumand and J. Fridrich, “Deep Learning for Detecting Processing History of Images,” *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2018*, San Francisco, CA, January 29–February 1, 2018.

[IMD] M. Goljan, J. Fridrich, and M. Kirchner, "Image Manipulation Detection Using Sensor Linear Pattern," *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2018*, San Francisco, CA, January 29–February 1, 2018.

[SRNet] M. Boroumand, M. Chen, and J. Fridrich, “Deep Residual Network for Steganalysis of Digital Images,” *IEEE TIFS*, 14(5), pp. 1181–1193, May 2019.

[Diver] J. Butora and J. Fridrich, “Detection of Diversified Stego Sources with CNNs,” *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2019*, San Francisco, CA, January 14–17, 2019.

[RWTY] M. Boroumand, J. Fridrich, and R. Cogranne, “Are we there yet?” *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2019*, San Francisco, CA, January 14–17, 2019.

[RefC] Mo Chen, M. Boroumand , and J. Fridrich, “Reference Channels for Steganalysis of Images with Convolutional Neural Networks,” *7th IH&MMSec. Workshop*, Paris, France, July 3–5, 2019.

[A1] Y. Yousfi, J. Butora, Q. Giboulot, and J. Fridrich, “Breaking ALASKA: Color Separation for Steganalysis in JPEG Domain,” *7th IH&MMSec. Workshop*, Paris, France, July 3–5, 2019.

[IMLu] M. Darvish Morshedi Hosseini and M. Kirchner, "Unsupervised Image Manipulation Localization With Non-Binary Label Attribution," *IEEE Signal Processing Letters*, 26 (7), pp. 976–980, July 2019.

[HDR] M. Darvish Morshedi Hosseini and M. Goljan, “Camera Identification from HDR Images,” *7th IH&MMSec. Workshop*, Paris, France, July 3–5, 2019.

[ArbS] C. Fuji-Tsang and J. Fridrich. “Steganalyzing images of arbitrary size with CNNs,” In

A. Alattar and N. D. Memon, editors, Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2018, Burlingame, CA, January 29–February 2, 2018.

[RJCA] J. Butora and J. Fridrich, "Reverse JPEG Compatibility Attack," *IEEE TIFS*, 15(1), pp. 1444–1454, December 2019.

[QFsec] J. Butora and J. Fridrich, "Effect of JPEG Quality on Steganographic Security," *7th IH&MMSec. Workshop*, Paris, France, July 3–5, 2019.

[QFscale] Y. Yousfi and J. Fridrich, "JPEG Steganalysis Detectors Scalable With Respect to Compression Quality," *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2020*, San Francisco, CA, January 26–30, 2020.

[SB] M. Goljan and M. Darvish Morshedi Hosseini "Semi-Blind Image Resampling Factor Estimation for PRNU Computation," *Proc. IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2020*, San Francisco, CA, January 26–30, 2020.

[Trunc] J. Butora and J. Fridrich, "Steganography and its Detection in JPEG Images Obtained with the 'Trunc' Quantizer," *IEEE ICASSP*, Barcelona, Spain, May 4–8, 2020.

[OH] Y. Yousfi and J. Fridrich, "An Intriguing Struggle of CNNs in JPEG Steganalysis and the OneHotConv Solution," *IEEE SPL*, 2020.

[Cost] J. Butora, Y. Yousfi, and J. Fridrich, "Turning Cost-Based Steganography into Model-Based," *8th IH&MMSec. Workshop*, Denver, CO, June 20–22, 2020.

[DC] J. Butora and J. Fridrich, "Extending the Reverse JPEG Compatibility Attack to Double Compressed Images," *IEEE SPL*, July 2020, submitted.

[A2] Y. Yousfi, J. Butora, E. Khvedchenya, and J. Fridrich, "ImageNet Pre-trained Models for JPEG Steganalysis," *IEEE WIFS*, New York City, NY, December 6–11, 2020, in preparation.

[CNNPRNU] M. Darvish Morshedi Hosseini, M. Goljan, H. Zeng, and Xiaohua Li, "Wavelet-Based Design of CNN for PRNU Extraction," *IEEE TIFS*, submitted.

## LIST OF ACRONYMS & GLOSSARY

**ALASKA** – public steganalysis evaluation aimed at detecting steganographic content in real life conditions

**BN** – Batch normalization

**BOSS** – Break our steganographic system dataset

**BOWS2** – Break our watermarking system dataset

**CNN** – Convolutional neural network

**DCT** – Discrete cosine transformation

**EA** – Edge adaptive stego method [EA]

**FA50** – False alarm for 50% detection rate

**FFT** – Fast Fourier transform

**HDR** – high dynamic range

**HILL** – High Low Low embedding method [HILL]

**HUGO** – Highly undetectable steGO method [HUGO]

**JPEG** – Joint photographic expert group image format

**JRM** – JPEG rich model [JRM]

**JXuNet** – Steganalysis CNN for JPEG images designed by Xu [JXuNet]

**LSBF** – generic least significant bit flipper

**LSBM** – Least significant bit matching

**LSBR** – Least significant bit replacement

**MBS** – Model based steganography [MBS]

**MD5** – Missed detection rate for 5% false alarm rate

**MiPOD** – Minimizing the power of the most powerful detector method [MiPOD]

**ML** – Maximum likelihood

**PCE** – Peak-to-correlation energy ratio

$P_E$ – Total detection error under equal class priors

**PRNU** – Photo-response non-uniformity

**RJCA** – Reverse JPEG compatibility attack [RJCA]

**ROC** – Receiver operating characteristic

**SCRM** – Spatio-color rich model [SCRM]

**SNR** – Signal-to-noise ratio

**SPN** – Sensor pattern noise

**SRM** – Spatial rich model [SRM]

**SRNet** – Steganalysis residual network [SRNet]

**e-SRNet** – Steganalysis residual network trained on rounding errors

**eY-SRNet** – Steganalysis residual network trained on rounding errors and decompressed images

**R-SRNet** – Steganalysis residual network with reference channel

**UED** – Uniform embedding distortion steganography [UED]

**UNIWARD** – Universal wavelet relative distortion [UNI]

**WOW** – Wavelet obtained weights [WOW]

**XuNet** – Steganalysis CNN designed by Xu [XuNet]

**YCrCb** – luminance, chrominance, chrominance color representation

**YeNet** – Steganalysis CNN for the spatial domain designed by Ye [YeNet]