Real or virtual: a video conferencing background manipulation-detection system

Nowroozi, E., Mekdad, Y., Conti, M., Milani, S., & Uluagac, S. (2024). Real or virtual: a video conferencing background manipulation-detection system. Multimedia Tools and Applications. Advance online publication. https://doi.org/10.1007/s11042-024-20251-6

Published in:
Multimedia Tools and Applications

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2024 The Authors.
This is an open access article published under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date:04. Oct. 2024
Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20251-6

Real or virtual: a video conferencing background manipulation-detection system

Ehsan Nowroozi1,2 · Yassine Mekdad3 · Mauro Conti4 · Simone Milani5 · Selcuk Uluagac6

Received: 26 February 2023 / Revised: 6 August 2024 / Accepted: 2 September 2024
© The Author(s) 2024

Abstract
In the past few years, the popularity and wide use of video conferencing software enjoyed exponential growth in market size. This technology enables participants in different geographic regions to have a virtual face-to-face meeting. Additionally, it allows participants to utilize virtual backgrounds to hide their real environment with privacy concerns or to reduce distractions, particularly in professional settings. In scenarios where the users should not hide their actual locations, they may mislead other participants into assuming that the displayed virtual backgrounds are real. In this paper, we propose a new publicly-available dataset of virtual and real backgrounds in video conferencing software (e.g., Zoom, Google Meet, Microsoft Teams). The presented archive was evaluated by an exhaustive series of tests and scenarios using two well-known features extraction methods: CRSPAM1372 and six co-mat. The first verification scenario considers the case where the detector is unaware of manipulated frames (i.e., the forensically-edited frames are not part of the training set). A model trained on zoom frames that were tested with Google Meet frames can detect real background images from virtual ones in video conferencing software with 99.80% detection accuracy. Furthermore, it is possible to distinguish virtual from real backgrounds in videos created for videoconferencing software at a high detection rate of approximately 99.80%. According to our conclusions, the proposed method greatly enhanced the detection accuracy and resistance against diverse adversarial conditions, making it a reliable technique for classifying actual as opposed to virtual backgrounds in video communications. Given the described dataset provided and some preliminary experiments that we performed, we expect that it will lead to more future research in this domain.

Keywords Artificial intelligence · Video conferencing · Virtual backgrounds · Real background detection · Background manipulation · Multimedia forensics and security · Adversarial machine learning · Privacy in video conferencing · Video background verification · Forensic detection systems · Adversarial multimedia forensics

1 Introduction
Applications such as Google Meet, Microsoft Teams, and Zoom have become essential, especially in situations1 where face-to-face meetings are impractical. While enabling social

---

1 The dataset is available at [1].

Extended author information available on the last page of the article

Published online: 25 September 2024
Multimedia Tools and Applications

distancing, these systems facilitate smart work, online gatherings, and virtual conferences [3, 7, 27]. Owing to their popularity, privacy and security measures need to be stronger than ever to fend off attacks, such as JPEG compression or image post-processing. Users of these platforms can set virtual backgrounds to hide their real locations, remove distractions, or mask people behind them. However, this creates a situation where people can deceive others by making them believe that they are in different places, thus creating a challenge for verifying the authenticity of the background [41]. The detection of virtual backgrounds is difficult because robust video features are required for accurate detection. This paper addresses the problem of detecting when someone is using a fake background during video conferencing, which misrepresents their whereabouts. Unfortunately, the current detectors fail under adverse conditions. Therefore, we seek to provide a new dataset and detection algorithms that can differentiate between real and fake backgrounds, even under adversarial attacks, bridge the gap left by existing detection methods, and enhance security for video conferencing.

In recent years, the widespread adoption of video conferencing software has underscored the need for effective privacy and security measures. While virtual backgrounds provide privacy by concealing the user’s actual surroundings, they also open avenues for deception, where users might mislead others by pretending to be in a different location. For example, when interviewing someone for a job remotely, they could use a virtual background that makes it appear like they are in an office setting while they are actually in a coffee shop or their bedroom. Similarly, in an online classroom, a student might put up a virtual background so nobody knows they are not paying attention or are somewhere inappropriate for learning. Moreover, during virtual court hearings, witnesses could lie about their location if being in hostile environments affects how credible their testimonies will be considered by judges presiding over such cases; thus underlining why it is important to come up with effective strategies for detecting and dealing with abuses associated with this technology. Existing datasets and detection methods do not sufficiently address this specific challenge under various adversarial conditions. Therefore, our primary motivation for proposing this dataset is to bridge this gap by providing a comprehensive, publicly available dataset of real and virtual backgrounds recorded under diverse conditions. This dataset aims to facilitate the development and benchmarking of robust detection algorithms capable of distinguishing between real and virtual backgrounds, even under adversarial settings, thereby enhancing the reliability and security of video conferencing applications. The study has been further enhanced by referring to more significant works in image and video representation, like the ones suggested by [38, 42] and [37] which give ideas on advanced techniques of image and video analysis thereby promoting a comprehensive knowledge about the subject.

Additionally, the authors considered two Generative Adversarial Network (GAN) image datasets based on unpaired image-to-image translations: the styleGAN [44] and the starGAN [9] datasets. In MF, residual images are frequently used in calculating the co-occurrence matrices to identify or localize the detection [4, 5]; these approaches often consider the SPAM [34], or the CSRMQ [14] features, which are initially used for steganalysis applications. In [24], the color components are used to generate co-occurrence matrices from high-pass filtering residuals, as well as for each truncated residual image. Then, the co-occurrences are obtained by combining the color channels into a feature to train the Support Vector Machine (SVM). The solutions in [3, 11] achieve significant results by feeding six co-occurrence features.

\[^{2}\] In this paper, the authors considered the name CSPAM, which stands for the color spam. However, in our work, we consider CRSPAM1372 where CR stands for ColoR, SPAM stands for the features extraction method, and 1372 stands for the feature dimension.
matrices (hereafter, we refer to as six co-mat) into the CNN from spatial and spectral bands. This approach computes co-occurrence matrices by considering cross-band co-occurrences (spectral) and co-occurrences (spatial) on the grey level computed separately on the single bands. Although only tested on ProGAN [18] and StyleGAN [19] datasets, the authors argue that it can be employed in any MF detection task. Compared to the three co-occurrence matrices approach that has spatial features [29], the six co-occurrence matrices approach with both spatial and spectral features significantly enhance the resilience against various types of post-processing and laundering attacks [3]. In this context, we consider two popular methods recently employed in MF for our dataset verification: (i) the CRSPAM1372 method [5], and (ii) the six co-mat method [3]. We mention that the CRSPAM1372 approach is similar to the CSRMQ method [14]. However, the last one considers all color channels separately and across the color channels. Despite the studies conducted so far, the advancements in this field are insignificant, given the lack of additional datasets. In this context, we are motivated to fill this gap by presenting novel datasets focusing on real and virtual backgrounds across different video conferencing software.

To verify the validity of our dataset, we conducted a series of a test by considering different scenarios when the detector is unaware and aware of an adversary. Moreover, we evaluate all detectors’ robustness in the presence of post-processing, indoor lightening conditions against software mismatches (e.g., Google Meet, Microsoft Teams), and post-processing operations before compression. We also captured different videos using additional cameras for this purpose. In particular, we considered the MSI GF65 10UX laptop (low-quality camera) and Apple IPAD Pro (high-quality camera). Further studies will consider more cameras to increase the dataset’s size. In our verification study, we found that the detection accuracy is relatively high if the camera’s fingerprints are close to the subject. Additionally, for better performance against different cameras, we found that either the detector needs to be fine-tuned or another methodology should be used to detect different camera models based on sensor pattern noise [12]. Overall, our experimental results demonstrate a good performance over a wide variety of scenarios considered in this study [1]. Table 1 lists all the abbreviations and notations used in our study.

1.1 Comparison with previous works

Our approach improves the basis prepared by previous articles, particularly the papers shown in [10] and [28]. The article by [10] suggests a general video manipulation detection system based on pixel co-occurrence matrices while ours incorporates advanced spectral as well as spatial feature extraction approaches targeted at video conferencing scenarios. The dataset described in paper [28], however, does not have an elaborate analysis under different environmental conditions. In contrast to this approach, we also perform robustness of our detector in different lighting situations and adversarial settings which makes it more appropriate for applying such methods to real-world applications. This way, the method refines its accuracy and usability for practical purposes in various settings of video conferencing by addressing these specific scenarios.

Our method considers robustness by conducting tests that are consistent with optical spectrum conditions and the adversarial network replication of light spectrum, which are not sufficiently covered in [28]. These additions highlight the strength of our approach in detecting counterfeit backgrounds when the conditions are difficult, making it practically useful in real-life situations.
| Acronym       | Description                                      |
|--------------|--------------------------------------------------|
| CRSPAM       | Color Rich Subtractive Pixel Adjacency Matrix    |
| SRMQ1        | Spatial Rich Model                               |
| AI           | Artificial Intelligence                          |
| ML, DL       | Machine and Deep Learning                       |
| CNN          | Convolutional Neural Network                     |
| GAN          | Generative Adversarial Network                   |
| SGD          | Stochastic Gradient Descent                      |
| MF           | Multimedia Forensics                             |
| RWD          | Real-World Data                                  |
| SVM          | Support Vector Machine                           |
| $I$          | Image                                            |
| $\alpha$     | Horizontal displacement                          |
| $\beta$      | Vertical displacement                             |
| $\delta$     | Color channel                                    |
| $W$          | Width of the image                               |
| $L$          | Height of the image                              |
| $\Gamma$     | Displacement vector for spatial co-occurrence    |
| $\Gamma'$    | Displacement vector for inter-channel co-occurrence |
| $C_{\Gamma}(I)$ | Spatial co-occurrence matrix                     |
| $C_{\Gamma'}(I)$ | Cross-band co-occurrence matrix                  |
| $x$          | Row Pixel intensity value                        |
| $y$          | Column Pixel intensity value                      |
| Adv-ML       | Adversarial Machine Learning                     |
| $H_0$        | Real background                                  |
| $H_1$        | Virtual background                               |
| $R, G, B$    | Red, Green, Blue                                 |
| CLAHE        | Clip Limit Adaptive Histogram Equalization       |

**Comparison of computational costs:** Our dataset is about 4.551 GB in total and we will take new videos with different cameras and backgrounds in the future, which has high-resolution video files. Our data set covers more ranges of conditions and settings, making it more suitable for general usability testing. Our data was processed on a machine with approximately 64 GB RAM including feature extraction and model training. On the other hand, [10] and [28] are mainly concerned with static image processing computational complexity; this generally requires less computing power than dynamic video analysis does. However, even though the computational expenses are higher, this approach has improved robustness and applicability to various cases.

### 1.2 Problem definition

In our day-to-day professional life, video conferencing meetings have become ubiquitous, especially when physical meetings are impossible. Several companies have allowed their employees to work remotely and cooperate globally through video conferencing software.
Such benefits enable the digital workforce and promote collaboration. However, video conferencing software might pose privacy threats. For instance, the visual content transmitted during a video conferencing call can directly be shared from users’ private environments to third parties, thus exposing sensitive information. To that end, video conferencing software solves the privacy leakage by replacing the observable environment of a video conferencing call with a virtual background that covers the sensitive living environment. Nevertheless, this solution is inefficient as it might leak some parts of the real background in video frames. In [17], the authors explored the exploitation of these leaks in virtual backgrounds of Zoom video conferencing software and reconstructed parts of the real background. Assuming the videos are recorded in high quality without pixel leakage from the real background, the question that remains unanswered is whether the recorded video is considered a virtual or real background. To the best of our knowledge, most of the existing detectors rely on Artificial Intelligence (AI) to obtain detection accuracy. In such cases, we are motivated to investigate the possibility of affecting the detection performance by the adversary. On the other hand, the study proposed by [17] showed that when considering a high-quality camera, a superficial, tiny leakage in virtual background frames happens if the subject moves slowly. In this context, we aim to investigate the detector’s performance for detecting real background from virtual one if the subject in the recorded video moves slowly without pixel leakage. In such a scenario, the users can mislead other participants by claiming their virtual background as a real one. To the best of our knowledge, the existing research on the robustification of DL detectors against different attacks is limited [25, 26]. In our work, we construct a set of novel datasets that consists of real and virtual background videos. Then, we verify them under two specific scenarios:

- the detector is unaware of the attacks that may be considered by an adversary to degrade the detection accuracy (unaware scenario);
- the detector is aware of the attacks, and therefore, it trains the analyzing software with adversarial manipulated data. (aware scenario).

### 1.3 Research aims and questions

Recently, significant research efforts have been dedicated to developing forensic tools to retrieve data and detect tampered multimedia documents [39]. However, existing tools often fail under adversarial settings, which is a major challenge in multimedia forensics [6, 31, 35]. This research aims to address this gap by providing a novel dataset specifically designed for video conferencing scenarios. The motivation for proposing this dataset includes:

- The a need to provide a comprehensive set of video conferencing backgrounds (real and virtual) to support the development of robust forensic tools.
- To address the security and privacy concerns associated with virtual backgrounds, which can be exploited to deceive participants about the user’s actual location.
- To enable rigorous testing and verification of detection algorithms under various adversarial conditions, ensuring their reliability and robustness in real-world applications.

Our research is driven by the following questions:

- RQ1: How does the proposed dataset compare with state-of-the-art datasets in terms of applicability to multimedia forensics (MF) [30, 31] applications?
- RQ2: How can we verify and test the validity of our dataset under adversarial settings?
- RQ3: What is the performance and robustness of the detector against different attack scenarios when trained with our proposed dataset?
1.4 Contributions

The contributions of this paper are summarized as follows:

- We construct a large-scale video conferencing dataset of real and virtual backgrounds and make them publicly available for MF applications.
- We verify and test our dataset under different attack scenarios, and demonstrate its applicability via a detector by proposing an experimental evaluation of the classifier that can identify the type of backgrounds used in the video conferencing software.
- We evaluate the detection performance of our dataset under various types of attack scenarios to determine its robustness (e.g., changing the user’s background, different lighting conditions, and considering different post-processing operations), and show that we can perfectly identify a real from a virtual background in video conferencing software with a detection accuracy of 99.80% in an unaware scenario and 99.66% using the six co-mat [3] technique.

1.5 Organization

The paper is organized as follows. In Section 2, we present the state-of-the-art and a description of real and virtual video conferencing backgrounds. In Section 3, we provide the proposed system and threat models with the feature extraction methods: CRSPAM1372 [5] and six co-mat [3]. Then, we describe the methodology used in our research in Section 4. We report and discuss our experimental results in Section 5. Finally, we conclude the paper in Section 6 with some remarks and promising future work.

2 State of the art

In this section, we provide the fundamental concept of backgrounds used by several video conferencing software, then we discuss the related works on the security of virtual backgrounds.

2.1 Video conferencing backgrounds

In the following, we describe the types of video conferencing backgrounds used by participants during a video conferencing call.

- **Real Background**: It consists of the background captured by the camera of the user during a video conferencing meeting. This background displays the real scene where the user is located, which has not been changed, altered, or edited by any additional software.
- **Virtual Background**: One of the techniques used by several video conferencing software is to allow the user to hide or change the real background with a selected image. This image is known as the virtual background, and aims to cover the real background for different purposes (e.g., privacy, avoiding distraction during the meeting, etc).

Although replacing the real background with a virtual one has several benefits, it triggers security and privacy concerns for the participants. On the one hand, an adversary can potentially reveal or recover the hidden parts [17]. In this case, a user can use a virtual background as if it is their real background, possibly to claim being somewhere (e.g. office) when they
are not. To detect the latter case, there is a need to develop an efficient tool that can detect a virtual background from a real one.

### 2.2 Related work

Due to the amount of time in video conferencing, nearly half of the users prefer to work from their personal environments. In [10], the authors noticed similar privacy concerns about the virtual background; that is, users could conceal where they are with the virtual background to cheat other participants during the meeting. Nearly half of the users would rather work from their personal environments owing to the amount of time spent in video conferencing [33]. Similar privacy concerns related to the virtual background were raised in [10], where participants might use it to hide their location and eventually deceive others. Similar to the aforementioned video conferencing platforms, online conference or streaming video tools offer virtual background capabilities and some customization options. Users can, for instance, load their virtual backgrounds. In this case, the program removes the real background in the video streaming platform and substitutes it with a virtual background by semantically segmenting techniques. With the popularity of video meetings in everyday work, there is a greater necessity for user privacy protection [16].

Detecting the authenticity of a virtual background whether it’s real or manipulated can be referred to as detecting [3, 29, 40]. Prior works in MF propose several techniques to identify natural images. It is worth mentioning that the most current methods are based on Convolutional Neural Networks (CNNs). For instance, the method developed in [29] computed from three color image bands the co-occurrence matrices and fed them to the CNN model. Another approach proposed in [3] demonstrated that although Generative Adversarial Networks (GANs) can produce high-quality images with likely invisible traces, it is challenging to reconstruct a consistent relationship among the color bands. As an extension of the method proposed in [29], the authors improved the detection of images generated through GANs, and thus by considering as input the cross-band co-occurrences and the gray-level co-occurrences to the CNN detector. These co-occurrences are separately computed on the single bands. Consequently, the proposed CNN detector achieves good detection accuracy, and robustness for the intra-band method against post-processing. However, these approaches applied for GAN content detection are not robust against different attacks and post-processing [3, 29].

Additional research efforts have been conducted to investigate the privacy leakage of virtual backgrounds in video conferencing software [17]. The authors demonstrated the possibility of partial reconstruction of the visual environment despite the virtual background. They evaluated computer-vision based attacks against several video conferencing software (e.g., Zoom, Webex, and Google Meet), and explored the possibility of identifying objects in the user’s virtual background. Another work investigated privacy issues of the virtual background in video conferencing platforms [16]. However, these works cannot perfectly identify a real from a virtual background under adversarial attacks [16, 17]. More specifically, when the videos are captured with a high-quality (e.g., MacBook-Pro camera) with a virtual background, and the subject moves slowly, we observe the absence of tiny leakage pixels in the real background. Consequently, the work proposed in [17] is not efficient in this case. Moreover, the reconstruction of the real background from a virtual background will be a challenging task that we did not consider in this study. On the other hand, we found in [17] the possibility of fooling the detector through different attacks. In our study, we are motivated to evaluate the robustness of our detector against various attack scenarios, where we aim to distinguish between videos with real and virtual backgrounds. In particular, we
develop a novel high-quality video conferencing dataset where it is difficult to distinguish between videos that are recorded with real and virtual backgrounds.

In our study and differently from previous studies, we are interested in the security of Machine Learning and Deep Learning [17, 26], referred to as Adversarial Machine Learning (Adv-ML), and where the adversary can fool the detector easily through different attacks so-called adversarial attacks. Additionally, we noticed similar privacy concerns about the virtual background, where the users could conceal their location with the virtual background and cheat other participants in the meeting, such as considering the real background as a virtual one, or different lighting conditions.

3 System and threat model

In this section, we describe our proposed system and threat model considered to detect real background from a virtual one in video conferencing videos by considering different baseline methods and different ML techniques. Then, we present the feature extraction methods namely CRSPAM1372 [5] and six co-mat [3], which we considered as two main baseline feature extraction methods for verification.

3.1 System model

Our system considers a detection mechanism that distinguishes virtual backgrounds from real background videos by exploiting inconsistencies among spectral bands in all the frames. Specifically, we consider CNN as a detector that is fed by six co-occurrence matrices [3]; therefore, the CNN accepts the cross-band and spatial co-occurrence matrices, and then we train the detector to identify videos with a real background from a virtual background. The problem addressed in this work is schematically depicted in Fig. 1, i.e., distinguishing videos with real background $H_0$ from videos with virtual background $H_1$. After capturing videos from Zoom or other video conferencing software, we leverage the visibility of the pixels that are non-static and appear only occasionally in video frames in the real environment when the transition between foreground and background changes (e.g., during the movement of the

![Fig. 1](https://example.com/fig1.png)

(a) The detector is trained with CRSPAM1372 features [5] (b) The detector is trained with six co-mat features [3]

Fig. 1 Detection tasks for CRSPAM1372 and six co-mat features. (a) the detector is trained with CRSPAM1372 features, (b) the detector is trained with six co-mat features
person in the front, although there is not any pixel leakage). Afterward, we address recent feature extraction baseline methods considered in MF: CRSPAM1372 [5], and six co-mats strategy [3]. We note that we recorded the videos with different cameras and with different compression formats. Therefore, regarding the evaluation of the robustness of detectors, we considered JPEG compression with different quality factors, although we can consider other formats as well. In this study, we consider two types of classifiers in ML/DL: the Support Vector Machine (SVM) and the CNN. Specifically, the CNN detector has only been trained we trained the CNN detector in two modes: unaware and aware case. However, we trained the SVM only in an unaware modality, since the performance drops in this case; therefore, the aware case is not significant regarding the performance of the detector. In this context, the training was built according to the same scheme for the unaware detector without compression at the end. Here, the detector is unaware of all attacks [30, 31]. In this case, we applied different attacks to the unaware detector that degraded the performance to understand which kind of attack strategies degraded the performance of the detector. Therefore, in an aware case, we fine-tune the unaware detector with the attacks that degrade its performance.

3.2 Threat model

In this study, we address different scenarios that an adversary might consider to fool the detector:

- We assume that an adversary can access the real background and then consider it a virtual one. This scenario can be viewed as a new strategy and a harmful attack that the adversary can consider fooling the detector.
- We assume that the adversary recorded different videos with changing environmental lighting conditions to fool the detector.
- We considered different post-processing operations such as resizing, zooming, and rotation, to evaluate the robustness of the detector.
- We further conducted tests so as to ascertain the effectiveness of our method under various circumstances where the optical spectrum of an object is consistent from day to day. These involved recording videos nearby with the same light conditions over different days. The purpose was to see if, in spite of a constant spectral property in a scene, there still was a difference between virtual and real backgrounds this way. Our research findings show that even though the method maintains high accuracy in these cases, there are some small variations in detection rates which may require fine-tuning.
- We tested whether an adversarial network could be used to simulate light spectra for fake backgrounds such that their detector will reveal them. Specifically, we trained a neural network classifier on data from one location with the original background image to adjust the light spectrum of an object from another location to match its spectrum obtained at some other setting. Consequently, it can be inferred that while robustness is crucial when dealing with adversarial techniques that are more sophisticated than simple changes to lighting conditions, it might not be enough alone and hence further improvements are needed towards attaining accurate detection levels.

3.3 Feature extraction methods

In what follows, we describe the two main feature extraction approaches that we considered as a baseline method, namely: CRSPAM1372 [5] and six co-mat [3].
3.3.1 The CRSPAM1372 method

In terms of the detection task, we must select a large number of features capable of capturing different types of relationships between neighboring pixels. In contrast, the feature dimensionality must be limited when training with SVM. Indeed, utilizing a high number of feature sets may improve modeling performance. However, it would require the use of multiple classifier techniques such as ensemble classifiers [21]. In this case, the training process is challenging especially in adversary-aware modalities. In general, residual-based features have been used to identify a wide range of global manipulations [13, 34].

The feature set is generated by evaluating the residual in all directions (e.g., ←, →, ↑, ↓, ↗, ↘, ↙, ↖), then values truncating with a specific number of $T = 3$ (See Fig. 2), and finally, co-occurrences are estimated with order $d = 2$. Figure 2 demonstrates how to calculate residuals in every direction for one pixel, for simplicity. Then it is carried out with respect to all pixels and all color channels. Moreover, an SVM classifier is used to separate real from virtual backgrounds based on these characteristics.

To analyze the relationships between color channels, we consider CSRMQ1, a comprehensive model for color images introduced in [14] for steganalysis, and which consists of two components: The Spatial Rich Model component (SRMQ1) [13] and the 3-D color co-occurrence component. The features of SRMQ1 are calculated individually for each channel to maintain the same dimensionality and then combined together. The same noise residuals SRMQ1 are considered to produce features in 3-D color components but cross the color channels. The resulting feature space has a dimensionality of 12753 and cannot be employed with a single classifier such as SVM. Thus, for this problem, the authors in [5] used the SPAM686 [34] feature set with the same method as in CSRMQ1, and they refer to CRSPAM1372.

The first component in CRSPAM is derived by evaluating second-order co-occurrences of the first-order residuals $d = 2$. The features are then truncated with 3 ($T = 3$), computed separately for each channel, and then combined. Following that, in the second component, the residual co-occurrences concerning the three channels are computed. In the end, the CRSPAM feature set has 1372 dimensions in total. In this study, the CRSPAM1372 [5] is one of the baseline feature extraction techniques that we considered for extraction features from the real and virtual backgrounds and then fed to SVM and CNN only in an unaware scenario.

![Fig. 2](image_url)

*Fig. 2* Process the residuals for each pixel in all directions across all color channels which is shown here in an illustration for one pixel. The values should truncate with a number 3.
3.3.2 The six co-mat method

It has recently been proved that generated images may be revealed by evaluating inconsistencies in color channels [29]. The authors in [3] compute cross-band co-occurrences (spectral) and gray-level co-occurrences (spatial), then feed them to CNN to discriminate between authentic and malicious images. They claim that exploiting inconsistencies across the color bands (spectral) besides spatial ones, enhanced the robustness of the detector against different processing. In other words, the six co-occurrence matrices (six co-mats) method is a technique that entails computing co-occurrence matrices from the image. For each pixel, we calculate the occurrence of pixel intensity values in horizontal, vertical, and diagonal directions within a given window. This process is repeated for each color channel independently and then across the color bands resulting in six different co-occurrence matrices that represent spatial relationships in the image. These matrices are fed into a CNN that classifies background as either real or virtual. The CNN is trained using an SGD optimizer with a learning rate of 0.001 for 50 epochs; cross-entropy loss is employed to optimize model parameters during training phase of this algorithm. The proposed technique captures well spatial dependencies and it’s highly robust against various manipulations thus applicable for our detection task.

Therefore, the detector focuses more on spectral-pixel interactions rather than the features with spatial. Given an image \( I = (\alpha, \beta, \delta) \) with a size \( W \times L \times 3 \), where each color channel has an offset or displacement \( \Gamma = (\Gamma_{\alpha}, \Gamma_{\beta}) \) that is used with (1, 1) to compute the spatial co-occurrence matrix, and \( \Gamma' = (\Gamma'_{\alpha}, \Gamma'_{\beta}) \) is used with (0, 0) for inter-channel. The spatial co-occurrence matrices of red, blue, and green channels are expressed as follows:

\[
W_{\Gamma}(x, y; I_{Red}) = \sum_{\alpha=1}^{W} \sum_{\beta=1}^{L} \left\{ \begin{array}{ll}
1 & \text{if } \Gamma(\alpha, \beta) = x \text{ and } I_{Red}(\alpha + \Gamma_{\alpha}, \beta + \Gamma_{\beta}) = y \\
0 & \text{otherwise} \end{array} \right. \tag{1}
\]

We note that a similar (1) is applicable for Green and Blue channels. The CNN network’s input is provided by the tensor \( T_{\Gamma, \Gamma'} \), which has a dimension of \( 256 \times 256 \times 6 \) and consists of three spatial co-occurrence matrices for color channels or \( [C_{\Gamma}(I_{Red}), C_{\Gamma}(I_{Green}), C_{\Gamma}(I_{Blue})] \), as well as three cross-band co-occurrence matrices for the pairs \( [I_{R-G}], [I_{R-B}], \text{ and } [I_{G-B}] \) or \( [C_{\Gamma'}(I_{R-G}), C_{\Gamma'}(I_{R-B}), C_{\Gamma'}(I_{G-B})] \). \( I_{R-G}, I_{R-B}, \text{ and } I_{G-B}, \) refer to the colors Red and Blue, Green and Blue, and Red and Green, respectively. The construction of the cross-co-occurrence matrix (spectral) for the channels Green, Red, and Blue for an image \( I = (\alpha, \beta, 1) \), where \( x, y \) are integers between 0 and 255 is expressed as follows:

\[
W_{\Gamma'}(x, y; I_{R-G}) = \sum_{\alpha=1}^{W} \sum_{\beta=1}^{L} \left\{ \begin{array}{ll}
1 & \text{if } I(\alpha, \beta, 1) = x \text{ and } I(\alpha + \Gamma'_{\alpha}, \beta + \Gamma'_{\beta}, 2) = y \\
0 & \text{otherwise} \end{array} \right. \tag{2}
\]

Similarly, the same (2) is applicable for RB and GB. Additionally, the six tensors that include the Cross-Co-Net network’s \( (\Gamma_{\tau, \tau'}) \) input consist of the three co-occurrence matrices for the channels \( [I_{Red}, I_{Green}, \text{ and } I_{Blue}] \) and the three cross-co-occurrence matrices for the couple \( [I_{R-G}, I_{R-B}, I_{G-B}] \), and defined as follows:

\[
\Gamma_{\tau, \tau'}(x, y) = [W_{\Gamma}(x, y; I_{Red}), W_{\Gamma}(x, y; I_{Green}), W_{\Gamma}(x, y; I_{Blue})], \tag{3}
\]

As a result, six co-mat is another baseline feature extraction approach that we considered for our dataset verification in this study. Given the high detection performance achieved with
CNN in both unaware and aware scenarios, we did not examine the SVM in this case. The CNN detector scheme is shown in Fig. 3.

4 Experimental methodology

In this section, we discuss our dataset [1] and the experimental method that was examined in this research to determine how well the detector performs under different verification situations.

4.1 Proposed dataset

Due to the lack of a dataset regarding videos with real and virtual backgrounds for video conferencing software, we generated our dataset and made it publicly available for reproducibility, further applications, and a benchmarking reference for MF researchers applications [1]. In these datasets, we recorded videos using different video-conferencing applications (e.g., Zoom, Google Meet, and Microsoft Teams). The videos were recorded using different cameras of different qualities (e.g., MacBook Pro, Apple IPAD Pro, and MSI GF65 10UX laptop). Using this dataset, we recorded diverse real and virtual backgrounds. Moreover, to evaluate the effectiveness of our detectors, we recorded a video when a real background was considered as a virtual background.

4.1.1 Lighting conditions

We adjusted the indoor lamps manually with respect to lighting conditions and captured the video using different video conferencing applications. In various environmental settings, we filmed videos so as to evaluate how the performance of a detector might vary.

![Fig. 3 The six co-mat scheme of the CNN-based detector](image-url)
4.1.2 Subject and environment variations

Different videos have different subjects wearing various clothes in diverse colors. When it comes to the environment, several locations are chosen when shooting videos with real and virtual backdrops. In the future, we will consider more difficult situations for video recording with real and computer-generated backgrounds, which will involve the use of more than one camera.

4.1.3 Frame extraction

We extracted video frames of 1280 × 720 pixels. These frames have the same resolution to simplify the verification step. The frame examples for videos on real and virtual backgrounds are shown in Fig. 4.

4.2 Dataset generation standards

To make our dataset robust and generalizable, we followed specific guidelines and standards.

Complexity A diverse set of scenes was included at various levels of visual complexity. These include cluttered environments with many objects and simple backgrounds with few elements. Complexity is measured by the number of distinct objects and their arrangements within a given scene.

Location Videos were recorded at different locations to represent the different environmental settings. These included home offices, living rooms, public spaces, or professional settings. Each locality was selected to provide a wide range of contextual backgrounds.

Illumination Videos were captured under different lighting conditions to replicate different real-life situations. Thus, it varies from an environment that is well-lit using natural sunlight to a low-light setting with artificial lights. We also included situations with mixed lighting sources to test the resilience of this system to challenges posed by illumination.

Camera quality Because there are differences in the hardware used, the videos were filmed using different camera devices such as high-res webcams, integrated laptop cameras, and mobile device cameras. This variation helps explore the robustness of the system when dealing with different video qualities.

Virtual background types This dataset comprises various types of virtual backgrounds, including still images, moving animations, and blurring effects, and accounts for popular background types employed in video-conferencing software.

Environmental factors Factors such as background noise, movement within the frame, and the presence of more than one person have also been considered to simulate actual video conference occasion scenarios. These factors are important for examining whether the system can differentiate between a fake virtual background and a real background under normal usage conditions.

4.2.1 Computational cost

An overview of the memory requirements and computational time required to create and process the dataset is given:
Fig. 4  Examples of video frames with real and virtual backgrounds. (a) Examples of frames with real background; (b) Examples of frames with virtual background (users considered virtual background as a real background)
Computers used A MacBook Pro, an Apple IPAD Pro, and an MSI GF65 10UX laptop with 64GB of RAM were mentioned in the paper. They were used for video recording under various conditions.

NVIDIA system used This study states that NVIDIA GeForce RTX 3060 was used to process the dataset, including feature extraction and training detection models.

4.3 Network architecture

For the network architecture, we took into consideration the network proposed in [3] and decreased the number of layers to fit our target application. The network is composed of four convolutional layers, followed by two fully connected layers, as shown in Fig. 5. The first input layer contains a six-band input co-occurrence. In what follows, we explain the network architecture:

- Thirty-two filters of size $3 \times 3$ are considered for a convolutional layer with stride one and followed by a ReLu;
- Thirty-two filters of size $5 \times 5$ are considered for a convolutional layer with stride one and followed by a $3 \times 3$ max-pooling layer, plus considering 0.25 for dropout;
- Sixty-four filters of size $3 \times 3$ are considered for a convolutional layer with stride one and followed by a ReLu;
- Sixty-four filters of size $5 \times 5$ are considered for a convolutional layer with stride one and followed by a $3 \times 3$ max-pooling layer, plus considering 0.25 for dropout;
- A two dense layers by considering 256 nodes, followed by a dropout of 0.5 and a sigmoid activation.

4.4 Robustness analysis

Concerning the evaluation of the detector robustness, we recorded various videos with a real and virtual background while the subject moved slowly to avoid a pixel leakage. The baseline feature extraction algorithms CRSPAM1372 and six co-mat have been used for all frames. In this study, the SVM and CNN classifiers are fed separately with a CRSPAM1372 and six co-mat features. We verify the detectors’ robustness against several kinds of post-processing operations (see Fig. 1) by performing operations after each class $H_0$ and $H_1$. More precisely, we assume that the detector was trained with real and virtual background; post-processing (e.g., CLAHE, Resizing, and so on) was applied on both real and virtual background data.

![Fig. 5](https://example.com/fig5.png) Pipeline of the proposed network architecture. In the input, we consider $H_0$ and $H_1$ which are computed by the six co-mat method for a given background video frame, and the output distinguishes whether the background video frame is real or virtual.
to verify the robustness of the detector. Table 2 provides a list of post-processing operations that we considered for evaluating the detector’s robustness, and Table 4 provides a list of scenarios that we considered for verification and detailed in the following.

– **Scenario 1:** we examine the post-processing’s geometric alterations, such as resizing, zooming, and rotation. We employ median filtering, average blurring for filtering operations, and gamma correction, and we considered a Clip Limit Adaptive Histogram Equalization (CLAHE) method for contrast enhancement [43].

  Downscaling factors 0.8 and 0.5 are considered resizing scaling factors, whereas upscaling values of 1.4 and 1.9 are applied for zooming. For rotation operation, we employed degrees 5 and 10. Bicubic interpolation is employed for resizing, zooming, and rotation operations. Regarding median filtering and average blurring, different window sizes of $3 \times 3$, $5 \times 5$, and $7 \times 7$ are considered for both procedures. For gamma correction, we adjusted $\gamma$ to $\{0.9, 0.6, 1.3\}$, and the clip limit parameter for CLAHE was set to 2.0 and 4.0. Gaussian noise was generated with zero mean and standard deviations $\{0.8, 2\}$.

  Figure 6 shows the scheme that we considered for assessing the robustness of the detector. In this case, as we mentioned before, the post-processing operation is applied to the real and virtual background, and then the features are extracted by considering six co- mats. Finally, we evaluate the detector’s robustness against post-processing operations. Furthermore, we considered a challenging scenario in MF when two post-processing operations are applied sequentially; hence, in this case, we considered blurring followed by sharpening operations.

– **Scenario 2:** The internal lighting conditions are adjusted during video recording when 75% (low level of darkness), 50% (high level of darkness) of all lamps are turned on, and 100% (full illumination, means all lamps turned on), respectively. The key reason for examining this scenario is that contrast and lighting conditions are frequently performed during counterfeiting, which is useful for evaluating the detector’s performance.

– **Scenario 3:** We presented a challenging task to assess a detector’s robustness whenever an image of the real background is used as a virtual background in the video conferencing application (with an unaware detector).

  In this case, we suppose that the attacker has access to the subject’s real scene, either recording the scene in advance or reconstructing it by some estimation algorithm (like in [17]). To this purpose, we recorded a video with a real background without a subject and then utilized that video as a virtual background with a subject. For better video accuracy, we recorded a video in the same location with the same lighting circumstances, the same camera, and so forth. Figure 7 shows examples when an adversary can rebuild a real background or might gain access to the entire background and then uses the real background as a virtual one to deceive the detector in an unaware scenario.

| Method                        | Considered parameters |
|-------------------------------|-----------------------|
| CLAHE                         | 2.0, 4.0              |
| Resizing                      | 0.5, 0.8              |
| Zooming                       | 1.4, 1.9              |
| Rotation                      | 5-10                  |
| Median filtering and array blurring | $3 \times 3$, $5 \times 5$, $7 \times 7$ |
| Gamma correction             | $0.9, 0.6, 1.3$       |
– **Scenario 4:** a well-known issue with ML-based forensic tools is that they can be influenced by mismatch applications [22], or a generalization capability issue. Therefore, the classifier performs poorly when tested with a video that a detector has never seen during the training phase. We recorded several videos in Google Meet and Microsoft Teams

![Image of robustness procedure](image)

**Fig. 6** Robustness procedure considered in this paper to evaluate the detector’s performance for H0 and H1

![Examples of real background and attack video frames](image)

(a) Real background frame in location 1  
(b) Attack video frame considering real as a virtual background in location 1. In this case, the real and virtual background of location 1 is the same which can be considered as a novel harmful attack.

(c) Real background frame in location 2  
(d) Attack video frame considering real as a virtual background in location 2. In this case, the real and virtual background of location 2 is the same which can be considered as a novel harmful attack.

**Fig. 7** Examples of real background and attack video frames in different locations. (a) Real background frame in location 1, (b) Attack video frame considering real as a virtual background in location 1, (c) Real background frame in location 2, (d) Attack video frame considering real as a virtual background in location 2
to generalize the detector’s performance. Then, we test the detector’s performance that is trained with Zoom videos. Figure 8 shows examples of Google Meet and Microsoft Teams video frames with real and virtual backgrounds. As we mentioned, we assumed all video frames of the same height and weight. Moreover, in this case, all videos are captured with one camera. In the following scenario, we followed the same process with different cameras.

– **Scenario 5:** we also experimented with the mismatch of the software when videos were recorded in different cameras. In this scenario, we considered only Zoom video conferencing software that is recorded with different cameras. In this case, we captured videos on Apple IPAD Pro and MSI GF65 10UX laptops. Table 3 provides the accuracies achieved on different model cameras in the unaware scenario. We also considered a real-time condition (i.e., videos with real backgrounds) when we captured a video in the same lighting conditions and the same place. Afterward, we separated the frames in a real condition and computed the six co-mats. This scenario was employed in an aware scenario, and we achieved 100% accuracy if the video was captured with a MacBook Pro. Unfortunately, in a real-world scenario, if the video is captured in different cameras such as Apple IPAD Pro and MSI GF65 10UX laptop, the detector’s performance of Apple IPAD PRO decreases to 70%. However, for the MSI, the detector’s performance is around 97%. According to the experimental results, we remark a significant drop in the detector’s performance when the videos are captured with different cameras. To solve such issue, we suggest fine-tuning the aware detector with different videos recorded with different cameras.

![Fig. 8 Examples of real and virtual background frames in Google Meet and Microsoft Teams.](image)

(a) Real background in Google-Meet  
(b) Real background in Microsoft  
(c) Virtual background in Google-Meet  
(d) Virtual background in Microsoft
Table 3  Accuracies on different camera models for the unaware case scenario

| Model camera | IPAD Pro | MSI GF65 10UX |
|--------------|----------|---------------|
| Resolution   | 1080p    | 720p          |
| Accuracy (Test unaware model) | 70.45% | 97.63% |

4.4.1 Consistent optical spectrum analysis

We performed more tests to see how well our procedure handles situations where the optical spectrum of the object remains constant no matter what day it is. This included taking videos on subsequent days at the same spot with uniform lighting conditions. We wanted to establish whether, despite this spectral consistency in the scene, it could still differentiate between true and false backgrounds effectively. Our results show that though there are slight variations in detection performance revealing potential areas for further optimization, the method maintains high accuracy under these circumstances.

5 Experimental analysis

In this section, we discuss the analysis of our experiments by providing the considered settings, then we present the experimental results.

5.1 Experimental setup

In order to build our detector, the model was trained on 100000 patches for training (and validation) and 5000 patches for testing per class. All recorded frame videos under different conditions have a width of 1280 and a height of 720. The selection of 100000 patches was based on some preliminary experiments which show that this size balances computational feasibility against model performance. In order to build our detector, we trained the model with 100000 patches for training and validation (and 5000 patches for testing) per class. All recorded frame videos under different conditions have a width of 1280 and height of 720. We selected 100000 patches because some initial experiments suggested this size as a good trade-off between computational efficiency and model performance. To address the concern that using all frames from videos with the same background might be redundant, we changed our approach to sampling frames from each video instead of using every frame. This makes the training dataset more diverse and reduces potential overfitting. What we did was create six comats (six co-occurrence matrices) of these frames, where each comat is sized at $256 \times 256 \times 6$. Our choice of image resolution for this first baseline technique was guided by previous work [3] which showed that $256 \times 256$ worked well in extracting fine-grained spatial relationships between objects. On another note, for the second baseline technique, we used CRSPAM1372 as a split strategy for SVMs. The reason why CRSPAM1372 was chosen is that it had shown its effectiveness in capturing residual-based features during earlier steganalysis research [5]. All videos were recorded on a MacBook Pro with high quality but low noise. The aim behind this setup being higher quality recording is to reduce noise interference, which can affect feature extraction accuracy during model training.

For the optimizer, we employed stochastic gradient descent (SGD) [20] with a momentum of 0.9, a learning rate of 0.001, 50 training epochs, and a batch size of 20. We developed the network by employing the Keras API and TensorFlow as the backend [15]. We post-processed
the frames in Python using the OpenCV library. Following that, we evaluated each operation on 500 frames from the testing set (see Table 2). We deployed 3600 frames for training, 700 for validation, and 450 for testing to generate the aware model for identifying dangerous attacks (e.g., utilizing the real background as a virtual one). As a result, the CNN aware model was trained over 50 epochs with the SGD optimizer and 0.001 as the learning rate. Concerning the SVM classifier, the LibSVM library package [8] in the Matlab environment is considered and fed with CRSPAM1372 features. Additionally, we considered five cross-validations using a Gaussian kernel.

5.2 Experimental results

The performances of six co-mats and CRSPAM1372 in the different scenarios are given below. As shown in Table 4, these scenarios include post-processing effects, adversarial attacks, and illumination variations. To perform a comprehensive comparative analysis, we compared our method with other baselines under different conditions, including post-processing operations, lighting variations, and adversarial scenarios. Table 5 provides the accuracy comparisons for the six co-mats, CRSPAM1372, and other baseline methods across various settings. Our approach consistently outperforms the basic methods, particularly when there is a complex background or variability in light conditions, thus highlighting the robustness and generalizability of our methodology.

5.2.1 Performance of detectors in aware and unaware scenarios

When the detector is unaware of different attack scenarios, as shown in Table 5, the six co-mats matrices test set achieves 99.80% accuracy for a CNN detector. However, the CRSPAM1372’s accuracy for an SVM detector is only 50.00%, which is significantly lower than the six co-mat methods. We also explored a CNN unaware detector fed with CRPSAM1372 features and achieved 83.23% performance. As a result, as shown in the Table 5 unaware scenario, CNN always outperforms well than SVM. In this unaware scenario, we considered only the videos recorded in the Zoom video conferencing platform, therefore, the detector was only fed with features extracted from Zoom videos. To further support our claims, we also included quantitative metrics that show how well the proposed method improves data quality and regulatory decision-making. The detector was accurate 99.83% of the time under full

Table 4 Summary of the considered scenarios

| Scenarios | Summary |
|-----------|---------|
| Scenario 1 | The CLAHE method with post-processing’s geometric alteration (e.g., median filtering, average blurring, rotation, etc.). |
| Scenario 2 | Different lighting conditions, when 75%, 50%, and 100% of all lamps are turned on. |
| Scenario 3 | Challenging detection task where the real background is used as a virtual background. |
| Scenario 4 | Testing the detector against several videos captured from a wide range of video conferencing software (e.g., Google Meet, Zoom, Microsoft Team, etc.). |
| Scenario 5 | Testing the detector against several videos captured in a different cameras (e.g., Apple IPAD Pro, MSI GF65 10UX). |
| Scenario 6 | The optical spectrum must remain identical over different days in order to assess the ability of this method to differentiate between virtual and real backgrounds. |
lighting conditions (with every lamp turned on), 95.00% of the time when 75.00% of lamps were on, and 90.66% of the time with just half of the lamps turned on (Table 7). Similarly, to test the performance of the different camera models, we used a detector that achieved an accuracy rate of 97.63% with the MSI GF65 10UX laptop camera, and the Apple IPAD Pro camera achieved a percentage accuracy rate of 70.45%. Thus, these measures indicate that this technique can work in many different environments and with various hardware configurations, showing its relevance for practical applications in real-world situations as well. Therefore, even challenging situations do not present any difficulties for our method in recognizing physical or virtual backgrounds because, according to our results, it has always performed better than any other system under such circumstances.

To elaborate on six co-mat, the co-occurrence matrices are computed independently in each color (spatial) and then across the color bands (spectral), resulting in six co-occurrence matrices in this scenario. In this case, we have three matrices for a spatial band and three matrices for a spectral band, for a total of six matrices. We illustrate the inconsistencies between the spectral bands in a single real frame and spatial frame independently in Fig. 9, by considering the matrices size $250 \times 250$ in the six co-mat technique. Additionally, for frames with virtual backgrounds, the same matrices size $250 \times 250$ are considered, as shown in Fig. 10. As seen in both figures, spectral features can gather more information than spatial features. Hence, the detector can

|                      | Unaware scenario | Aware scenario |
|----------------------|------------------|----------------|
| Six co-mat (CNN)     | 99.80%           | 50.00%         |
| CRSPAM1372 (SVM)     | 83.23%           | 99.66%         |
| CRSPAM1372 (CNN)     |                  |                |

**Table 5** Six co-mat and CRSPAM1372 accuracy

![Illustration of different channel co-occurrences in one real frame. The channels are: (a) Red channel, (b) Green channel, (c) Blue channel, (d) Red-Green channel, (e) Red-Blue channel, and (f) Green-Blue channel. (the size of the matrices is $250 \times 250$)](image-url)
differentiate between real and virtual backgrounds by taking spectral properties into account. We performed these experiments using an unaware scenario when anti-forensically-edited frames were not included in the training set.

Aside from the unaware scenario, we performed simulations when the detector is slightly aware of attacks that an adversary may employ to deceive the detector (referred to as the aware scenario). It is worth mentioning that in an aware scenario, we only have the CNN classifier when fed by six co-mat features, since in an unaware scenario, the CRSPAM1372 with an SVM classifier, the detector performance significantly degrades to 50%. As a result, we only considered six co-mat with a CNN detector, which had a significant performance accuracy of 99.80% in the unaware scenario (see Table 5); Hence, we achieved the performance accuracy of 99.66% in the aware scenario. In another simulation scenario, we evaluated the robustness of the unaware CNN detector that is fed by six co-mat against different post-processing operations (e.g., resizing, zooming, and rotation) and with different parameters (e.g., median filtering, gamma correction, etc). The list of all post-processing and parameters we considered in this study are reported in Table 6. Concerning the experiment results, six co-mat achieves substantially superior robustness in all post-processings, even when the operation is applied with strong parameters. The worst-case scenario is Gaussian noise with standard deviation 2 which is one of the challenging tasks nowadays in MF, and where the accuracy reduces to 71.60% or somewhat less. Looking at the results of six co-mat, we remark that the CNN network’s accuracy is often around 99.00%. The main reason is that performing post-processing procedures nearly influences the spatial bands while having little effect on intra-channel interactions; consequently, the network may train with more robust features, resulting in a robust detector.

In the robustness analysis, the detector’s performance is efficient against post-processing operations such as contrast manipulations, geometric transformations, and filtering. In this
Table 6  Detector robustness accuracy in the presence of post-processing (six co-mat CNN))

| Operation                        | Parameter | Accuracy |
|----------------------------------|-----------|----------|
| Median filtering                 | 3 × 3     | 100%     |
|                                  | 5 × 5     | 100%     |
|                                  | 7 × 7     | 100%     |
| Gamma correction                 | 0.9       | 100%     |
|                                  | 0.6       | 99.20%   |
|                                  | 1.3       | 99.80%   |
| Average blurring                 | 3 × 3     | 100%     |
|                                  | 5 × 5     | 99.18%   |
|                                  | 7 × 7     | 99.32%   |
| CLAHE                            | 2         | 99.80%   |
|                                  | 4         | 99.80%   |
| Gaussian noise                   | 2         | 71.60%   |
|                                  | 0.8       | 88.40%   |
| Resizing                         | 0.8       | 100%     |
|                                  | 0.5       | 98.20%   |
| Zooming                          | 1.4       | 99.33%   |
|                                  | 1.9       | 97.30%   |
| Rotation                         | 5         | 98.25%   |
|                                  | 10        | 95.20%   |
| Blurring followed by sharpening  | –         | 99.30%   |

case, we applied well-known post-processing operations recently considered in MF to original videos with a real and virtual background, and then we evaluated the detector’s performance.

We also employed geometric transformations (e.g., resizing, rotation, and zooming), operations of filtering (e.g., blurring and median filtering), and contrast adjustment (e.g., adaptive histogram equalization which is mentioned as AHE [45], and gamma correction). For resizing operations, downscaling factors 0.8 and 0.5 are considered, while we applied the zooming upscaling with factors 1.4 and 1.9 are considered with bicubic interpolation. Regarding rotation, we considered 5 and 10-degree angles with bicubic interpolation. Additionally, for filtering operations, we set the window size for blurring and median filtering to [3×3, 5×5, and 7×7]. For noise addition, we considered Gaussian noise with standard deviations γ = 0.8 and 2 with zero mean. For gamma correction, we set γ ∈ {0.9, 0.6, 1.3}, while for adaptive Histogram Equalization (in particular, its refined version, Contrast Limited, implementation CLAHE [45]), the clip-limit parameter is set to 2.0, and 4.0. Finally, we assessed the robustness against the challenging task when two operations were applied to each other. Therefore, we employed blurring with window size 3 × 3 when followed by sharpening with kernel [-1, -1, -1, [-1, 9, -1], [-1, -1, -1]].

5.2.2 Performance in indoor lighting conditions

We recorded different videos with real and virtual backgrounds when we had diverse indoor light conditions to assess the detector’s performance. This real scenario is remarkably similar
to the gamma correction, CLAHE, that we covered in Section 5.2.1, and we referred to this scenario as physical attacks. As a result of physical attacks caused by changing indoor light conditions, discriminating frames with real and virtual backgrounds in the face of changing lighting conditions is one of the difficult tasks in MF. Additionally, this kind of attack scenario is commonly used during forgery creation, when the adversary needs to conceal any subject in content by changing the illumination [2, 5]. It is worth mentioning that this type of attack is among the most challenging tasks in MF. To evaluate the detector’s robustness against different lighting conditions, we captured the videos when 75% ($S_1$), 50% ($S_2$), and 100% ($S_3$) of all lamps are turned on in the physical environment. In Table 7, we report the accuracy of the detector under various lighting conditions.

### 5.2.3 Detector performance while considering real background as a virtual background

In this case, we assume that an adversary can rebuild a real background or potentially gain access to the entire background, such as considering the recent approach proposed in [17]. Therefore, the attacker can then utilize the real background as a virtual one to mislead the unaware detector. As we mentioned before in Table 5, we achieved a high detection accuracy of 99.80% for the CNN architecture where the features are extracted using the six co-mat method. Hence, in this scenario, for a real background and virtual background that has a real background, the six co-mat method is considered as feature extraction. As a consequence, the spatial and spectral features in real and video backgrounds are quite close to each other; in this case, the attacker entirely fools the unaware detector, which we refer to as a harmful attack scenario. Here, our strategy is to fine-tune the unaware detector with a variety of videos in which real backgrounds substitute virtual backgrounds. The fine-tuned model is known as the aware model. Figure 11 illustrates a real background in (a) when the adversary considered as a virtual background (b). We achieved 99.66% training accuracy in the aware model, in which we fine-tuned an unaware model, and then the aware model was evaluated using the frames in Fig. 11. In this scenario, the aware model’s test accuracy is 90.25%, which is acceptable to some extent for this scenario.

### 5.2.4 Detector performance against software mismatch

A common issue with ML-based approaches is generalization capabilities against a mismatched dataset, and which may degrade the detector’s performance. In this case, the detector performs poorly when evaluated with samples that were not observed during the training phase, indicating that the detector did not generalize well and can be regarded as a difficult task [4]. To ensure that our detector is robust to software mismatch, we evaluated the aware model on videos captured on the same devices but using different video conferencing software (e.g., Google Meet and Microsoft Teams). Figure 12 shows the considered scheme for assessing detectors in mismatch video conferencing software. In this scenario, each video consists of 500 frames with real and virtual backgrounds with a resolution of 1280 × 720. Table 8 shows the detector’s performance in a mismatch scenario.

| Table 7 | Accuracy of different lighting conditions |
|---------|------------------------------------------|
| Percentage of lamps turned on | S1 | S2 | S3 |
| Detection accuracy | 95.00% | 90.66% | 99.83% |
From the experimental results, it can be seen that the detection accuracies on Google Meet are somewhat better than those achieved on Microsoft Teams. This could be explained by Microsoft Teams videos being stored with a high noise ratio, as our detector is sensitive to noisy images. In general, the noise has a greater impact on our detection model since it degrades important features (see Table 6 gaussian-noise for example). Whereas capturing videos in Google Meet, the storage of videos has less noise compared with Microsoft Teams.

5.2.5 Performance of the detector against post-processing employed before JPEG compression

In this scenario, we considered median filtering, resizing, and rotation operations with the window size $[3 \times 3, 0.8, \text{and } 5]$, respectively before compression to assess the aware detector’s performance, which this scenario considered a challenging task in MF (see Fig. 13). The results in Table 9 confirm that the accuracy drops in the presence of various Quality Factors (QFs) but somehow is acceptable since the detector did not consider these features during the training phase. To clarify more, we note that JPEG compression is one of the harmful laundering attacks that an adversary might perform to deceive the detector.

MPEG4 is one of the common storage formats that is used in video recording. This format is also considered one of the compression formats. Therefore, we employed it as another scenario to evaluate the detector’s performance in an aware case. To that end, we recorded
Multimedia Tools and Applications

Table 8  Accuracy of aware model against mismatch applications

|                | Google Meet | Microsoft Teams |
|----------------|-------------|-----------------|
|                | 99.80%      | 63.75%          |

a video in Zoom video conferencing software on a MacBook Pro. Our experimental results showed a detection accuracy of 98.76%. It is quite an interesting result that proves our detector’s robustness against various compression formats (e.g., JPEG, MPEG4).

5.3 Discussion

In pre-recorded video conferencing scenarios, our suggested method has demonstrated great accuracy in detecting real and virtual backgrounds. The results show that this detector can differentiate between real and virtual backgrounds with high precision. In video conferencing sessions, where individuals should not deceive others about their actual location, it is important to have such a capability. Our experiments were conducted under different environmental conditions, including various lighting settings and camera qualities, which implies that the proposed approach is robust enough for use in different situations. According to the experimental results, using a CNN detector with the six co-mat method gives better results than the CRSPAM1372 method with an SVM detector. This means that combining both spatial and spectral features makes our model more effective at distinguishing between different types of virtual backgrounds. Moreover, even after applying multiple post-processing operations or

Fig. 13  Scheme for evaluating aware detector against median filtering considered before compression. The same procedure is considered for resizing and rotation.
adversarial conditions, the model remained robust as its accuracy did not drop significantly in any case. These findings confirm that this technique can work well under normal conditions where people are likely to use it frequently.

We also considered extending our detection approach to live meetings. By integrating the algorithm into the video conferencing software itself, the system can analyze frames on the fly to establish whether the background is authentic or not. Such analysis will require continuous processing of video frames while applying trained models, thereby identifying areas where there might be discrepancies indicative of using a virtual background during the conversation. In addition to being applicable in a live setting, where every participant’s backdrop could change dynamically over time, computational efficiency is another advantage exhibited by this model since it provides immediate feedback concerning the genuineness of each person’s appearance during such gatherings.

Furthermore, pixel leakage may occur within virtual backgrounds during live meetings, thus providing an opportunity to reconstruct the real background based on these pixel leakages. This idea relies on the knowledge that virtual backdrops are sometimes not fully opaque, allowing small parts of the actual surroundings to be visible. Over time, collecting and examining all these leaked pixels can help create a more accurate representation of the actual environment where an individual is located. This approach holds promise for future research and could greatly improve the reliability associated with detecting virtual backgrounds during live video conferencing.

In general, not only does our method work well in pre-recorded situations, but it also has the potential for use during live meetings. Additionally, integrating pixel leakage analysis strengthens the ability to detect and reconstruct real backgrounds, providing another interesting avenue for further investigation.

In multimedia forensics and security, detecting real or virtual backgrounds from meeting screenshots taken by a person is a challenging task. This difficulty arises because screenshots capture only a single frame, lacking the temporal information available in video recordings. Variations in lighting, compression artifacts, and the quality of the screenshot can further complicate the detection process. This paper [11], and [32] discusses this challenge well. Therefore, in the future, we will apply the strategy in [11] and [32] to see if their methods work to detect screenshots whenever the user has used a virtual background or not.

### 6 Conclusion and future works

The widespread use of last-generation video conferencing technology nowadays resulted in an exponential increase in the market size. This technology enables people to have face-to-face virtual meetings in different locations. Besides, it allows users to consider a virtual background to conceal their real environment due to privacy concerns. Nevertheless, in scenarios where the users should not hide their real locations, they may mislead other participants by claiming their virtual background as a real one. Due to the lack of datasets in this field,
it is challenging to validate real and virtual backgrounds. Moreover, given the popularity of
different machines and deep learning techniques that can be considered for different tasks,
the adversary can easily fool the detector via different scenarios. In this paper, we presented
a new large dataset that consists of a set of real and virtual backgrounds from different video
conferencing software. We validated the proposed dataset by conducting exhaustive analysis
scenarios. In particular, we considered two recent baseline feature extraction methods: the
CRSPAM1372 method and the six co-mat method. These techniques are commonly used
in MF along with two classifiers that serve as detectors: the SVM and the CNN. Then, we
evaluated these two detectors against different scenarios by:
(i) evaluating the detector’s performance in aware and unaware cases.
(ii) evaluating the detector’s performance against different indoor lightning conditions.
(iii) considering the real background as a virtual background.
(iv) employing different post-processing operators to assess the detector’s performance.
(v) evaluating the detector’s performance against software mismatch.
(vi) assessing the detector’s performance when the adversary considers different post-
processing operations before JPEG compression.
(vii) employing MPEG4 compression methods to evaluate the detector’s performance. Dur-
ing the verification phase, our experimental results showed that using the six co-mat
approach increased the detector’s robustness against various scenarios. We discovered
that the CRSPAM1372 approach when compared to six co-mat, is not a satisfactory
approach, as we achieved 50.00% accuracy.

Future studies will be devoted to more additional scenarios that consist of:
(i) investigating the detector’s performance when the attacker tries to modify the cross-band
relationship in six co-mat to deceive the detector.
(ii) evaluating the detector’s performance against various types of adversarial attacks (e.g.,
FGSM, JSMA, C&W, etc).
(iii) exploring generalization capabilities when various videos are taken with various cam-
eras, and we will update our dataset in the near future as a second version. In this
direction, we aim to increase the size and diversity of our dataset.
(iv) assessing the detector’s robustness when the attacker uses a video rather than an image
as a virtual background.
(v) To ensure that our approach is valid, we used a dataset that we created. It consisted of
different video conferencing scenarios so it has various environmental conditions and
camera qualities. Even though our method was currently validated with the use of self-
developed datasets having both real and virtual backgrounds, we know the necessity of
trying out our solution with external sets. Nevertheless, common meeting video data
sets found in public like AMI Meeting Corpus [23] or ICSI MRDA Corpus [36] do not
have virtual background features. In future investigations; thus far for this paper only,
we propose to introduce new test cases into wider arrays across multiple types while
evaluating them against each other according to their robustnesses then take into account
generalizations based on these results (these additional sets may contain environments
created virtually).

Acknowledgements We would like to thank the students Mohammad Hajian Berenjestanaki, and Yoosef
Habibi from the Department of Mathematics, University of Padua, Italy, for helping us to create a dataset.
Moreover, we would like to acknowledge Prof. Berrin Yanikoglu from the Faculty of Engineering and Nat-
ural Sciences, Sabanci University, Istanbul, Turkey, for her fruitful comments to improve the quality of the
manuscript.

Funding The authors did not receive support from any organization for this work.
Data Availability  The datasets generated during and/or analyzed during the current study are available in the Zenodo repository: [https://zenodo.org/records/5572910]

Declarations

Conflicts of Interest  The authors declare they have no conflict of interest.

Open Access  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Dataset for Real and Virtual Backgrounds of Video Calls (2021). https://zenodo.org/record/5572910
2. Barni M, Costanzo A, Nowroozi E, Tondi B (2018) CNN-based detection of generic contrast adjustment with JPEG post-processing. In: 2018 25th IEEE International conference on image processing (ICIP), IEEE, pp 3803–3807
3. Barni M, Kallas K, Nowroozi E, Tondi B (2020) CNN detection of gan-generated face images based on cross-band co-occurrences analysis. In: 2020 IEEE International workshop on information forensics and security (WiFS), IEEE, pp 1–6
4. Barni M, Nowroozi E, Tondi B (2017) Higher-order, adversary-aware, double JPEG-detection via selected training on attacked samples. In: 2017 25th European signal processing conference (EUSIPCO), IEEE, pp 281–285
5. Barni M, Nowroozi E, Tondi B (2018) Detection of adaptive histogram equalization robust against JPEG compression. In: 2018 International workshop on biometrics and forensics (IWBFF), IEEE, pp 1–8
6. Böhme R, Kirchner M (2013) Counter-forensics: attacking image forensics. In: Digital image forensics, pp 327–366
7. Bonettini N, Bestagini P, Milani S, Tubaro S (2021) On the use of Benford’s law to detect gan-generated images. In: 2020 25th International conference on pattern recognition (ICPR), pp 5495–5502. https://doi.org/10.1109/ICPR48806.2021.9412944
8. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3). https://doi.org/10.1145/1961189.1961199
9. Choi Y, Choi, Kim M, Ha JW, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
10. Conti M, Milani S, Nowroozi E, Orazi G (2021) Do not deceive your employer with a virtual background: a video conferencing manipulation-detection system. arXiv:2106.15130
11. Ferreira A, Nowroozi E, Barni M (2021) VIPPrint: validating synthetic image detection and source linking methods on a large scale dataset of printed documents. J Imaging 7(3):50. https://doi.org/10.3390/jimaging7030050
12. Filler T, Fridrich J, Goljan M (2008) Using sensor pattern noise for camera model identification. In: 2008 15th IEEE International conference on image processing, IEEE, pp 1296–1299
13. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inf Forensics Secur 7(3):868–882. https://doi.org/10.1109/TIFS.2012.2190402
14. Goljan M, Fridrich J, Cogranne R (2014) Rich model for steganalysis of color images. In: 2014 IEEE International workshop on information forensics and security (WiFS), IEEE, pp 185–190
15. Gulli A, Pal S (2017) Deep learning with Keras. Packt Publishing Ltd
16. He S, Lei Y (2021) The privacy protection effectiveness of the video conference platforms’ virtual background and the privacy concerns from the end-users. arXiv:2110.12493
17. Hilgefort JM, Arp D, Rieck K (2021) Spying through virtual backgrounds of video calls. Proceedings of the 14th ACM workshop on artificial intelligence and security. https://doi.org/10.1145/3474369. https://doi.org/10.1145/3474369.3486870
18. Karras T, Laine S, Lehtinen J (2019) Analyzing and Improving the Image Quality of StyleGAN. CoRR abs/1912.0
19. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2017) Enlarging the training dataset is beneficial for training generative adversarial networks. CoRR abs/1710.1
20. Ketkar N (2017) Stochastic gradient descent. In: Deep learning with Python, Springer, pp 113–132
21. Kodovsky J, Fréchot J, Holub V (2012) Ensemble classifiers for steganalysis of digital media. IEEE Trans Inf Forensics Secur 7(2):432–444.
22. Kodovsky J, Sedighi V, Fréchot J (2014) Study of cover source mismatch in steganalysis and ways to mitigate its impact. In: Alattar AM, Memon D, Heitzenrater CD (Eds) Media watermarking, security, and forensics 2014, vol 9028, p 90280J. International society for optics and photonics, SPIE. https://doi.org/10.1117/12.2039693
23. Kraaij W, Hain T, Lincoln M, Post W (2005) The ami meeting corpus. In: Proc international conference on methods and techniques in behavioral research, pp 1–4
24. Li H, Li B, Tan S, Huang J (2020) Identification of deep network generated images using disparities in color components. Signal Process 174:107616. https://doi.org/10.1016/j.sigpro.2020.107616
25. Mekdad Y, Bernieri G, Conti M, El Fergougui A (2021) The rise of ics malware: a comparative analysis. In: European Symposium on Research in Computer Security, Springer, pp 496–511
26. Milani S, Tagliasacchi M, Tubaro S (2014) Discriminating multiple jpeg compressions using first digit features. APSIPA Trans Signal Inform Process 3:e19. https://doi.org/10.1109/ATSP.2014.19
27. Nowroozi E, Dehghantanha A, Parizi RM, Choo KKR (2021) A survey of machine learning techniques in adversarial image forensics. Comput Sec 100:102092
28. Nowroozi E, Habibi Y, Conti M (2024) Spritz-ps: validation of synthetic face images using a large dataset of printed Images. Multimed Tools Appl pp 1–29
29. Nulab: Adjusting to remote work during the COVID-19 crises (2020). https://nulab.com/learn/collaboration/adjusting-to-remote-work/
30. Roy A, Dixit R, Naskar R, Chakraborty RS (2020) Digital Image Forensics: Theory and Implementation. Springer
31. Shriberg E, Dhillon R, Bhagat S, Ang J, Carvey H (2004) The icsi meeting recorder dialog act (mrda) corpus. In: Proceedings of the 5th SIGdial workshop on discourse and dialogue at HLT-NAACL 2004, pp 97–100
32. Tang H, Liu J, Yan S, Yan R, Li Z, Tang J (2023) M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In: Proceedings of the 31st ACM international conference on multimedia, pp 1719–1728
33. Tang H, Yuan C, Li Z, Tang J (2022) Learning attention-guided pyramidal features for few-shot fine-grained recognition. Pattern Recogn 130:108792
34. Verde S, Bestagini P, Milani S, Calvagno G, Tubaro S (2021) Focal: a forgery localization framework based on video coding self-consistency. IEEE Open J Signal Process 2:217–229
35. VideoZeus (2020) How To FAKE A Zoom Meeting–Clone Yourself On Video. https://www.youtube.com/watch?v=eprkLhdqh4U
42. Xing P, Tang H, Tang J, Li Z (2024) Adps: Asymmetric distillation postsegmentation for image anomaly detection. IEEE Trans Neural Netw Learn Syst

43. Yadav G, Maheshwari S, Agarwal A (2014) Contrast limited adaptive histogram equalization based enhancement for real time video system. In: 2014 International conference on advances in computing, communications and informatics (ICACCI), IEEE, pp 2392–2397

44. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

45. Zuiderveld K (1994) Contrast limited adaptive histogram equalization. Graphics gems pp 474–485

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ehsan Nowroozi is a senior IEEE/ACM professional member and IEEE IFS-TC with expertise in AI for Security. He currently serves as a senior lecturer (associate professor) in cybersecurity at Ravensbourne University, London, United Kingdom. He received his Ph.D. from the University of Siena, Italy. He holds several postdoctoral positions from Queen’s University Belfast, the United Kingdom, Siena, and Padua Universities in Italy, and Sabanci University in Turkey. His main research interests include AI for Security, Adversarial Machine Learning, and Multimedia Forensics and Security. He is also a grant reviewer for EPSRC-UKRI from 2024 to the present. Additionally, he became an editorial board member of IEEE Transactions on Network and Service Management in July 2024 and is an active reviewer for prestigious journals such as IEEE TIFS, IEEE TNSM, IEEE TNNLS, and others.

Yassine Mekdad Received the Ph.D. degree in cybersecurity from Moulay Ismail University of Meknes, Morocco. He has been awarded with a fellowship by Fondazione Ing. Aldo Gini and holds a guest researcher position with the SPRITZ research group at the University of Padua, Italy. He has also been awarded with a Fulbright fellowship. Currently, he works as a cybersecurity researcher at the Cyber-Physical Systems Security Lab (CSL) at Florida International University, Miami, FL, USA. His research interest principally covers security and privacy problems in the Internet of Things (IoT), Industrial Internet-of-Things (IIoT), and Cyber-physical systems (CPS). Furthermore, he works on research problems at the intersection of the cybersecurity and networking fields with an emphasis on their practical and applied aspects. He is a member of the IEEE Cybersecurity Community and IEEE Young Professionals.
Mauro Conti Received the Ph.D. degree from the Sapienza University of Rome, Italy. He is a Full Professor with the University of Padua, Italy, and an Affiliate Professor with TU Delft and University of Washington, Seattle. He was a Postdoctoral Researcher with Vrije Universiteit Amsterdam, The Netherlands. In 2011, he joined as an Assistant Professor with the University of Padua, where he became an Associate Professor in 2015, and a Full Professor in 2018. He has been a Visiting Researcher with GMU, UCLA, UCI, TU Darmstadt, UF, and FIU. His research is also funded by companies, including Cisco, Intel, and Huawei. He has been awarded with a Marie Curie Fellowship by the European Commission, and with a Fellowship by the German DAAD.

Simone Milani Was born in Camposampiero, Italy, in 1978. He received the Laurea degree in telecommunication engineering and the Ph.D. degree in electronics and telecommunication engineering from the University of Padova, Padova, Italy, in 2002 and 2007, respectively. He was a Visiting Ph.D. Student with the University of California at Berkeley, Berkeley, CA, USA, in 2006. He was a Postdoctoral Researcher with the University of Udine, Udine, Italy, the University of Padova, and the Politecnico di Milano, Milan, Italy, from 2007 to 2013. He has been with STMicroelectronics, Agrate, Italy. He is currently an Assistant Professor with the Department of Information Engineering, University of Padova. His research interests include digital signal processing, image and video coding, 3D video processing and compression, joint source-channel coding, robust video transmission, distributed source coding, multiple description coding, and multimedia forensics.

Selcuk Uluagac Leads the Cyber-Physical Systems Security Lab at Florida International University, focusing on security and privacy of Internet of Things and Cyber-Physical Systems. He has a Ph.D. and M.S. from Georgia Institute of Technology, and M.S. from Carnegie Mellon University. In 2015, he received the US National Science Foundation CAREER award and US Air Force Office of Sponsored Research’s Summer Faculty Fellowship, and in 2016, Summer Faculty Fellowship from the University of Padova, Italy. Currently, He serves on the editorial boards of the IEEE Transactions on Mobile Computing, Elsevier Computer Networks, and the IEEE Communications and Surveys and Tutorials.
Authors and Affiliations

Ehsan Nowroozi\textsuperscript{1,2} · Yassine Mekdad\textsuperscript{3} · Mauro Conti\textsuperscript{4} · Simone Milani\textsuperscript{5} · Selcuk Uluagac\textsuperscript{6}

Ehsan Nowroozi
ehsan.nowroozi65@ieee.org; e.nowroozi@rave.ac.uk
Yassine Mekdad
ymekdad@fiu.edu
Mauro Conti
mauro.conti@unipd.it
Simone Milani
simone.milani@dei.unipd.it
Selcuk Uluagac
suluagac@fiu.edu

1 Queens University Belfast, Belfast, United Kingdom
2 Department of Business and Computing, Ravensbourne University London, London, United Kingdom
3 Cyber-Physical Systems Security Lab, Florida International University, Miami, Florida, USA
4 Department of Mathematics, University of Padua, Padua, Italy
5 Department of Information Engineering, University of Padua, Padua, Italy
6 Cyber-Physical Systems Security Lab, Florida International University, Miami, Florida, USA