

FOCAL: A Forgery Localization Framework Based on Video Coding Self-Consistency

SEBASTIANO VERDE ¹ (Student Member, IEEE), EDOARDO DANIELE CANNAS ² (Student Member, IEEE),
PAOLO BESTAGINI ² (Member, IEEE), SIMONE MILANI ¹ (Member, IEEE),
GIANCARLO CALVAGNO ¹ (Member, IEEE), AND STEFANO TUBARO ² (Senior Member, IEEE)

¹Department of Information Engineering, University of Padua, 35131 Padova, Veneto, Italy

²Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy

CORRESPONDING AUTHOR: SEBASTIANO VERDE (e-mail: sebastiano.verde@dei.unipd.it)

This work was supported in part by the University of Padova project Phylo4n6 prot. BIRD165882/16 and in part by DARPA and Air Force Research Laboratory (AFRL) under Agreement FA8750-16-2-0173.

ABSTRACT Forgery operations on video contents are nowadays within the reach of anyone, thanks to the availability of powerful and user-friendly editing software. Integrity verification and authentication of videos represent a major interest in both journalism (e.g., fake news debunking) and legal environments dealing with digital evidence (e.g., courts of law). While several strategies and different forensics traces have been proposed in recent years, latest solutions aim at increasing the accuracy by combining multiple detectors and features. This paper presents a video forgery localization framework that verifies the self-consistency of coding traces between and within video frames by fusing the information derived from a set of independent feature descriptors. The feature extraction step is carried out by means of an explainable convolutional neural network architecture, specifically designed to look for and classify coding artifacts. The overall framework was validated in two typical forgery scenarios: temporal and spatial splicing. Experimental results show an improvement to the state of the art on temporal splicing localization as well as promising performance in the newly tackled case of spatial splicing, on both synthetic and real-world videos.

INDEX TERMS Forgery detection, multimedia forensics, video codecs, video forensics.

I. INTRODUCTION

The assessment of authenticity for video sequences is nowadays a paramount task in a variety of contexts, such as citizen journalism and fake news debunking, as well as evidence validation in legal procedures and fraud detection. This concern has gained importance during the last years because of the wide availability of powerful and easily-operable video editing programs (e.g., Adobe Premiere, Apple Final Cut, etc.) and the wide-spread use of video data in communication and documenting activities. Moreover, the development of deep learning solutions for the automatic creation and editing of image and video contents have posed new challenges to forensic analysts, since a malicious user has the opportunity to create fake contents that overcome most of the existing detectors.

As a matter of fact, forensic analysts have been constantly investigating innovative and accurate solutions for forgery detection and localization. Among the first strategies being

proposed, we can find detectors that identify the acquisition device [1], [2], physical inconsistencies [3], video recapturing [4], frame deletion and insertion [5], [6], or codec-related operations [7]–[9]. Most of these detectors verify the *self-consistency* [10], [11] of video processing footprints, i.e., the uniformity of traces left on the signal across different frames and regions of the video sequence. Whenever an external element is included within an original image or video, the forensic footprints in the altered region change with respect to untouched ones. Revealing such a discrepancy allows detecting the possible presence of a forgery.

Extending the preliminary work in [12], the current paper proposes a FORgery loCALizer (FOCAL) that checks the self-consistency of multiple and independent forensic traces related to video coding (Fig. 1). Differently from the previous work where forgeries consisted in concatenating video sequences from different sources (*temporal splicing*), this new

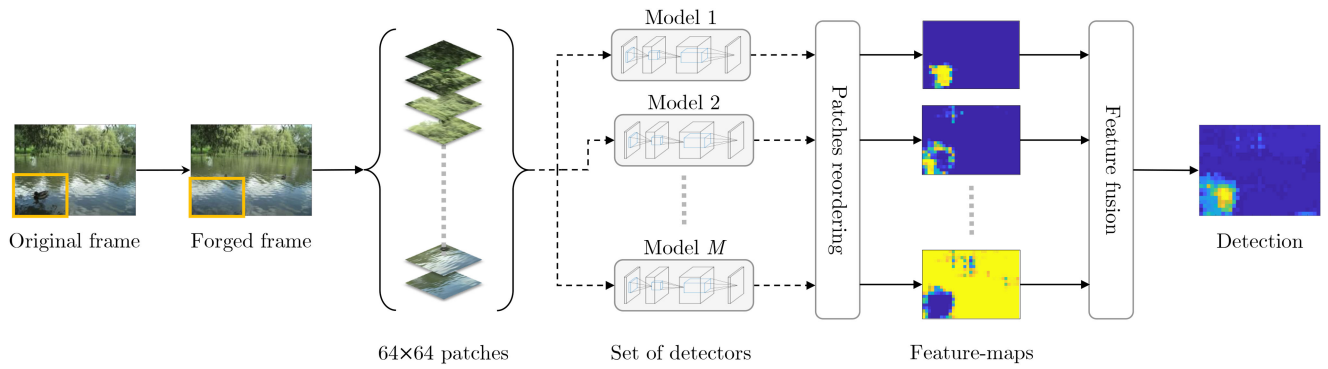


FIGURE 1. Forgery loCALizer (FOCAL) framework. A forged video-frame is split into 64×64 patches and fed to a set of pre-trained detectors (e.g., classifiers of the video coding standard and quality). Extracted features are rearranged into feature-maps and a fusion function merges them into a single detection heat-map. Dashed and solid lines are used to denote patch-wise and frame-wise operations, respectively.

approach is also able to precisely localize an altered region within a single frame (*spatial splicing*) as well as along time dimension.

Given an input video, each frame is split into smaller patches, and a feature vector is extracted from each one of them. The set of features corresponds to the output values of the final softmax layers from multiple convolutional neural networks (CNNs) dedicated to the classification of different coding parameters, such as coding standard and quality level. These CNNs share an *explainable* architecture, which was specifically designed to look for coding artifacts by aligning the receptive fields of the network filters to the quantization block boundaries, where the most significant traces are typically visible.

An unsupervised fusion technique was designed to merge the outputs of these heterogeneous feature descriptors into a human-readable heatmap, which characterizes each frame-patch from the analyzed video with a likelihood measure that models the probability of being forged. This approach also makes the framework scalable and extendible at will, allowing the introduction of additional detectors and feature descriptors to contribute to the overall heatmap.

Experimental validation takes into account different forgery setups, such as temporal and spatial splicing in controlled and uncontrolled environments. Results show that the proposed solution is able to improve the performance of [12], thanks to the newly adopted network architecture, and to obtain convincing results in the detection of local forgeries as well, with an area under the curve (AUC) of the receiver operating characteristic (ROC) curve of 0.94.

The rest of the paper is organized as follows. Section II briefly overviews the literature on video forgery detection and localization, distinguishing between temporal and spatial forgeries. Section III formally defines the problem addressed by the paper and the notation used. Section IV presents the proposed CNN for extracting coding-related features, with special emphasis on the architectural choices. Section V illustrates our forgery localization framework, from the feature extraction step to the final feature fusion and heatmap generation, in both temporal and spatial forgery cases. Section VI

reports all the details about the experimental setup, the training phase, the generation of the synthetic dataset and the obtained results. Finally, Section VII concludes the paper and outlines possible future work.

II. RELATED WORK

In recent years, video authentication has emerged as a novel and challenging research field [13], [14] leading to the development of algorithms and tools capable of estimating whether a video sequence is original or not. Most of the proposed approaches identifies two different types of forgeries: temporal and spatial splicing. The first case refers to videos that have been modified through the inclusion or deletion of some frames into or from the original sequence. In the second one, the content of individual frames is modified, e.g., with a cut-and-paste of a region (inclusion/removal of an object from the scene) or performing an upscale-crop editing (where an object located in an outermost part of the video is removed by cropping the frames). Furthermore, it is worth mentioning double/multiple compression. Since video sequences are usually available in compressed format and any modification must be performed in the pixel domain, videos need to be re-encoded every time a forgery is operated. For this reason, forged sequences typically exhibit the presence of multiple compression artifacts.

Among the strategies for detecting the insertion/deletion of frames, some algorithms exploit the correlation and similarity between frame characteristics [15], [16]; if some temporal patterns do not follow the expected trend, the algorithm raises an alarm. Similarly, other solutions identify regular patterns in the camera noise signals; whenever there are repetitions [17] or oddities [18], [19] due to the fact that forged frames were taken by a different camera, the algorithm reports an anomaly. Deletions can be highlighted by spotting irregularities in the video motion statistics, obtainable through the analysis of the optical flow [20], [21], interpolation [22], or standard block matching [23]. The correlation in prediction residual information [24], [25], in texture patterns [26] and in brightness [27] can be exploited as well.

Spatial forgeries can be detected by checking the consistency of forensic traces left on the video sequence by the acquisition device or by different encoding algorithms. The strategy proposed in [28] exploits the scaling invariance of the minimum average correlation energy Mellin radial harmonic (MACE-MRH) correlation filter to unveil traces of upscale-crop forgeries. Similarly, the impact of a spatial splicing on the encoding of motion vectors in interlaced videos can be analyzed to reveal traces of possible alterations [29]. Object removal can be exposed by detecting discrepancies in the motion vectors too [30] and also through a combination of different steganalysis features [31]. Copy-move object removal can be revealed by exploiting local descriptors, such as histogram of oriented gradients (HoG) [32] or scale-invariant feature transform (SIFT) descriptors [33]. The same type of forgery can also be detected by analyzing the spatial and temporal correlation among frames [34], Zernike moments [35], and optical flow similarities [36]. Together with intra-frame similarities and discrepancies, by comparing a plausible model with what is estimated directly from the pixels it is possible to expose physical inconsistencies in scene illumination and object motion [3].

Revealing traces of multiple compression on the analyzed video sequence allows an effective detection of editing operations. A first insight was provided in [37], followed by several researches extending to videos the coding footprints identified for images [9], [38]. In the context of double MPEG detection, the misalignment of the group-of-pictures (GOP) structures related to the first and second encoding can be informative as well; whenever the coding parameters change between two consecutive encoding steps, a superposition of heterogeneous artifacts appears on the video and can be detected [39]. Similarly, the simultaneous presence of traces related to incompatible coding parameters or formats is investigated in several papers [8], [40], [41]. Furthermore, whenever a video sequence is compressed twice, it is possible to observe some peculiar noise patterns. In [42] the authors propose a first-order Markov statistics for the differences between quantized discrete cosine transform (DCT) coefficients along different directions, while the solution in [43] employs a modified Huber Markov random field (HMRF) model; these methods permit assessing whether a whole sequence of frames is authentic or tampered (e.g., compressed twice) but, differently from the proposed one, do not allow to precisely localize a forgery operation.

Recent approaches are setting up a new trend in video forgery detection by using deep neural networks. In [44] the authors propose a recursive autoencoder (implemented with the LSTM architecture, to exploit temporal dependencies) to learn a feature representation of pristine videos and detect forgeries as outliers of the learned model. In [12] two CNNs are independently trained to extract codec and quality related features with the purpose of detecting temporal inconsistencies, showing that the combination of heterogeneous detectors enhances the overall performance. Some studies have also been addressed to expose the newly appeared threat of

AI-generated highly-realistic forgeries, also known as DeepFakes. The detection is carried out by means of eye-blinking analysis [45] and combinations of deep learning models such as CNNs and recurrent neural networks (RNNs) [46]. However, these methods are specifically tailored to the problem of detecting fake faces, thus they fail in spotting general video forgeries.

Following the trend of fusing multiple features to increase the overall accuracy, the proposed strategy combines a set of coding-related features obtained from different CNNs. The architecture of these networks was designed in awareness of where and how compression artifacts appear, as described in Section IV.

III. PROBLEM DEFINITION

The purpose of the FOCAL framework is detecting and localizing temporal and spatial splicing operations on video sequences. Here we provide a formal definition of the tackled problem and the notation used throughout the paper.

Let us define a video sequence \mathbf{X} as an array of N frames denoted by \mathbf{X}_n ,

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N], \quad (1)$$

where each frame is a matrix of pixels X_{uv} ,

$$\mathbf{X}_n = [X_{uv}]_n. \quad (2)$$

Pixel coordinates are $(u, v) \in \mathcal{U} \times \mathcal{V}$, where $U = |\mathcal{U}|$ and $V = |\mathcal{V}|$ are the amount of pixel per column and row, respectively.

Definition 1: Let \mathbf{X} and \mathbf{Y} be two sequences of frames with size U_X, V_X, N_X and U_Y, V_Y, N_Y , respectively. The two sequences are *spliceable* if $U_X = U_Y$ and $V_X = V_Y$.

Without loss of generality, we will define the two types of forgeries addressed by this work for spliceable sequences only.

Definition 2: Let \mathbf{X} and \mathbf{Y} be two spliceable sequences. A *temporal splicing* is a function \mathcal{T} that concatenates \mathbf{X} and \mathbf{Y} into a single sequence:

$$\mathcal{T}(\mathbf{X}, \mathbf{Y}) = [\mathbf{X}_1, \dots, \mathbf{X}_{N_X}, \mathbf{Y}_1, \dots, \mathbf{Y}_{N_Y}]. \quad (3)$$

The resulting sequence is called *temporally-spliced* and the frame-index $N_X + 1$ is the *splicing point*.

Definition 3: Let X_{uv} and Y_{uv} be two frames from spliceable sequences \mathbf{X} and \mathbf{Y} , respectively. Let $\mathcal{R} \subset \mathcal{U} \times \mathcal{V}$ be a subset of pixel coordinates. A *spatial splicing* is a function \mathcal{S} that substitutes pixels in \mathcal{R} of one frame with pixels in \mathcal{R} of the other frame:

$$\mathcal{S}(X_{uv}, Y_{uv}, \mathcal{R}) = \begin{cases} Y_{uv}, & (u, v) \in \mathcal{R} \\ X_{uv}, & \text{otherwise} \end{cases}. \quad (4)$$

The resulting sequence is called *spatially-spliced* and \mathcal{R} is the *spliced region*.

Given a video sequence under analysis, with no additional information available except for the pixel values, we aim to localize possible splicing points or spliced regions.

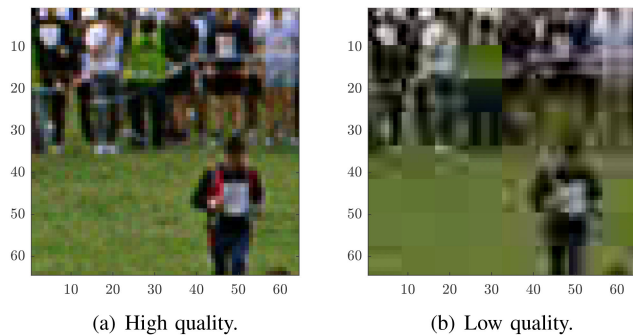


FIGURE 2. Block-artifacts at different encoding qualities in a 64×64 patch. The grid of 8×8 blocks is particularly evident in 2(b).

IV. CODING FEATURES

The core of our forgery localizer consists in a convolutional neural network specifically designed to detect and classify traces left by video-coding algorithms. Understanding which coding scheme and parameters were used to encode a video clip, by only looking at the pixel domain, represents a challenging task even for a human observer. However, almost anyone is able to perform a rough classification on the perceptive quality of a video, usually by looking for the presence of block-artifacts (typically more evident in lower quality videos, as Fig. 2 shows).

Block-artifacts are introduced by any coding algorithm adopting the block-based transform principle. This coding paradigm first splits the set of frames into groups-of-pictures (GOPs) that are encoded independently of one another. Each frame within a GOP is encoded according to a pattern of predefined type:

- *Intra* or I type: the frame is coded independently from all the others; the first frame of each GOP must be intra-coded.
- *Predictive* or P type: the frame is encoded with motion compensation, using the previous I or P frames as references.
- *Bidirectionally predictive* or B type: the frame is encoded with motion compensation, using the previous and the following I or P frames as references.

The customary encoding procedure involves a domain transform (most likely DCT) applied to a block of pixels, typically 8×8 in size. The obtained coefficients are then quantized and fed to an entropy encoder. The lower the bitrate, the coarser is the coefficients representation, resulting in blurry blocks and evident discontinuities at the block boundaries.

Since block-artifacts appear to be quite distinctive for the human vision, we attempted to design a network whose attention is focused on these particular features. Specifically, we wanted our network to analyze the regions nearby the *corners* of the block-grid, as each one of them allows to observe four blocks and boundaries at the same time. To accomplish that, the network architecture was designed to align with the block-grid and to extract a descriptor for each corner and its associated neighborhood. To better understand this, we need to introduce the concept of *receptive field*.

TABLE 1. Network Architecture

Layer	Kernels	w	s	z	Activation
Conv-1	64	4×4	1	0	BN + ReLU
Conv-2	64	3×3	2	0	BN + ReLU
Conv-3	64	4×4	1	0	BN + ReLU
Conv-4	64	3×3	2	0	BN + ReLU
Conv-5	64	3×3	2	1	BN + ReLU
FC-1	64				
FC-2	K				Softmax

Network's layers hyperparameters: w = kernel size; s = stride; z = padding.

The receptive field denotes the region of the input that a particular network neuron is looking at. It is described by its center position and its size. Pixels contribution to the calculation of the output feature grows exponentially towards the center of the receptive field. Given the input size, the layers of a CNN can be designed in order to produce features with the desired receptive field.

The CNN architecture we propose consists of: five convolutional layers, each one followed by a batch normalization (BN) and a rectified linear unit (ReLU) activation; two fully-connected layers; a softmax activation layer. Table 1 reports the complete list of layers, specifying for each convolutional one the number of kernels, the kernel size w , the stride s and the padding size z .

The input to the CNN is a luminance patch of 64×64 pixels. Chrominance components are neglected since they do not add relevant information to block-artifacts and are often subsampled. Assuming a block-grid of 8×8 transform blocks, each patch contains exactly $7 \times 7 = 49$ corners.

Fig. 3 provides a visualization of the spatial layout of the filters for each one of the five convolutional layers. In addition, the values of the following geometrical parameters are provided:

- the size m of the output feature map, based on the size of the input feature map and the layer's properties,

$$m_{out} = \left\lfloor \frac{m_{in} + 2z - w}{s} \right\rfloor + 1; \quad (5)$$

- the jump factor j in the output feature map, representing the spatial displacement in the receptive field between consecutive layers,

$$j_{out} = j_{in} \cdot s; \quad (6)$$

- the receptive field size r of the output feature map,

$$r_{out} = r_{in} + (w - 1) \cdot j_{in}; \quad (7)$$

- the center position c of the receptive field of the first output feature,

$$c_{out} = c_{in} + \left(\frac{w - 1}{2} - z \right) \cdot j_{in}. \quad (8)$$

All parameters are computed with respect to those of the previous layer and, since we are dealing with square filters and symmetrical stride and padding, the dimensions are equal in

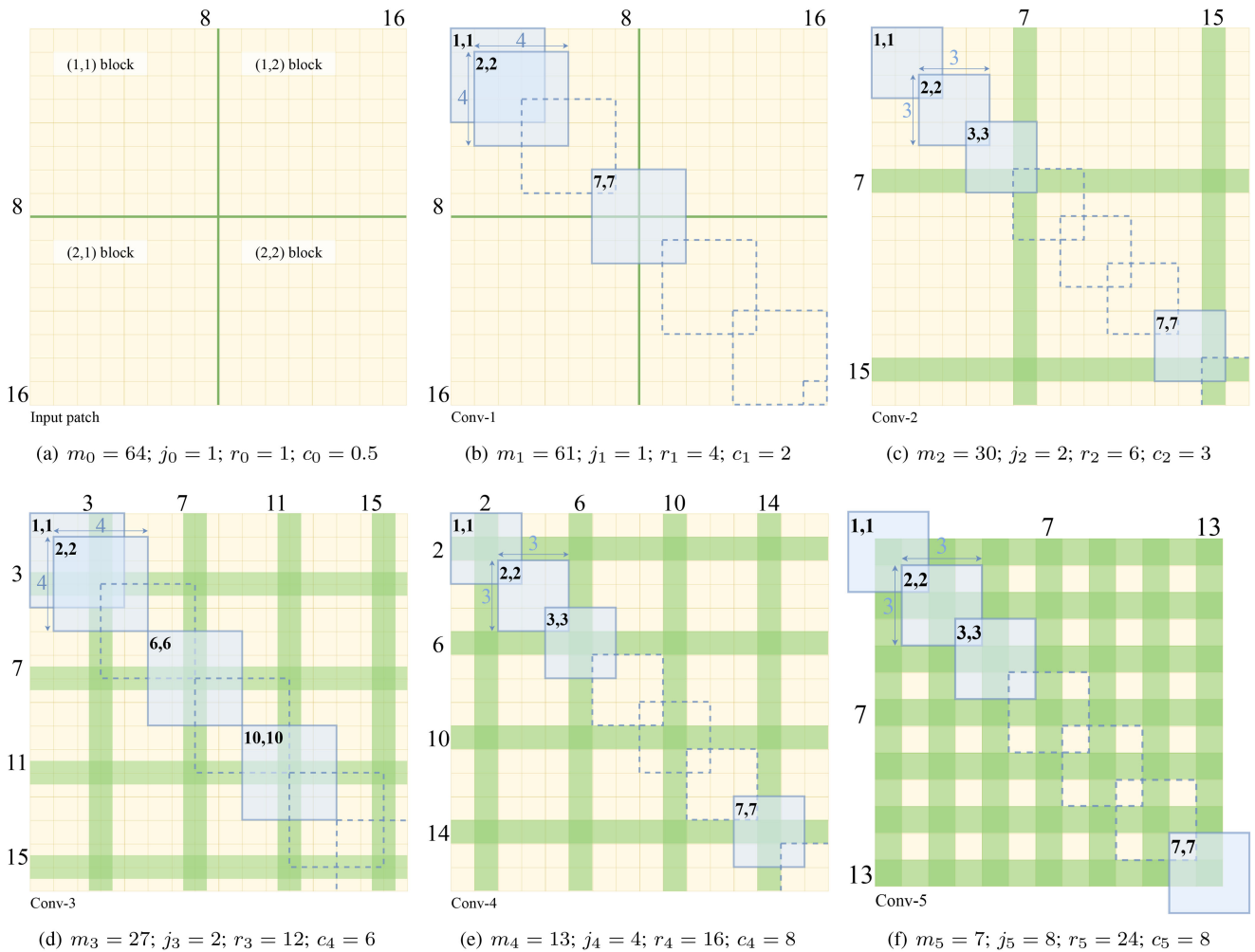


FIGURE 3. Architecture of the convolutional layers, with feature map size (m_i), jump factor (j_i), receptive field size (r_i) and center position (c_i). Green areas denote feature activations related to block boundaries. The output of Conv-5 is a 7-by-7-by-64 tensor, consisting of one 64-length feature vector for each corner of the block grid.

both directions. In the input layer, we have $m_0 \times m_0 = 64 \times 64$ features (the input size), jump factor and receptive field size equal to one pixel ($j_0 = r_0 = 1$) and the center position is the center of the first pixel ($c_0 = 0.5$). Green areas in Fig. 3 represent those parts of the feature map carrying information related to block boundaries. Yellow areas are associated with pixels within the blocks.

With this particular design, the network progressively condenses the block-grid, without blending together the descriptors associated to different corner points. The output of the last convolutional layer is a 7-by-7-by-64 tensor, forming a map of 64-elements descriptors, one for each corner of the input patch. This tensor is fed to a fully connected network that returns the final K -length patch descriptor, where K depends on the chosen number of classification outputs.

The feature descriptors calculated by this CNN can be exploited in a variety of forensic applications. In the following section, we discuss the design of our forgery localization framework, which checks the self-consistency of these coding features to detect temporal and spatial splicing operations.

V. FORGERY LOCALIZATION

FOCAL employs the CNN described in Section IV to extract from an input frame-patch a descriptor associated to its coding standard and quality. The idea is that patches or frames coming from different video sequences will exhibit different coding traces. Detecting descriptor inconsistencies may therefore lead to localizing forgeries.

The proposed CNN was trained to solved a 4-class *codec classification* task, within the following closed set of coding standards,

$$\{\text{MPEG-2, MPEG-4, H.264, H.265}\},$$

and a 4-class *quality classification* task, where the encoding quality is determined by the following values of the quantization step,

$$\Delta = \{5, 10, 20, 40\}.$$

We will refer to the four quality levels throughout the paper as *high*, *medium-high*, *medium-low* and *low*, respectively.

The two trained models were kept separate and used as independent feature extractors. Details on the training phase are provided in Sections VI-A, VI-B. Note also that this framework is scalable to an arbitrary number of trained models, given that they output a vector-shaped feature descriptor.

In the following paragraphs, we discuss the feature extraction phase and the algorithms designed to detect inconsistencies in the descriptors, with the purpose of solving two typical video forensics scenarios: temporal and spatial splicing localization.

A. FEATURE EXTRACTION

Let \mathbf{X} be a video sequence of N frames, as defined in Section III. Each frame \mathbf{X}_n is split into 64×64 patches \mathbf{X}_n^p , $p = 1, \dots, P$, where the number of patches P depends on the video resolution and on the stride used for patch extraction. Note that, in the case of overlapping patches, the stride must be a multiple of the dimension of the coding blocks (8 pixels), in order to maintain the alignment described in Section IV. The extracted patches are then converted to YCbCr color space and their luma components are fed into the two trained CNNs.

For each patch \mathbf{X}_n^p , the output of each network is a four-element feature vector,

$$\mathbf{f}_C^p(n) = [f_{H264}^p(n), f_{H265}^p(n), f_{MPEG2}^p(n), f_{MPEG4}^p(n)],$$

$$\mathbf{f}_Q^p(n) = [f_{low}^p(n), f_{m-low}^p(n), f_{m-high}^p(n), f_{high}^p(n)],$$

where each element $f_{\{\cdot\}}^p(n)$ represents the likelihood of the p -th patch from the n -th frame being encoded with one of the four considered codecs/qualities. Due to the final softmax activation, feature vectors are non-negative and sum to one. As well as considering these vectors as probability distributions over codec/quality classes, one can interpret them as general descriptors capturing local coding traces. As a matter of fact, for forgery detection we are not required to exactly detect the adopted codec and the related coding parameters, but rather observe some sort of feature inconsistency between and within frames.

Given our patch-level descriptors, obtained from heterogeneous feature extractors, we can design different algorithms that leverage such information to detect anomalies, which in turn raise an alarm on the possible presence of forgeries.

B. TEMPORAL SPLICING LOCALIZATION

The proposed temporal splicing localization algorithm relies on the presented features to calculate a descriptor for each frame of the video and localizes inconsistencies between adjacent descriptors. Without loss of generality, we considered temporally-spliced videos composed by only two shots, since this can be easily extended by iterating the same procedure. Additionally, we considered the case of spliced videos obtained with sequences encoded with different codecs and/or different quality parameters. This simulates the case of compilations of shots coming from different devices, broadcasting

sources and social media, as well as shots compressed multiple times or re-encoded as a whole after being spliced.

Given the patch-level features obtained with the procedure described in Section V-A, the desired frame-level feature vectors, $\mathbf{f}_C(n)$, $\mathbf{f}_Q(n)$, are obtained through a standard average,

$$\mathbf{f}_C(n) = \frac{1}{P} \sum_{p=1}^P \mathbf{f}_C^p(n), \quad (9)$$

$$\mathbf{f}_Q(n) = \frac{1}{P} \sum_{p=1}^P \mathbf{f}_Q^p(n), \quad (10)$$

where all operations are performed element-wise.

Finally, the two vectors are concatenated into a general eight-element frame descriptor,

$$\mathbf{f}(n) = [\mathbf{f}_C(n), \mathbf{f}_Q(n)]. \quad (11)$$

To automatically detect inconsistencies over time, we compute the squared Euclidean distance between adjacent feature vectors,

$$\Delta \mathbf{f}(n) = \|\mathbf{f}(n) - \mathbf{f}(n+1)\|^2, \quad (12)$$

and we feed $\Delta \mathbf{f}(n)$ to a threshold-detector.

Fig. 4 reports two examples of temporal splicing localization, applied to 200-frame videos with a splicing point at frame $n = 100$. Fig. 4(a) shows a 100-frame MPEG-2 medium-low-quality video, spliced with a 100-frame MPEG-4 low-quality video. We can observe an evident feature inconsistency at the splicing point, which is correctly detected in the Euclidean distance domain. Note that the $\Delta \mathbf{f}$ axis is displayed in logarithmic scale and the splicing peak is actually two orders of magnitude higher than the background. Fig. 4(b) highlights the sensitivity of the algorithm to intra-coded frames. In the second half of the compilation, the system detects a strong inconsistency once every 30 frames (the GOP size), producing a series of false positives (even though the actual splicing point still yields a significantly higher peak). However, the pattern is regular by its very nature, and thus easy to neglect automatically. Interestingly, note how such inconsistencies are detected only by quality features (lower four), while codec ones remain, correctly, idle: the codec itself is not changing in presence of an I frame, but I frames are coded independently from the others, so coding parameters are indeed different.

C. SPATIAL SPLICING LOCALIZATION

Differently from temporal forgeries, the problem of identifying and localizing a spatial splicing has to be solved within a single frame. The proposed spatial splicing localization algorithm relies on the presented features to calculate local descriptors within the frame, and then look for possible coding inconsistencies. The presence of altered regions comes in the form of activation maps, containing the likelihood of being altered for each patch in the frame. Note how this scenario is significantly more challenging than temporal forgeries: having not the chance of averaging feature vectors, as in the case of frame-level descriptors, translates into reasonably

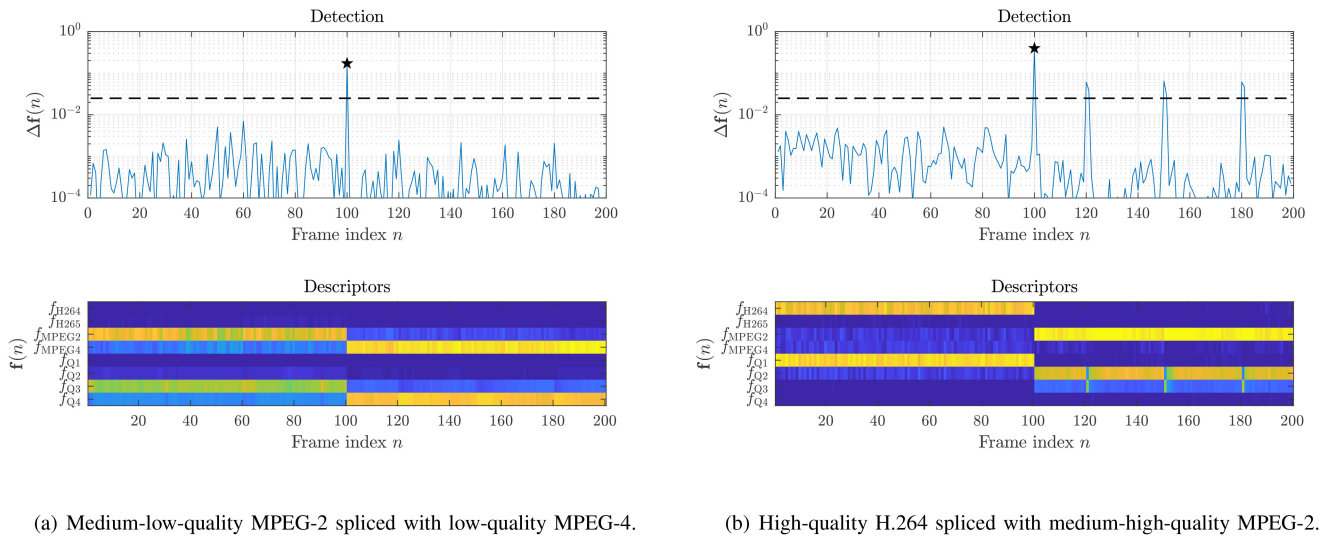


FIGURE 4. Two examples of temporal splicing localization. Feature descriptors (below) are analyzed by means of the Euclidean distance (above) between adjacent vectors. Fig. 4(b) also shows the presence of false positives due to intra-coded frames, with periodicity given by the GOP size.

less accurate features, and thus requires a more sophisticated processing.

1) PATCH-LEVEL FEATURES

Given a frame \mathbf{X}_n , patch-level descriptors $\mathbf{f}_C^p(n)$ and $\mathbf{f}_Q^p(n)$ are extracted as described in Section V-A. We recommend an 8-pixel stride to have a dense description of the frame, while remaining aligned with the quantization grid. The two vectors are concatenated into a general eight-element patch descriptor,

$$\mathbf{f}^p = [\mathbf{f}_C^p(n), \mathbf{f}_Q^p(n)]. \quad (13)$$

Let P_U and P_V be the number of patches extracted along dimensions U and V , respectively. The patch-level descriptors are then arranged in a $P_U \times P_V \times 8$ frame-level *feature tensor*,

$$\mathbf{F}(n) = \begin{bmatrix} \mathbf{f}^{1,1}(n) & \dots & \mathbf{f}^{1,P_V}(n) \\ \vdots & \ddots & \vdots \\ \mathbf{f}^{P_U,1}(n) & \dots & \mathbf{f}^{P_U,P_V}(n) \end{bmatrix}. \quad (14)$$

We call a *feature map*, $\mathbf{F}_k(n)$, $k = 1, \dots, 8$, each $P_U \times P_V \times 1$ matrix in the tensor.

Fig. 5(a) shows an example of feature tensor, with the eight feature maps plotted separately. In this example, the splicing is located at the bottom-left corner of the frame.

The second step consists in transforming the feature tensor into an *activation tensor*, i.e., a set of eight activation maps. Each map $\mathbf{F}_k(n)$ of the feature tensor is fed separately into a suitable activation function $h(\cdot)$ to produce an activation map $\mathbf{H}_k(n)$. With the purpose of highlighting regions that differ from the general trend, we chose as activation function the pixel-wise squared distance from the average value,

$$\mathbf{H}_k(n) = h(\mathbf{F}_k(n)) = \|\mathbf{F}_k(n) - E(\mathbf{F}_k(n))\|^2, \quad (15)$$

where $E(\cdot)$ denotes the expectation operator.

Fig. 5(b) shows the activation tensor obtained from the feature tensor in Fig. 5(a). Note the significance of the activation function for feature map $k = 4$, in particular.

2) FEATURE FUSION

The last and most delicate step consists in fusing the activation tensor into the final activation map. The main issue is that not all feature maps are equally informative, in general. Since each map activates in presence of a specific coding trace, it follows that maps carrying the highest information content will be those related to the coding parameters closest to what is actually present in the analyzed frame. In 5(b) for instance, we observe that: (i) feature maps 1, 2, 4, 5 and 8 correctly agree on an activation at the bottom-left corner; (ii) map 3 does not activate at all; (iii) maps 6 and 7 show a noisy and wide activation, probably due to a background-foreground coding difference in the original video.

In order to automatically select the most informative features, we devised a twofold criterion accounting for the possible presence of idle and/or widely-activated maps.

- *High variance*: a useful map should contain diversity in its values, if an activation is present. This condition helps filtering out idle maps.
- *Low entropy*: a useful activation should be localized to the tampered region. This condition helps filtering out noisy or widely-activated maps.

With these two conditions in mind, we defined a metric called *variance-to-entropy ratio (VER)*,

$$\Lambda(x) = \frac{\text{var}(x)}{H(x)}, \quad (16)$$

which merges the high-variance (var) and low-entropy (H) conditions into a single scoring value.

Fig. 5(c) reports the VERs obtained for the activation maps in the current example. Note that the highest VER values are

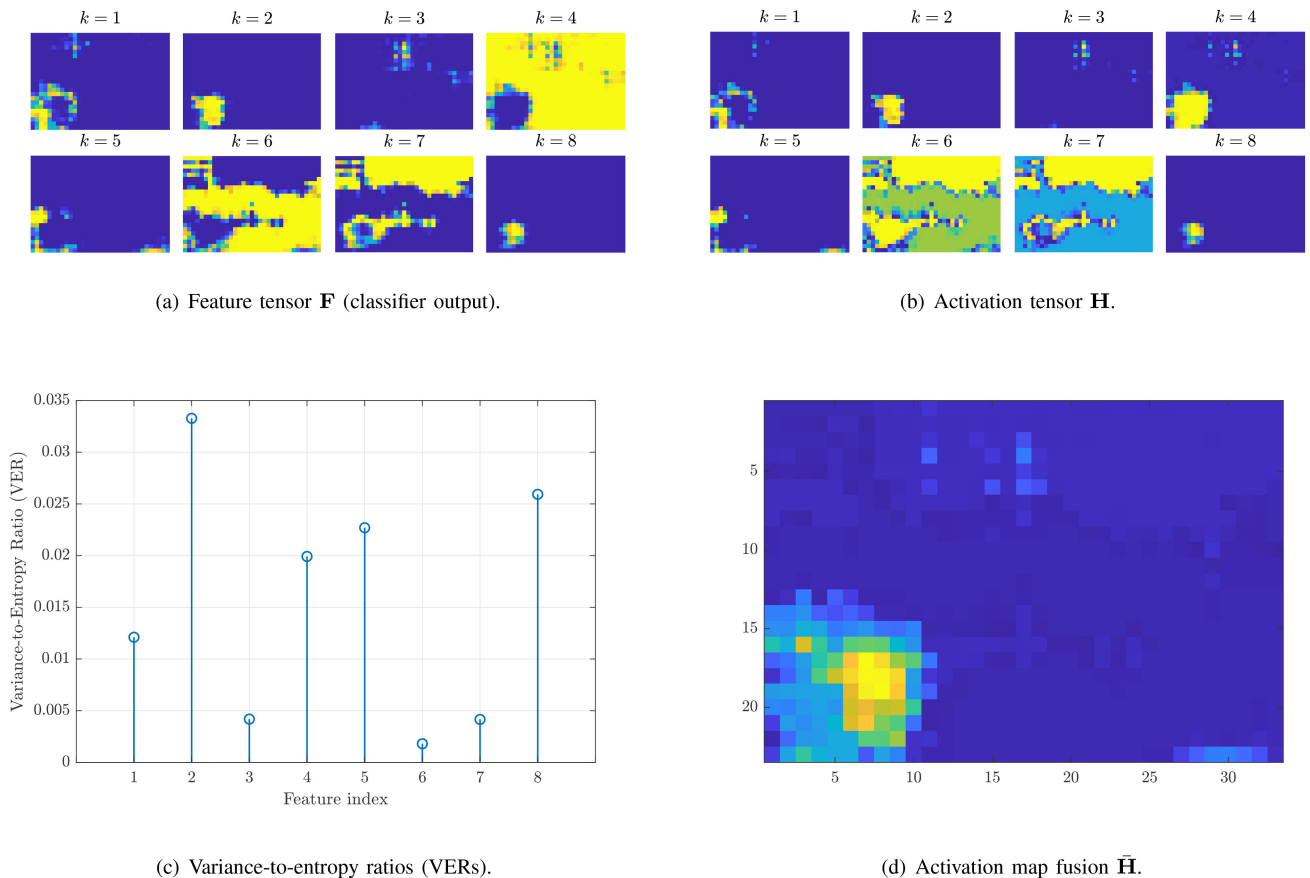


FIGURE 5. Example of spatial splicing localization. The output maps of the CNN classifiers in 5(a) are processed by the activation function in (15), obtaining the activation maps in 5(b). The latter are then averaged using the VERs defined in (16) as weights, providing the final activation map in 5(d).

related to the maps resulting the most informative according to the criteria outlined above, i.e., $k = 2, 4, 5, 8$.

The final fused activation map $\bar{\mathbf{H}}(n)$ is then obtained as

$$\bar{\mathbf{H}}(n) = \frac{\sum_{k=1}^K \Lambda(\mathbf{H}_k(n)) \cdot \mathbf{H}_k(n)}{\sum_{k=1}^K \Lambda(\mathbf{H}_k(n))}, \quad (17)$$

which denotes an element-wise weighted average of the eight activation maps, with weights equal to the VERs.

Fig. 5(d) shows the fusion result for the activation maps of the current example. As desired, all noisy and flat activation maps are discarded, while meaningful ones are retained and merged together into a human-readable output.

VI. EXPERIMENTS AND RESULTS

A distinctive trait of FOCAL framework consists in training the same CNN described in Section IV to solve different classification tasks. As described previously in Section V, in this work we trained two independent coding-related models:

- a 4-class codec classifier;
- a 4-class quality classifier.

This section describes the training of the proposed CNN and the testing campaign carried out to evaluate the system in realistic scenarios, along with the obtained experimental results. Subsection VI-A and VI-B report the training details for the two employed models; VI-C describes the generation

of the testing dataset; VI-D, VI-E and VI-F outline the performed experiments in controlled scenarios, a state-of-the-art comparison and the tests with uncontrolled videos.

A. CODEC-CNN TRAINING

To develop the codec-detector network, we considered four coding standards, namely H.264, H.265, MPEG-2 and MPEG-4, and resorted to a classic training-validation-test approach.

A training dataset of 300 high-resolution videos was built, starting from five uncompressed video sequences [47]: *duckstakeoff* (720p), *stockholm* (720p), *ice* (4CIF), *harbour* (4CIF), *parkrun* (720p). Each video was encoded with FFmpeg software to obtain 60 different versions, combining the four codecs with different coding configurations: fixed quality parameter q , ranging from 1 to 10; constant bitrate (CBR) set to 2 Mb/s, 4 Mb/s and 6 Mb/s; variable bitrate (VBR) set to 2 Mb/s, 4 Mb/s and 6 Mb/s; GOP of 30 frames.

For validation, we built a similar dataset of 300 high-resolution videos following the same procedure of the training phase, but starting from a different set of original video sequences: *parkjoy* (720p), *shields* (720p), *soccer* (4CIF).

For testing the codec classification network on an unrelated dataset, we built a collection of 1672 low-resolution videos, starting from 19 sequences at CIF resolution: *akiyo*,

crew, mother, soccer, bridgeclose, flower, news, table, city, foreman, paris, tempete, coastguard, hall, salesman, water-fall, container, mobile, signirene. Each video was encoded with FFmpeg mixing again codecs and qualities: fixed quality parameter q , ranging from 1 to 31 with step 2; CBR set to 500 Kb/s, 1 Mb/s and 2 Mb/s; VBR set to 500 Kb/s, 1 Mb/s and 2 Mb/s; GOP of 10 frames.

The network was trained using a categorical cross-entropy loss function and Adam optimizer [48], with standard hyperparameters and learning rate. We selected the model by minimizing the validation loss over 50 epochs.

B. QUALITY-CNN TRAINING

For the quality-detector network, the model was trained following again a training-validation-test pipeline considering four quality levels identified by the quantization step Δ . Such value is not directly accessible in the majority of codecs, which present differences in the implementation of the quantization procedure, and it is usually controlled by a higher-level quality parameter q . As a matter of fact, the relation between Δ and q is exponential in H.264 (18), and piece-wise linear in MPEG-2 and MPEG-4 (19):

$$\Delta_{\text{H264}} = \frac{5}{8} \cdot 2^{q/6}, \quad (18)$$

$$\Delta_{\text{MPEG}} = \begin{cases} 8, & 1 \leq q \leq 4 \\ 2q, & 5 \leq q \leq 8 \\ q + 8, & 9 \leq q \leq 24 \\ 2q - 16, & 25 \leq q \leq 31 \end{cases} \quad (19)$$

To generate the training, validation and test sets, we considered three different codecs, namely MPEG-2, MPEG-4 and H.264, and tuned the respective quality parameters according to (18) and (19) in order to have the same quantization step.

Seven raw videos in YUV format and 4CIF quality were used: *crew, crowdrun, duckstakeoff, harbour, ice, parkjoy, soccer.* Each video was encoded with FFmpeg, using three codecs (MPEG-2, MPEG-4, H.264), four quantization steps ($\Delta = \{5, 10, 20, 40\}$), the same VBR parameters of the first experiment and GOP of 30 frames, yielding a total of 84 video sequences. From each video, we extracted 30 frames and 99 non-overlapping 64×64 patches per frame, obtaining a total of 249 480 patches. We observed a clear performance improvement – both in training and testing – using only high-variance patches, since “flat” ones tend to look alike in every codec-quality configuration. We set a variance threshold of 10^3 , ending up retaining 122 473 patches (about 50% of the total). The set of patches was partitioned as follows: 70% for training, 20% for validation and 10% for testing.

The network was trained using a categorical cross-entropy loss function and SGDM optimizer, with an initial learning rate of $5 \cdot 10^{-3}$, a 0.5 drop factor every 5 epochs and batch size of 256 patches. We selected the model by minimizing the validation loss over 50 epochs.

C. TEST DATASET

The generation process of the final testing dataset was split into two steps. First, we created a set \mathcal{D} of encoded videos, using different codecs and qualities. Then, we used the videos in \mathcal{D} to produce two datasets: one set $\mathcal{D}_{\text{temp}}$ of temporally-spliced videos and one set $\mathcal{D}_{\text{spat}}$ of spatially-spliced videos.

Dataset \mathcal{D} was generated from five uncompressed videos that were not included in the previous sets used for training, validation and testing: *four people, in to tree, johnny, kristen and sara, old town cross.* Each video is 210 frames long and 720p resolution. Using FFmpeg, each sequence was encoded with MPEG-2, MPEG-4 and H.264 codecs, with GOP set to 30 frames and four fixed values for the quality parameter, $q = \{3, 8, 13, 18\}$, producing 12 different versions of each video. Overall, the dataset consisted of 60 encoded videos, or 12600 frames. The choice of using q for the final system evaluation is motivated by two reasons: first, it allows to test the algorithm in a scenario closer to a real-world case, since in everyday applications the encoding quality is tuned by means of the quality parameter and not the quantization step; secondly, it produces a testing set that is even more uncorrelated with that used in the CNN training and testing phases.

Denoting with $v_i, i = 1, \dots, 5$, a video from the five originals, dataset \mathcal{D} consists of

$$\mathcal{D} = \bigcup_{i=1}^5 \mathcal{D}_{v_i},$$

where each \mathcal{D}_{v_i} is the subset containing the different versions of v_i , hence $|\mathcal{D}_{v_i}| = 12$ and $|\mathcal{D}| = 60$.

Dataset $\mathcal{D}_{\text{temp}}$ for temporal splicing localization was obtained by splicing the first 100 frames of each video in \mathcal{D}_{v_i} with the second 100 frames of any other video in \mathcal{D}_{v_i} , for each i . Given that the number of possible pairs in a set of 12 elements is $\binom{12}{2} = 66$, dataset $\mathcal{D}_{\text{temp}}$ consists of $5 \times 66 = 330$ temporally-spliced videos, corresponding to $330 \times 200 = 66000$ frames.

Dataset $\mathcal{D}_{\text{spat}}$ for spatial splicing localization was obtained by substituting a CIF window (288×352) of each video in \mathcal{D}_{v_i} with the same window of any other video in \mathcal{D}_{v_i} , for each i . The window was placed at the center of the frame, with the top-left corner aligned with the patch extraction grid, and kept fixed throughout the length of the video. Similarly to $\mathcal{D}_{\text{temp}}$, dataset $\mathcal{D}_{\text{spat}}$ consists of 330 spatially-spliced videos, corresponding to 66000 frames, or 1 452000 patches.

To simulate a more realistic scenario, videos in $\mathcal{D}_{\text{temp}}$ and $\mathcal{D}_{\text{spat}}$ were re-encoded with high-quality H.264 after the forgery. Note that there are no scene changes or content inconsistencies in the forged sequences: temporally-spliced videos have the first 100 frames encoded differently from the subsequent ones; spatially-spliced videos have a CIF window in the middle of the frame encoded differently from the rest. This is necessary to assess the capability of our algorithm to properly localize changes in coding rather than content.

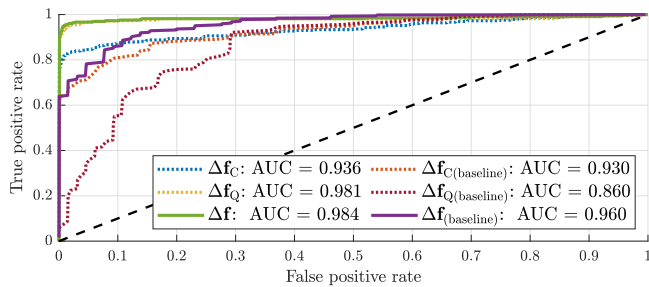


FIGURE 6. ROC curves for frame-wise temporal splicing localization. Comparison of codec-related, quality-related and combined features, for the proposed method and the baseline [12].

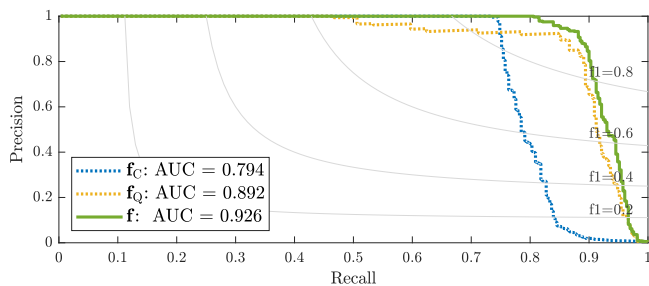


FIGURE 7. PR curves for frame-wise temporal splicing localization. Comparison of codec-related, quality-related and combined features.

D. EXPERIMENTS IN CONTROLLED ENVIRONMENT

For each forensic scenario, namely temporal and spatial splicing, we run three separate tests: one using codec-related features alone; one using quality-related features alone; one using both features concatenated.

Dataset \mathcal{D}_{temp} was analyzed with the algorithm outlined in Section V-B. The detection of splicing points was evaluated frame-wise: a true-positive consisted of a splicing point correctly identified in the transition between two adjacent frames. For this experiment, false positives related to the GOPs were discarded, as discussed in Section V-B.

Fig. 6 shows the ROC curves obtained for temporal splicing localization with the three different set of features. For each case, a direct comparison with the baseline work in [12] is provided. All three descriptors outperform the baseline ones, with an AUC peaking at 0.984 for the concatenated features. For better appreciating the small differences between using the proposed descriptors separately or concatenated, we report the same results displayed as PR curves in Fig. 7. We can observe how quality-related features f_Q provide better results than codec-related ones f_C in this scenario. However, the concatenated features f still lead to an improvement with respect to f_Q alone. Tests run on the concatenated descriptor f show a 100% precision up to a recall of roughly 80%, denoting a good robustness of the algorithm to false positives, and an AUC of 0.926. The optimal operating point of the green curve corresponds to an F1-measure of 0.902.

Dataset \mathcal{D}_{spat} was analyzed with the algorithm outlined in Section V-C. The detection was evaluated patch-wise; a true-positive consisted in a 64×64 patch correctly classified

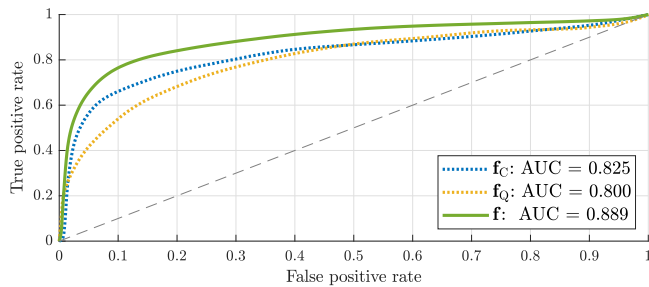


FIGURE 8. ROC curves for patch-wise spatial splicing localization, using descriptors from a *single* frame. Comparison of codec-related, quality-related and combined features.

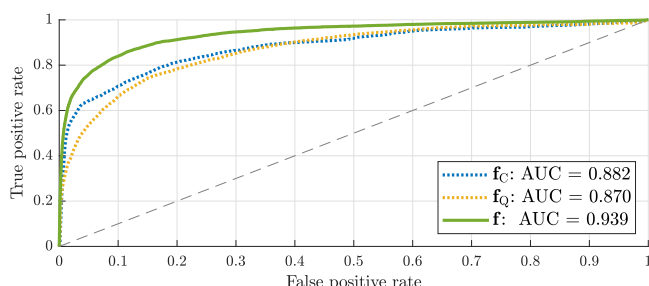


FIGURE 9. ROC curves for patch-wise spatial splicing localization, using descriptors averaged over *multiple* frames. Comparison of codec-related, quality-related and combined features.

as forged. We run two separate tests: one with patch-level descriptors obtained frame-by-frame; one with descriptors obtained by averaging throughout the video frames.

Fig. 8 shows the ROC curves obtained for spatial splicing localization on a single frame, with the three different set of features. Again, the benefits of using concatenated descriptors are clearly visible. Note also how in this case f_C performs better than f_Q , when taken individually. We can assume this is due to the fact that quality is trickier to assess at patch-level than at frame-level; low-variance patches, for instance, typically look very similar at different encoding qualities. However, f_Q features still provide useful information in combination with f_C , as shown by the results improvement associated with f .

Since in \mathcal{D}_{spat} the forged region is fixed in time, we were able to average descriptors throughout all the 200 frames of the video sequences. Fig. 9 shows the ROC curves obtained for spatial splicing localization averaged over multiple frames with the three different descriptors. As expected, we observe a clear improvement in all of them with respect to the single-frame case. In a real-case scenario however, the assumption of a non-moving forged region does not hold in general. Nevertheless, it is still possible to resort to this performance-enhancing strategy by averaging descriptors over short-time windows, assuming that the motion of the altered region is slow enough compared to the frame-rate.

All three experiments show that concatenating feature descriptors associated to different classification tasks lead to a clear performance improvement. As long as we are able to identify additional classes of forensics traces, and to obtain a

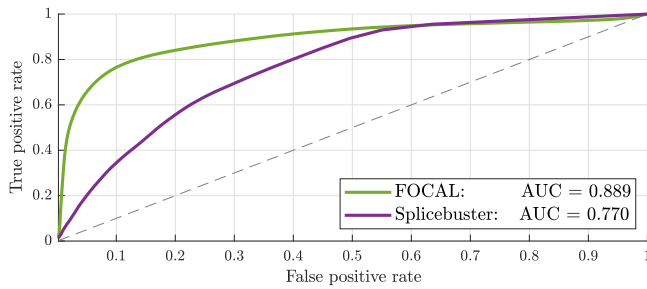


FIGURE 10. ROC curves comparison of FOCAL and Splicebuster [49] for spatial splicing detection, in single-frame setup (Fig. 8).

specific feature descriptor for them, we can assume that the proposed framework would keep taking advantage from the combination of a greater number of feature extractors.

E. COMPARISON WITH THE STATE OF THE ART

In this experiment, the spatial splicing detection capabilities of FOCAL were tested against an image forgery localization state-of-the-art technique, i.e., *Splicebuster* [49]. We adopted the single-frame approach (the same of Fig. 8) and compared the detection heatmaps produced by the two approaches frame by frame.

As for FOCAL, we employed the combined codec- and quality-related descriptors, with the same setup of the experiment described in Section VI-D. As for Splicebuster, the official available implementation for still image analysis was used. Both methods were tested on dataset $\mathcal{D}_{\text{spat}}$ of spatially-spliced videos (Section VI-C).

Fig. 10 shows a comparison of the two ROC curves obtained for FOCAL and Splicebuster on dataset $\mathcal{D}_{\text{spat}}$. Detection accuracy is computed pixel-wise with respect to the ground-truth, independently for each frame, and it is possible to observe an improvement of about 0.12 in terms of AUC.

Table 2 reports some qualitative examples of detection heatmaps for the two methods under analysis. The displayed heatmaps are related to five videos from dataset $\mathcal{D}_{\text{spat}}$ (described in Section VI-C), with one realization for each original sequence. The first column reports the RGB frames, with a dashed box highlighting the altered region (remember that only coding parameters are different, i.e., there is no content change); the second and third columns report FOCAL and Splicebuster’s heatmaps, respectively.

The first three videos (obtained from *old_town_cross*, *johnny* and *into_tree*) present a splicing whose coding parameters are highly different from the rest of the frame (both codec and quality change); in fact, we can observe that, at least for the first two cases, both methods correctly activates in correspondence of the altered region at the center of the frame. The third one however is not detected by Splicebuster: by looking at the activation at the top of the frame, which quite resemble the trees-sky transition, we may infer that the detector is spotting a background-foreground coding difference or a change in the overall pixels statistics.

TABLE 2. Comparison of Detection Maps

RGB frame	FOCAL	Splicebuster [49]

Comparison of detection heatmaps obtained with FOCAL and Splicebuster [49] from five videos of dataset $\mathcal{D}_{\text{spat}}$ (Section VI-C), one realization for each original sequence. The altered region (dashed box) only differs in terms of coding parameters, i.e., no content change.



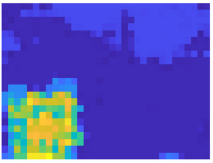
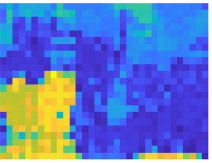
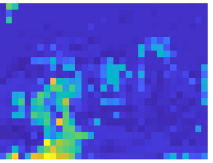
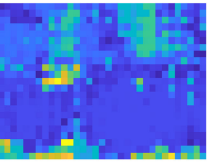


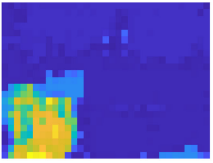
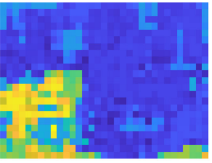
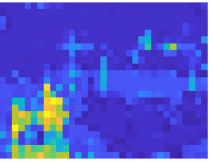
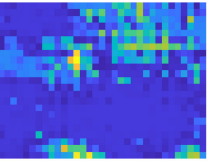


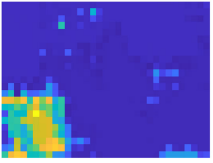
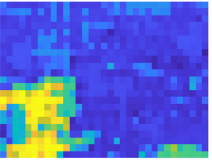
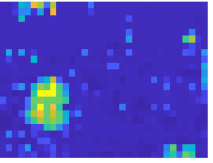
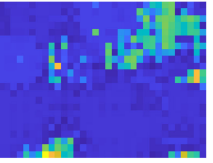
The last two videos (obtained from *kristen_and_sara* and *four_people*) are trickier for both methods, as the altered region is closer to the rest of the frame in terms of coding parameters. One can observe that FOCAL is still able to identify the splicing, at the cost of a slightly noisier heatmap. As for Splicebuster, we notice again the behavior of the third case, in that heatmaps appear to be affected by the content of the scene (see the contours of the two rightmost people in the fifth case).

F. EXPERIMENTS ON UNCONTROLLED VIDEOS

As a last experiment, we run FOCAL on videos from the online dataset of the REVerse engineering of audio-VISual coNtent Data (REWIND) project [24]. These sequences contain photo-realistic forgeries, similar to those encountered in a practical forensic scenario. Original videos were recorded using low-end devices, with a resolution of 320×240 pixels and a framerate of 30 fps. Each forged sequence is available in four encoding configurations: lossless H.264 and lossy H.264 with $q = 10, 20, 30$.

Given the availability of multiple coding qualities for the same forged video, we used this dataset to assess how the robustness of the algorithm is impaired as the encoding quality decreases. Table 3 shows the activation maps calculated by FOCAL on three frames of sequence *01* of the REWIND dataset. Each map was obtained from a single frame, with no temporal averaging. In this example, the forgery consists in a spatial splicing at the bottom-left corner of the frame: the

TABLE 3. Forgery Localization At Different Re-Encoding Qualities

Original frame	Forged frame	Lossless	$q = 10$	$q = 20$	$q = 30$
					
					
					

Decaying localization performance at lower re-encoding qualities of the same forged video. From top to bottom, frames 52, 78 and 144 of sequence 01 of the REWIND dataset [24]. From left to right, the original frame, the forged frame, the localization heatmaps for lossless and lossy H.264 encoding, with quality parameter $q = 10, 20, 30$.

bigger duck is spliced-out by copy-moving an empty portion of the water surface and a second duck is spliced-in. The resulting content still appears photo-realistic and spotting the forgery turns out being a challenging task even for a human observer. As we can see, the localization capabilities of the algorithm remain satisfying up to a quality parameter $q = 10$. Lower encoding qualities progressively erase any useful forensic trace, impairing the detection reliability. However, since the content itself of the video become progressively less discernible, it is arguable whether a forger should ever adopt such low encoding qualities with the purpose of creating a convincing fake.

VII. CONCLUSION

In this paper we presented FOCAL, a framework for video forgery localization based on the self-consistency of coding traces. The main contributions that come along with this strategy consist in the design of an ad-hoc CNN architecture for learning codec-related features and a fusion technique that merges different descriptors into a general likelihood map, using the newly designed variance-to-entropy ratio metric. The resulting framework is scalable and generalizable at will. A first implementation, featuring two independently-trained coding-related descriptors, was here proposed and tested over two different forgeries situations. Experimental results showed a clear performance improvement with respect to the previous work on temporal splicing localization, and promising results in the newly tackled scenario of spatial splicing localization.

Being able to capture local coding traces over small patches of a video-frame, paves the way to several possibilities in the context of forgery detection. Future research will be devoted to assessing the scalability of the proposed framework through

the addition of further models and by upgrading the existing ones with a higher number of coding parameters. New strategies to fuse and leverage feature information could be explored as well, with the purpose of enabling the detection of increasingly complex types of forgeries.

ACKNOWLEDGMENT

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.

REFERENCES

- [1] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Source digital camcorder identification using sensor photo-response nonuniformity," in *Proc. SPIE Electron. Imag.*, 2007, Art. no. 65051G.
- [2] S. Bayram, H. T. Sencar, and N. Memon, "Video copy detection based on source device characteristics: A complementary approach to content-based methods," in *Proc. ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 435–442.
- [3] V. Conotter, J. O'Brien, and H. Farid, "Exposing digital forgeries in ballistic motion," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 1, pp. 283–296, Feb. 2012.
- [4] M. Visentini-Scarzanella and P. L. Dragotti, "Video jitter analysis for automatic bootleg detection," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2012, pp. 101–106.
- [5] M. Stamm, W. Lin, and K. Liu, "Temporal forensics and anti-forensics for motion compensated video," *IEEE Trans. Inf. Forensics Secur. (TIFS)*, vol. 7, no. 4, pp. 1315–1329, Aug. 2012.
- [6] P. Bestagini, S. Battaglia, S. Milani, M. Tagliasacchi, and S. Tubaro, "Detection of temporal interpolation in video sequences," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 3033–3037.
- [7] S. Bian, W. Luo, and J. Huang, "Exposing fake bit rate videos and estimating original bit rates," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 12, pp. 2144–2154, Dec. 2014.

- [8] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Codec and GOP identification in double compressed videos," *IEEE Trans. Image Process. (TIP)*, vol. 25, no. 5, pp. 2298–2310, May 2016.
- [9] S. Milani, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Multiple compression detection for video sequences," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2012, pp. 112–117.
- [10] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 101–117.
- [11] M. Mathai, D. Rajan, and S. Emmanuel, "Video forgery detection and localization using normalized cross-correlation of moment features," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, Mar. 2016, pp. 149–152.
- [12] S. Verde, L. Bondi, P. Bestagini, S. Milani, G. Calvagno, and S. Tubaro, "Video codec forensics based on convolutional neural networks," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 530–534.
- [13] S. Milani *et al.*, "An overview on video forensics," *APSIPA Trans. Signal Inf. Process.*, vol. 1, no. e2, pp. 1–18, 2012.
- [14] K. Sitara and B. Mehtre, "Digital video tampering detection: An overview of passive techniques," *Digit. Investigation*, vol. 18, pp. 8–22, 2016.
- [15] J. Yang, T. Huang, and L. Su, "Using similarity analysis to detect frame duplication forgery in videos," *Multimedia Tools Appl.*, vol. 75, no. 4, pp. 1793–1811, Feb. 2016.
- [16] G. Lin, J. Chang, and C. Chuang, "Detecting frame duplication based on spatial and temporal analyses," in *Proc. 6th Int. Conf. Comput. Sci. Educ.*, Aug. 2011, pp. 1396–1399.
- [17] A. De, H. Chadha, and S. Gupta, "Detection of forgery in digital video," in *Proc. 10th World Multi Conf. Systemics Cybern. Informat.*, 2009, pp. 229–233.
- [18] N. Mondaini, R. Caldelli, A. Piva, M. Barni, and V. Cappellini, "Detection of malevolent changes in digital video for forensic applications," in *Proc. SPIE 6505, Secur., Steganogr., Watermarking Multimedia Contents IX*, Feb. 2007, Art. no. 65050T. [Online]. Available: <https://doi.org/10.1117/12.704924>
- [19] M. Kobayashi, T. Okabe, and Y. Sato, "Detecting forgery from static-scene video based on inconsistency in noise level functions," *IEEE Trans. Inf. Forensics Secur.*, vol. 5, no. 4, pp. 883–892, Dec. 2010.
- [20] W. Wang, X. Jiang, S. Wang, M. Wan, and T. Sun, "Identifying video forgery process using optical flow," in *Proc. Digit.-Forensics Watermarking*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 244–257.
- [21] S. Kingra, N. Aggarwal, and R. D. Singh, "Inter-frame forgery detection in h.264 videos using motion and brightness gradients," *Multimedia Tools Appl.*, vol. 76, no. 24, pp. 25 767–25 786, Dec. 2017.
- [22] P. Bestagini, S. Battaglia, S. Milani, M. Tagliasacchi, and S. Tubaro, "Detection of temporal interpolation in video sequences," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 3033–3037.
- [23] Y. Su, J. Zhang, and J. Liu, "Exposing digital video forgery by detecting motion-compensated edge artifact," in *Proc. Int. Conf. Comput. Intell. Softw. Eng.*, Dec. 2009, pp. 1–4.
- [24] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Local tampering detection in video sequences," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process.*, Sep. 2013, pp. 488–493.
- [25] H. Liu, S. Li, and S. Bian, "Detecting frame deletion in H. 264 video," in *Information Security Practice and Experience: 10th International Conference, ISPEC 2014*, Fuzhou, China, May 5–8, 2014, Proceedings, Springer, vol. 8434, 2014.
- [26] Z. Zhang, J. Hou, Q. Ma, and Z. Li, "Efficient video frame insertion and deletion detection based on inconsistency of correlations between local binary pattern coded frames," *Secur. Commun. Netw.*, vol. 8, pp. 311–320, Jan. 2015.
- [27] L. Zheng, T. Sun, and Y.-Q. Shi, "Inter-frame video forgery detection based on block-wise brightness variance descriptor," in *Digital-Forensics and Watermarking: 13th International Workshop, IWDW 2014*, Taipei, Taiwan, Oct. 1–4, 2014. Revised Selected Papers, Springer, vol. 9023, 2015.
- [28] D.-K. Hyun, S.-J. Ryu, H.-Y. Lee, and H.-K. Lee, "Detection of upscale-crop and partial manipulation in surveillance video based on sensor pattern noise," *Sensors*, vol. 13, no. 9, pp. 12 605–12 631, Jul. 2013.
- [29] W. Wang and H. Farid, "Exposing digital forgeries in interlaced and deinterlaced video," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 438–449, Sep. 2007.
- [30] L. Li, X. Wang, W. Zhang, G. Yang, and G. Hu, "Detecting removed object from video with stationary background," in *Digital-Forensics and Watermarking: 11th International Workshop, IWDW 2012*, Shanghai, China, October 31–November 3, 2012, Revised Selected Papers, Springer, vol. 7809, 2013.
- [31] S. Chen, S. Tan, B. Li, and J. Huang, "Automatic detection of object-based forgery in advanced video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2138–2151, Nov. 2016.
- [32] A. V. Subramanyam and S. Emmanuel, "Video forgery detection using hog features and compression properties," in *Proc. IEEE 14th Int. Workshop Multimedia Signal Process.*, Sep. 2012, pp. 89–94.
- [33] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A SIFT-based forensic method for copy-move attack detection and transformation recovery," *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 3, pp. 1099–1110, Sep. 2011.
- [34] R. C. Pandey, S. K. Singh, and K. K. Shukla, "Passive copy-move forgery detection in videos," in *Proc. Int. Conf. Comput. Commun. Technol.*, Sep. 2014, pp. 301–306.
- [35] L. D'Amiano, D. Cozzolino, G. Poggi, and L. Verdoliva, "Video forgery detection and localization based on 3-D patchmatch," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jun. 2015, pp. 1–6.
- [36] A. Bidokhti and S. Ghaemmaghami, "Detection of regional copy/move forgery in mpeg videos using optical flow," in *Proc. Int. Symp. Artif. Intell. Signal Process.*, Mar. 2015, pp. 13–17.
- [37] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double mpeg compression," in *Proc. 8th Workshop Multimedia Secur.*, 2006, pp. 37–47.
- [38] X. Junyu, Y. Su, and Q. Liu, "Detection of double mpeg-2 compression based on distributions of DCT coefficients," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 27, no. 2, 2013, Art. no. 1354001.
- [39] D. Vazquez-Padin, M. Fontani, T. Bianchi, P. Comesana, A. Piva, and M. Barni, "Detection of video double encoding with gop size estimation," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, Dec. 2012, pp. 151–156.
- [40] P. Bestagini, A. Allam, S. Milani, M. Tagliasacchi, and S. Tubaro, "Video codec identification," in *IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 2257–2260.
- [41] Z. Huang, F. Huang, and J. Huang, "Detection of double compression with the same bit rate in mpeg-2 videos," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2014, pp. 306–309.
- [42] X. Jiang, W. Wang, T. Sun, Y. Q. Shi, and S. Wang, "Detection of double compression in mpeg-4 videos based on markov statistics," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 447–450, May 2013.
- [43] H. Ravi, A. V. Subramanyam, G. Gupta, and B. A. Kumar, "Compression noise based video forgery detection," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5352–5356.
- [44] D. D'Avino, D. Cozzolino, G. Poggi, and L. Verdoliva, "Autoencoder with recurrent neural networks for video forgery detection," *Electron. Imag.*, vol. 2017, no. 7, pp. 92–99, 2017.
- [45] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2018, pp. 1–7.
- [46] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2018, pp. 1–6.
- [47] C. Montgomery and H. Lars, "Xiph.org Video Test Media (Derf's Collection)," 2017. [Online]. Available: <https://media.xiph.org/video/derf>
- [48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [49] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2015, pp. 1–6.