# Multi-Scale Fusion for Improved Localization of Malicious Tampering in Digital Images

Paweł Korus, *Member, IEEE,* Jiwu Huang, *Fellow, IEEE*

***Abstract*—Sliding window-based analysis is a prevailing mechanism for tampering localization in passive image authentication. It uses existing forensic detectors, originally designed for full-frame analysis, to obtain the detection scores for individual image regions. One of the main problems with window-based analysis is its impractically low localization resolution stemming from the need to use relatively large analysis windows. While decreasing the window size can improve the localization resolution, the classification results tend to become unreliable due to insufficient statistics about the relevant forensic features. In this study, we investigate a multi-scale analysis approach which fuses multiple candidate tampering maps, resulting from the analysis with different windows, to obtain a single, more reliable tampering map with better localization resolution. We propose three different techniques for multi-scale fusion, and verify their feasibility against various reference strategies. We consider a popular tampering scenario with mode-based first digit features to distinguish between singly and doubly-compressed regions. Our results clearly indicate that the proposed fusion strategies can successfully combine the benefits of small-scale and large-scale analysis and improve the tampering localization performance.**

***Index Terms*—digital image forensics; tampering localization; result fusion; multi-scale analysis; first-digit-features; energy minimization; Markov random fields**

## I. Introduction

The increasing ease of editing digital photographs has spawned an urgent need for reliable authentication mechanisms, capable of precise localization of potential malicious forgeries. Though proactive image protection schemes can deliver precise identification of the tampered regions and even restore their original appearance with very high-quality [1–3], they can only be exploited in a strictly controlled environment, since they use a carefully designed digital watermark that needs to be available as side information.

The rapidly developing field of digital image forensics aims to deliver passive authentication mechanisms that analyze intrinsic fingerprints introduced on various stages of the image acquisition pipeline [4, 5]. As a result, such techniques can be applied to existing, non-watermarked images. However, since they are built on the foundations of machine learning and statistical signal analysis, reliable detection of forensic features renders precise tampering localization a challenging problem.

Most of existing localization methods are essentially extensions of conventional full-frame detectors operating on a sliding window basis. For the sake of sufficient statistical discriminability, a relatively large window size needs to be used. The typically reported size ranges between $64 \times 64$ px and $128 \times 128$ px, depending on the forensic feature at hand [6, 7]. The analysis window might be moved around the image in an overlapping [7], or non-overlapping [6] manner. In the former case, the final classification score of the current analysis window can either be used just for the central authentication unit of the window, or multiple candidate values available for every authentication unit might be fused together, e.g., with a voting mechanism [7].

One of the main problems with sliding window analysis is that on the one hand, large windows are expected to provide reliable classification accuracy, yet with limited localization resolution. On the other hand, small windows could potentially yield much better localization capability, but are expected to be impractically prone to errors. Hence, neither choice by itself fulfills the requirements for a reliable forensic tool.

Due to limited resolution of window-based analysis, the researchers have also explored alternative approaches to localization. Chen et al. proposed to use non-linear optimization, derived from robust principal component analysis, to find sparse clusters of outliers in the forensic feature space that are expected to constitute the tampered regions [8].

In the forensic analysis of JPEG images, localization can also be obtained by fine-grained analysis of JPEG quantization noise [9], or the histograms of DCT coefficients [10, 11], which exhibit different behavior depending on the compression history. The localization can also be performed by estimating the parameters of a mixture model for the DCT coefficients which is then used to calculate tampering probability for each image block [12]. While such analysis can be performed with resolution of $8 \times 8$ px blocks, its accuracy is still limited, especially for higher compression rates. To some extend, classification errors might be eliminated by incorporating prior knowledge about the tampering maps, e.g., by penalizing the differences between neighboring blocks [12]. Such an approach uses a Markov random field (MRF) to model the prior. Similar formulation has also recently been used in a Bayesian localization approach based on photo response non-uniformity (PRNU) analysis [7].

In this study, we address the problem by fusing the results from multiple analysis windows of various size. We will show that such an approach successfully combines the benefits of



This work was partly supported by NSFC (61332012, 61572329, and 61402295), Guangdong NSF (2014A030313557), and Shenzhen R&D Program (GJHZ20140418191518323).

P. Korus and J. Huang are with the College of Information Engineering, Shenzhen University, Shenzhen, China and Shenzhen Key Laboratory of Media Security. P. Korus is also with the Department of Telecommunications, AGH University of Science and Technology, Kraków Poland, E-mail: pkorus@agh.edu.pl, jwhuang@szu.edu.cn.

Supplementary materials are available online at http://ieeexplore.ieee.org.

small-scale and large-scale analysis and improves the performance of tampering localization.

In certain applications, e.g., wireless sensor networks, information fusion is a well studied problem [13–15]. It has been proven that in certain conditions, the optimal strategy involves summation of candidate measurements, weighted according to the performance of individual sensors [16]. Similar techniques - referred to as *ensemble* classification or boosting [17, 18] - are also widely used in pattern recognition for constructing a strong learner from a set of weak learners and for speeding-up computations for rich feature sets that are gaining popularity both in steganalysis [19] and in forensics.

In digital image forensics, decision fusion is an emerging trend aiming to combine the results from multiple detectors, preferably based on different forensic features. Examples of currently considered approaches range from simple decision-level fusion with majority voting or predefined logical rules [20], to more sophisticated approaches with measurement-level fusion based on the Dempster-Shafer theory of evidence (DSTE) [21] or fuzzy logic [22]. An evaluation of several fusion approaches for popular forensic detectors can be found in a recent study [23]. Extension of sophisticated fusion methods to tampering localization is an open problem. Both DSTE and fuzzy logic have been thoroughly analyzed only in a full-frame tampering detection scenario. However, preliminary work on a DSTE-based localization demonstrates feasibility of such an approach [24].

Multi-scale analysis is a well-established approach in various areas of image processing and computer vision. It has been used for object detection [25], image blending [26], depth of field blending [27], denoising [28, 29], speeding up convergence of optimization algorithms [30], finding similar patches (e.g., in super-resolution [31]), and many more. The most relevant works for the application at hand include object detection, image texture segmentation [32] and depth map estimation [33] which involve decision fusion. Object detection involves identification of only the bounding box of the object. Hence, simple voting mechanisms typically suffice. In the texture segmentation problem [32], the fusion was performed by averaging the candidate likelihood functions, weighted by discriminative capabilities of different scales. For patch-based depth map estimation [33], the authors used a MRF to model the multi-scale fusion problem.

Fusion techniques (not necessarily multi-scale) are also used in saliency detection. The input maps may represent various saliency estimation techniques [34], various types of saliency (static vs. dynamic) [35], or even co-saliency among many images [36]. A recent study compares the performance of many fusion methods [35], most of which are just variations of well known averaging, maximization, or multiplication approaches with custom weights for individual components. The best performance was obtained by the simplest unweighted averaging. Decision fusion in saliency can also involve combining bottom-up and top-down clues [37]. Bottom-up saliency refers to low-level image features, e.g., high-contrast regions that attract attention. Top-down saliency refers to high-level, task-oriented recognition tasks, e.g., face, object or text recognition.

Successful fusion of multi-scale tampering maps constitutes a challenging problem and requires new techniques to properly exploit inter-scale dependencies between the candidate maps. Existing methods commonly assume independence of the scores both in the spatial [7, 33] and in the scale dimension [33]. A brief argument about this simplification for single-scale PRNU-based localization can be found in [7]. Therefore, in multi-scale fusion of tampering maps, the aspect of inter-scale correlations calls for a separate, more detailed analysis.

In this paper we propose three novel fusion methods that exploit the dependencies between successive scales of analysis. Firstly, we consider an energy-minimization approach, which uses a MRF to model the prior knowledge about the tampering maps. Secondly, we consider two dual heuristic strategies, referred to as *bottom-up* and *top-down* fusion, which exploit the expected dependencies between the tampered regions in small and large-scale analysis. We compare the proposed strategies to a few reference methods, including decision-level fusion by majority voting, measurement-level fusion by candidate map averaging, supervised learning with support vector machines (SVM) and K-means clustering.

Our analysis is performed on the example case of a popular tampering scenario involving splicing of JPEG images that produces regions with different compression history. The forensic features of choice are the mode-based first digit features (MBFDF) [38]. While recent studies have already demonstrated that MBFDF can be used successfully for sliding window-based localization [6], we conduct a much more detailed analysis with densely sampled compression levels and emphasis on multi-scale localization.

This paper is organized as follows. In Section II we provide a detailed analysis of the multi-scale localization problem based on an popular forgery scenario involving splicing of JPEG images. The proposed fusion techniques are presented in Section III, and evaluated in Section IV. We discuss the limitations in Section V and conclude in Section VI.

## II. Multi-Scale Analysis in Digital Image Forensics

We consider a popular forgery, in which, as a result of content replacement, some fragments of the JPEG image have different compression history and exhibit either single or double compression artifacts. In this section, we provide a detailed analysis of the multi-scale localization problem based on MBFDFs. We make our discussion as general as possible, to facilitate easier generalization to other forensic features.

### A. Formal Statement of the Fusion Problem

We consider a classical sliding-window approach where a forensic detector analyzes successive windows of the image, and records the obtained classification scores in a real-valued *candidate map*. Without loss of generality, we assume that its values - the *candidate scores* - are in the range $[0, 1]$ that represents the confidence of the decision. A *tampering map* denotes the final binary decisions that indicate the locations of the tampered regions.

Multi-scale fusion involves computation of the final tampering map, given a set of candidate maps obtained from various

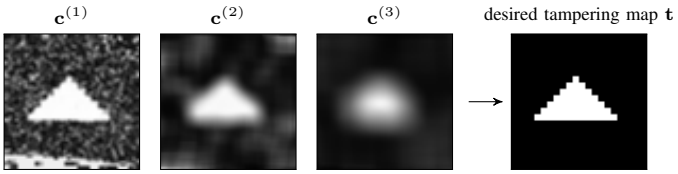



Fig. 1. Idealized fusion of 3 candidate maps $c^{(s)}$, corresponding to a small, medium, and large analysis windows, into a single accurate tampering map.

scales of analysis. Let $I$ denote an input image, divided into $N$ *authentication units*, in our study corresponding to non-overlapping $8 \times 8$ px blocks that are fundamental building blocks of JPEG images. While it is more intuitive to see the tampering map as a 2D matrix with $N_h$ rows and $N_w$ columns, for the sake of notation convenience the matrices will be indexed with a single variable. Hence, a tampering map constitutes a mapping $t : \{1, \ldots, N\} \rightarrow \{0, 1\}$ that can be conveniently represented in vector notation $\mathbf{t} \in \{0, 1\}^N$ with $t_i$ denoting the decision for the $i$-th authentication unit.

The problem is to find the optimal $\mathbf{t}$ given a set of candidate maps $\mathbf{c}^{(s)} \in [0, 1]^N$ for $s \in \{1, \ldots, S\}$ analysis window scales. In our investigation the candidate maps are of the same size. A general illustration of the problem is shown in Fig. 1 on an example of three real candidate maps obtained by sliding-window analysis using 16 px, 64 px, and 128 px windows.

A common problem with many forensic tools is their unreliable behavior in saturated, dark or flat regions of the image [6, 7]. In previous work on MBFDF-based tampering localization this aspect remained unaddressed [6]. In this study, in addition to the candidate tampering maps, we calculate their corresponding reliability maps $\mathbf{p}^{(s)}$. Proceeding in the same sliding window manner, we measure the average magnitude of the AC coefficients. As a result, we can easily detect saturated and flat blocks with zeroed histograms. Fusion methods may use these *reliability scores* as auxiliary input.

### B. Localization using First Digit Features

Our feature space contains 180 features - the MBFDFs of all 9 digits from first 20 AC coefficients. Classification is performed by a SVM classifier with a radial basis function (RBF) kernel and trained to yield probability estimates of the decisions. We refer to the obtained estimates as *classification scores*. Tampering localization capability is obtained by means of a sliding-window analysis. In contrast to the work by Amerini et al. [6] we use overlapping windows and move them one authentication unit at a time. We compute the final score for every authentication unit as the average probability estimate over all overlapping windows. Such an approach yields smooth tampering maps (Fig. 1). In order to reduce the computational complexity, it may be necessary to reduce the window overlap. Detailed analysis of this issue is out of scope of this study, but we discuss it briefly in Section V.

We use square analysis windows of size 16, 32, 48, 64, 80, 96, 112, 128 px, and consider a dense grid of possible JPEG quality levels $Q_1, Q_2 \in \{50, 51, \ldots, 100\}$. We assume the most practical scenario and train a separate classifier for every second compression level $Q_2$, and for every analysis window. Each classifier is trained on 20,000 example windows, chosen randomly from a training set generated from 1,338 uncompressed images from the UCID database [39]. The SVM parameters $C$ (misclassification penalty) and $\gamma$ (kernel scale) are determined by a grid-search with 3-fold cross validation.

The classifiers' accuracy is tested on a dataset generated from 1,000 uncompressed images from the BOSSbase dataset [40]. Each image is compressed to every possible pair of quality levels, and one window of every considered size is chosen randomly for testing. Fig. 2 shows the obtained classification rates for the decision in favor of double compression. Just as expected, large-window analysis yields more reliable results and is capable of successful classification for more compression combinations.

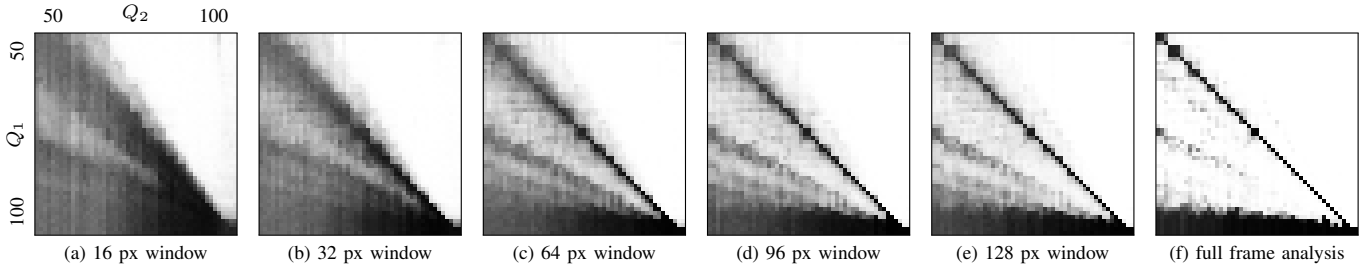
The obtained results are consistent with the phenomena already reported in the literature [11, 33]. For $Q_2 < Q_1$, the distribution of first digit features in DCT coefficients tends to be more similar, and correct classification requires more reliable statistics. Hence, the classification performance for $Q_2 < Q_1$ deteriorates rapidly with decreasing window size. The observed bands of better classification accuracy are most likely caused by accidental multiplicity of quantization steps in some of the sub-bands [41, 42]. Despite correct decisions of the classifier, the confidence tends to deteriorate and is typically insufficient for reliable tampering localization.

Classification problems around the diagonal stem from infinitesimal differences in the quantization steps which are insufficient to expose double compression. Similar behavior can also be observed when $Q_1 \gtrsim 95$ when the first JPEG compression is very close to lossless image representation. As a result, in order to generate a data set of reliable candidate maps, we will restrict our attention to the case of $Q_2 > Q_1$.

### C. Detailed Analysis of Candidate Maps

In general, forensic classifiers might exhibit asymmetry of the decision confidence when distinguishing between tampered and pristine regions. This phenomenon will also be visible in this study, as the decisions in favor of double compression tend to have higher confidence leading to slightly different behavior depending on the considered tampering scenario. In order to facilitate more general discussion and generalization to other forensic features, we consider two tampering scenarios corresponding to the presence of double JPEG compression either inside or outside of the tampered regions, in short referred to as the *double-inside* or *single-inside* scenarios, respectively. Regardless of the scenario, we expect the candidate maps to indicate tampered regions with scores $\approx 1$ and pristine regions with scores $\approx 0$. Hence, for the single inside scenario the candidate scores will correspond to 1 - classification scores.

Fig. 3 shows the behavior of the classification scores. The asymmetry of the decision confidence can be observed in the middle row, which shows normalized histograms of classification scores for the 16 px window. The histograms were obtained empirically based on 1,000 meaningful candidate maps. The successive rows correspond to the *double-inside* and *single-inside* tampering with a *rectangular* tampering pattern




(a) 16 px window (b) 32 px window (c) 64 px window (d) 96 px window (e) 128 px window (f) full frame analysis

Fig. 2. Classification rate in favour of the doubly-compressed decision for various analysis window sizes and a densely sampled compression grid; for reference the results for full frame analysis are shown in (f); high-resolution figures and full numeric data are available in supplementary materials.

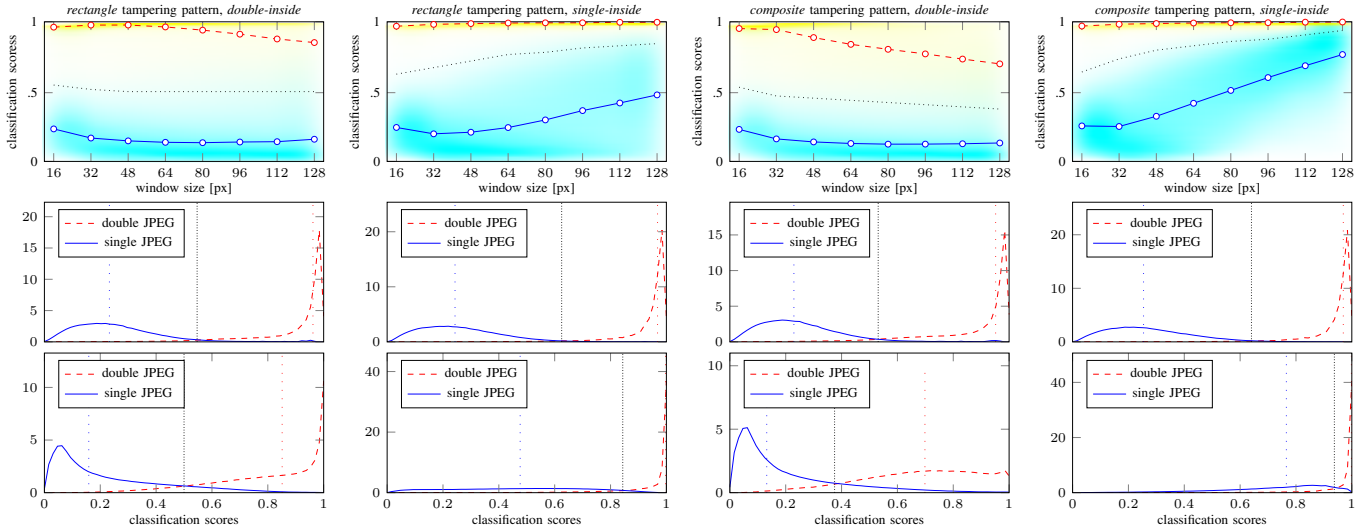



Fig. 3. Behavior of the classification scores for different tampering patterns and tampering scenarios: median of the candidate scores for various analysis windows with a color-marked distribution underlay (top); distribution of classification scores for a 16 px window (middle) and a 128 px window (bottom); dotted lines represent the medians of individual distributions (loosely dotted) and a hypothetical optimal decision threshold (densely dotted).

(1st and 2nd column) and a *composite* pattern (3rd and 4th column) with two disjoint regions of a rectangular and circular shape. Despite significant differences in the shape and size of individual regions, the distributions maintain virtually the same shape with a visible bias towards more reliable detection of double compression (red dashed line). This confidence asymmetry leads to different character of typical errors in the candidate maps. While in the double-inside scenario false positive errors are more likely, the single-inside scenario will be more prone to false-negative errors.

Large-scale analysis does not reveal such distinct similarities since the relatively larger overlap of the sliding windows will boost the differences stemming from decision confidence asymmetry, especially in the presence of small regions, or on the boundaries. The impact of these phenomena on the classification scores is shown in the bottom row of Fig. 3 which presents the corresponding histograms for a 128 px window. Consider for example the rectangular tampering pattern in the double-inside scenario (1st column). In the larger scale, we can observe better confidence in favor of the single compression decision (blue line) caused by a stronger statistical discriminability in the larger-scale window. We can also observe a heavier tail of the distribution of the scores in favor of double compression (red dashed line) caused by the larger overlap near region boundaries. In the single-inside scenario (2nd and 4th column), the smaller region with single compression and less reliable decisions can more easily become overwhelmed by the more confident doubly-compressed background.

These phenomena can also be observed in the medians of the decision scores for increasing window sizes (top row in Fig. 3). At the beginning, the medians tend to move towards more reliable decisions, which can be explained by increasing statistical discriminability of the forensic features. However, this improvement is dominated by the loss of confidence due to heterogeneous windows with both authentic and forged content. The impact of such regions is more severe when the tampered region corresponds to the less confident decision (single-inside scenario in the 2nd and 4th column).

### *D. Exploiting Inter-scale Dependencies*

The candidate maps are characterized by strong correlations both between the neighbors in the image plane, and between various scales of analysis. Exploitation of the latter dependencies is the key to successful multi-scale fusion.

Regardless of the tampering scenario, the candidate maps reveal some common behavior. Note that small-scale analysis

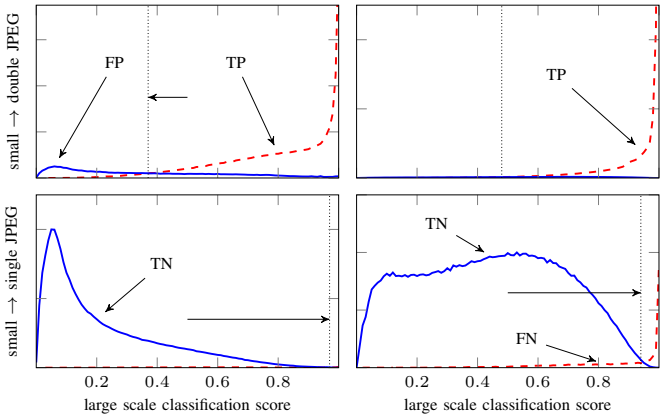



Fig. 4. Conditional distributions of large-scale classification scores predicated on small-scale scores for the *double-inside* (left) and *single-inside* (right) scenarios; the top (bottom) row shows 128 px scale scores in regions identified as doubly (singly) compressed by the 16 px scale; true labels are distinguished by the line style; hypothetical best threshold is marked with a dotted line; TP, TN, FP, and FN markers refer to the outcome of the small-scale decision.

indeed yields tampering maps with better resolution, but with a strong noise caused by worse decision confidence and classification errors. As the analysis scale increases, the maps become more reliable at the cost of the analysis resolution and the resulting inevitable problems with detecting small tampering and distinguishing separate regions in close proximity. However, as observed in Section II-C, it is practically impossible to accurately model the distributions of the candidate scores.

In order to exploit the correlations between various scales of analysis, we propose to predicate the decisions for larger scales based on the candidate scores from smaller scales. As already demonstrated in Section II-C, small-scale windows tend to have more stable distributions of the candidate scores. Fig. 4 shows conditional distributions of the scores of a 128 px window predicated on the decisions from a 16 px window (based on a hypothetical threshold of 0.5) - it is worth to compare to the original distributions in Fig. 3 conditioned on the true labels). The distributions were obtained analogously - from 1,000 meaningful candidate map sets for the *square* tampering pattern. The figure also labels the outcomes of the small-scale analysis as true positives, false positives, true negatives, and false negatives. Just as expected, there is a negligible amount of false negatives and false positives, in the *double-inside* and *single-inside* scenarios, respectively.

It can be observed that many errors originating from the smaller-scale analysis can easily be corrected in the larger scale where they exhibit confident correct decisions. Moreover, since the distributions are not symmetrical, further improvement could potentially be achieved by adjusting the decision thresholds. Such an approach would sacrifice correctly classified regions with low confidence (scores close to $0.5$) to obtain correct classification of more frequent errors in the second class. A new hypothetical threshold that minimizes the amount of classification errors is shown with dotted lines in Fig. 4 (along with a shift vector from the default 0.5 threshold used at the smaller scale).

### E. Determining Candidate Map Reliability

Small scale analysis yields less reliable candidate maps, which may unnecessarily introduce noise to the fusion process. Additionally, large scale analysis may produce empty candidate maps if the tampered region is smaller than the analysis window. Since such candidate maps do not contribute information useful for tampering localization, ideally they should be ignored by the fusion procedure. In this section, we describe a simple algorithm that allows to quickly estimate whether a candidate map is useful or not.

Due to potential problems with accurate modeling of valid maps, we follow a simpler approach and identify candidate maps that resemble Gaussian noise. For this purpose, we use maximum likelihood estimation to fit a Gaussian distribution to the candidate scores in reliable regions - determined by comparing the reliability scores to a threshold $p_{\text{th}}$. The obtained parameters $\mu, \sigma$ are then used to quickly identify meaningless low-variance noise ($\mu = 0.5 \pm 0.05, \sigma < 0.05$) and empty meaningless maps ($|0.5 - \mu| > 0.4, \sigma < 0.05$) that are immediately rejected. The mentioned thresholds were chosen empirically and validated by visual inspection of the obtained results. For the remaining maps we then measure the similarity between the empirical histogram $\hat{Q}$ and the accordingly sampled normal fit $Q$ using the Kullback-Leibler (KL) distance $D_{\text{KL}}(\hat{Q}||Q)$. In the performed experiments, the histogram is computed with 25 bins over the interval $(0, 1)$. The final decision is made by comparing the KL distance to a threshold $\theta_{\text{KL}}$ chosen as described in Section IV-B.

## III. Considered Fusion Strategies

This section presents the considered fusion methods. We begin by describing the proposed fusion mechanisms based on energy minimization (*EM fusion*, Sections III-A) and iterative top-down and bottom-up improvements (*TD/BU fusion*, Section III-B). Finally, in Section III-C, we describe the remaining methods based on majority voting (*MV fusion*), candidate map averaging (*AV fusion*), supervised learning (*SL fusion*), and clustering analysis (*CA fusion*) that are used as baseline for performance comparison.

All considered fusion methods identify unreliable regions of candidate maps based on the reliability scores $\mathbf{p}^{(s)}$. Authentication units with reliability scores below a threshold $p_{\text{th}}$ are considered unreliable. Each candidate scale can potentially identify different regions which increases the chance that at least some of the maps will contain meaningful candidate scores. Except for the EM and SL fusion, all other methods follow the widely accepted conservative approach [43] and label such regions as authentic (set $c_i = 0$) since they are not equipped with tools to resolve this ambiguity.

### A. Fusion by Energy Minimization

A Bayesian approach to tampering localization would involve finding the optimal tampering map $\hat{\mathbf{t}}$ that maximizes the posterior probability given a set of candidate maps:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}\in\{0,1\}^N}{\operatorname{argmax}} \; P(\mathbf{t}|\mathbf{c}^{(s)} : s = 1, 2, \ldots, S) \; . \tag{1}$$

Then, ignoring the irrelevant constant term, the problem can be rewritten as:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}\in\{0,1\}^N}{\operatorname{argmax}} \; P(\mathbf{c}^{(s)} : s = 1, 2, \ldots, S|\mathbf{t})P(\mathbf{t}) \; . \tag{2}$$

Due to analytical tractability issues, full independence of the observations for individual authentication units is commonly assumed, leading to a simpler formulation:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}\in\{0,1\}^N}{\operatorname{argmax}} \; \prod_{i=1}^{N} P(c_i^{(s)} : s = 1, 2, \ldots, S|t_i)P(\mathbf{t}) \; . \tag{3}$$

Analogously, we find it more practical to assume inter-scale independence of the candidate scales at this point, and introduce a simpler heuristic mechanism for exploiting these dependencies at a later stage. Then, the problem becomes:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}\in\{0,1\}^N}{\operatorname{argmax}} \; \prod_{i=1}^{N}\prod_{s=1}^{S} P(c_i^{(s)}|t_i)P(\mathbf{t}) \; . \tag{4}$$

The prior of the tampering map $P(\mathbf{t})$ can be conveniently modeled with a MRF. Then, the decision for each authentication unit will depend only on its direct neighborhood. Assuming a 1st order neighborhood, the decision regarding $t_i$ will depend only on up to 4 of its neighbors $t_j : j \in \Xi_i = \{i-1, i+1, i-N_h, i+N_h\}$, corresponding to the top, bottom, left, right neighbors, respectively. Obviously, at the image borders the set of neighbors $\Xi_i$ needs to be pruned accordingly.

In practice it is often more convenient to represent the MRF in terms of Gibbs potentials, and reformulate the problem into energy minimization [44, 45]. Gibbs potentials use probabilities in the form:

$$P(\mathbf{t}) = Z^{-1}e^{-U(\mathbf{t})} = Z^{-1}e^{-\sum_{c\in C} V_c(\mathbf{t})} \; , \tag{5}$$

where $Z$ is a normalizing constant, and $U$ is an energy function defined as a sum of potentials $V_c$ on individual *cliques* - small groups of neighboring nodes in a graphical model of the MRF (authentication units in the problem at hand). Similarly to Chierchia et al. [7], we resort to the popular Ising model [45] which considers single-element and two-element cliques.

Finally, by taking a negative logarithm of (4), the multi-scale fusion problem can be solved by minimizing the following energy function:

$$\frac{1}{S}\sum_{i=1}^{N}\sum_{s=1}^{S} E_\tau(c_i^{(s)}, t_i) + \alpha\sum_{i=1}^{N} t_i + \beta\sum_{i=1}^{N}\sum_{j\in\Xi_i} |t_i - t_j| \; . \tag{6}$$

In our work we use a graph cuts-based solver [46, 47] from the UGM toolbox [48] to quickly find the optimal tampering map. The first term in (6), referred to as the *data term*, is responsible for maintaining resemblance to the candidate maps, and is normalized by the number of candidate maps $S$ for the sake of better stability of the parameters $\alpha, \beta$. The second term introduces a penalty $\alpha$ for every tampered authentication unit, and thus encodes a preference towards sparser solutions. The third term penalizes the differences between neighboring authentication units, leading to a preference towards piece-wise constant solutions, and eliminating noisy and small unconnected components[1]. The choice of $\alpha, \beta$ will be discussed in Section IV-B.

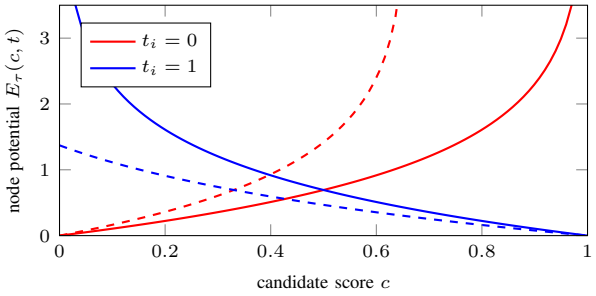



Fig. 5. Node potentials for the data term for both decision labels $t_i = 0$ and $t_i = 1$; $\tau = 0.5$ (solid); $\tau = 0.33$ (dashed).

The data term is composed of potentials that penalize the differences from the candidate tampering maps. Since the candidate scores correspond to the probability that a given block is tampered, the data terms could be obtained as $-\log(c_i)$ and $-\log(1-c_i)$, respectively. However, we use the following generalization:

$$E_\tau(c,t) = -\log\,\max(\Psi_{\text{min}}, \Psi_\tau(c,t)), \tag{7}$$

with $\Psi_{\text{min}} \in [0,1]$ and:

$$\Psi_\tau(c,t) = \begin{cases} 1 - \frac{c}{2\tau} & \text{for } t = 0, \\ 1 + \frac{c}{2(1-\tau)} - \frac{1}{2(1-\tau)} & \text{for } t = 1, \end{cases} \tag{8}$$

where $\tau \in (0,1)$ is a quasi-threshold that equalizes potentials for both decisions[2], i.e., $E_\tau(\tau, 0) = E_\tau(\tau, 1)$. The shape of the adopted energy function, and the impact of the quasi-threshold for both decisions and $\tau = 0.33$ and $0.5$ are shown in Fig. 5. Setting a minimal value $\Psi_{\text{min}}$ prevents the nodes from becoming fixed to certain decisions (due to infinite energy). We set $\Psi_{\text{min}} = 0.001$ based on preliminary experiments.

In order to exploit the dependencies between different scales of analysis, we propose a simple mechanism of *threshold drift*. As observed in Section II-D, adjustments of the decision thresholds based on the intermediate hypothetical decisions from smaller-scale analysis could yield significant improvement of the classification accuracy. Hence, the general idea of threshold drift is to vary the threshold for every authentication unit, and update the thresholds $\tau_i^{(s)}$ for large analysis windows

[1]The interaction strength $\beta$ could be chosen adaptively, e.g., based on the actual image content. Such an approach would encourage stronger propagation within similar objects or image segments, and could be seen as a variant of structure transfer to the tampering map - in essence similar to guided filtering [49]. Such an approach would allow to improve the resolution of the tampering map by exploiting real object boundaries on a pixel level. However, our study focuses on synthetic forgeries that do not alter semantic content of the image, which limits the potential gains in this setting. We leave this aspect for our future work.

[2]The tampering penalty $\alpha$ can also be seen as a decision bias with similar impact to the quasi-threshold; as such, it can also be used for controlling the classification trade-offs under fixed $\tau$; such an approach was used, e.g., in [7]. However, in our experiments it gave worse results than our generalization.

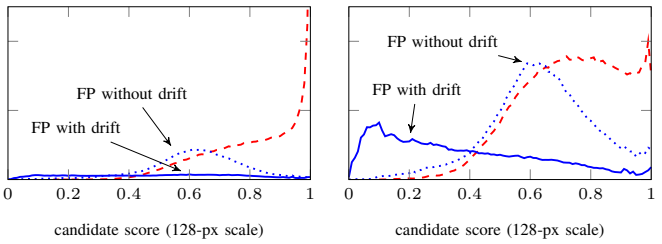



Fig. 6. Impact of the threshold drift on the conditional distributions of the large-scale candidate scores for the *rectangle* (left) and *composite* (right) tampering patterns.

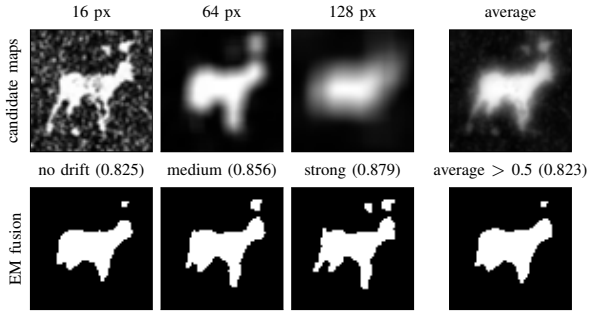



Fig. 7. Impact of the threshold drift on EM fusion of 8 multi-scale maps; numbers in brackets correspond to $F_1$ scores; the average candidate map and its thresholded version are shown for reference.

based on the values from smaller scales, i.e.:

$$\tau_i^{(s)} = \begin{cases} \tau^{(1)} & \text{if } s = 1\ , \\ \tau_i^{(s-1)} & \text{if } s > 1 \text{ and } p_i^{(s)} \leq p_{\text{th}}, \\ \tau_i^{(s-1)} + \delta & \text{if } s > 1 \text{ and } c_i^{(s-1)} \leq \tau_i^{(s-1)}, \\ \tau_i^{(s-1)} - \delta & \text{if } s > 1 \text{ and } c_i^{(s-1)} > \tau_i^{(s-1)}\ , \end{cases} \tag{9}$$

where $\delta \in \mathbb{R}_+$ is the strength of the drift and $\tau^{(1)}$ denotes an initial threshold, typically chosen around 0.5. The directions of the drift are consistent with the observations from Section II-D. In order not to discard confident scores from larger scales, we do not drift the threshold above 0.95 or below 0.05. The strength $\delta$ will be determined experimentally (Section IV-B).

The efficiency of this simple approach is demonstrated in Fig. 6 showing conditional distributions of candidate scores from a 128 px window predicated successively on all of the considered smaller window sizes, i.e., on a 112 px window, which was predicated on the 96 px window, etc. It can be observed that the threshold drift can effectively improve the separation between the two distributions, and reduce the amount of classification errors. As a result, it allows to partially recover small regions captured only by small-scale analysis. The impact of the threshold drift on the final tampering map is shown in Fig. 7.

In contrast to other fusion methods, EM fusion can use the MRF-based prior to propagate the decisions from reliable image regions into unreliable regions identified by the reliability maps. Unreliable candidate scores are ignored, and replaced with a predefined value $c_{\text{sat}}$. Choosing $c_{\text{sat}} = 0$ corresponds to the mentioned conservative approach, but still allows the solver to change this decision if appropriate. Choosing $c_{\text{sat}} \approx \tau$

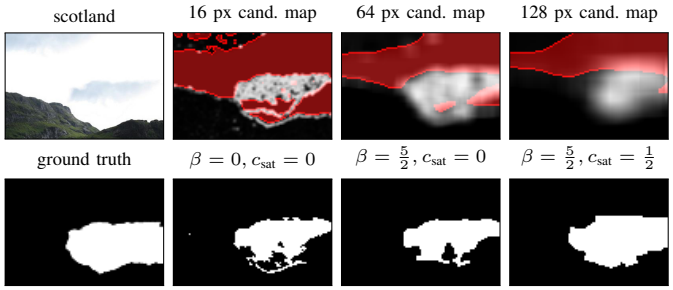



Fig. 8. Propagation of reliable decisions into unreliable saturated regions by the EM fusion; unreliable regions - determined by reliability maps $\mathbf{p}^{(s)}$ - are overlaid in red onto candidate maps; bottom row shows the impact of the parameters $\beta, c_{\text{sat}}$ leading to successive $F_1$ scores: 0.764, 0.746, and 0.837.

introduces more ambiguity and makes it easier for the solver to propagate reliable scores from the neighborhood into unreliable regions. All of the mentioned phenomena can be observed in Fig. 8 which shows the impact of both $c_{\text{sat}}$ and $\beta$ on the fusion result for an example highly saturated image.

For best results, $c_{\text{sat}}$ should be chosen individually for every image. Due to imperfect nature of unreliable region identification, it is not desirable to blindly set $c_{\text{sat}} \approx \tau$ as it may lead to propagation of unidentified unreliable misclassified regions. Especially since saturated regions may produce highly confident incorrect decisions. In our study saturated regions were typically classified as doubly-compressed with nearly perfect decision probabilities. Assuming that the estimation of unreliable regions is sufficiently accurate, such errors are typically easily eliminated by choosing sufficiently large $\beta$.

### B. Bottom-Up and Top-Down Fusion

The proposed bottom-up (BU) and top-down (TD) fusion are heuristic approaches exploiting the observation that large-scale analysis is expected to be more reliable, yet with a worse resolution of analysis. The BU fusion uses the candidate map from the smallest scale as an initial estimate, and successively eliminates classification errors by considering larger analysis windows. The TD fusion begins with the largest scale, and successively refines the shape based on small scale maps.

The operation of both TD and BU fusion is illustrated in Fig. 9. The relevant parameters for every step are marked with dotted arrows. Both methods begin by discarding unreliable candidate maps (Section II-E). The remaining maps are then clustered using K-means with 2 centroids to obtain estimated intermediate binary decision maps $\mathbf{t}^{(s)}$. Decisions in unreliable map regions are then set to 0, and the final pre-processed individual maps are obtained by removing connected components (CC) smaller than $\theta_{\text{c}}$ authentication units.

The main step of both algorithms involves successive updates of the tampering map's initial estimate $\mathbf{t}^e$, initialized as $\mathbf{t}^e = \mathbf{t}^{(1)}$ and $\mathbf{t}^e = \mathbf{t}^{(S')}$ for the BU and TD fusion, respectively. In the BU fusion, the algorithm considers the candidate maps in the order of their increasing window size. Each CC in the current estimate $\mathbf{t}^e$ is verified against the next candidate scale $\mathbf{t}^{(s)}$ by comparing the rate of re-detected authentication units with a threshold $\theta_{\text{fp}}$. Invalid CCs are removed from $\mathbf{t}^e$. In order to allow for detection of multiple

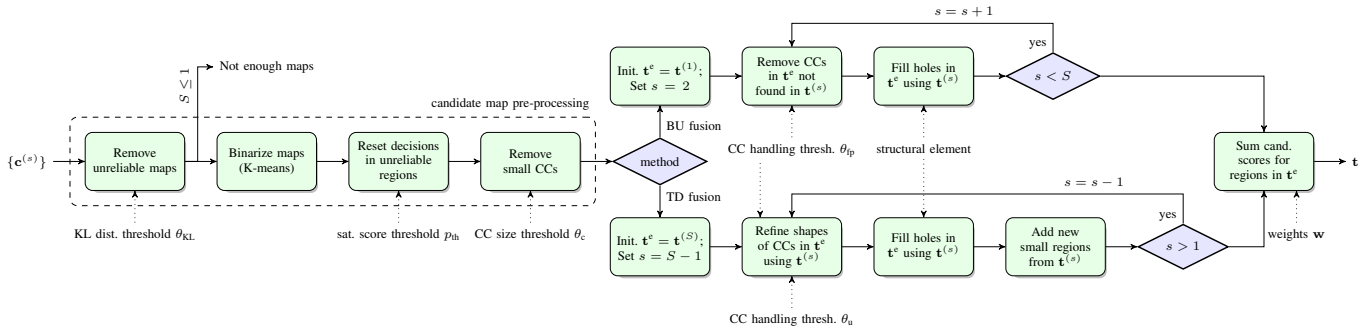



Fig. 9. Operation of the bottom-up and top-down fusion; relevant parameters for every step are marked with dotted arrows. Operation of the bottom-up and top-down fusion; relevant parameters for every step are marked with dotted arrows.

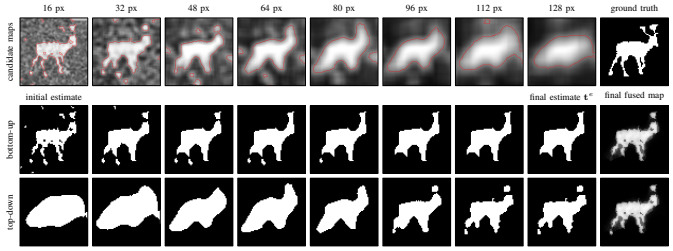



Fig. 10. Comparison of successive updates of $\mathbf{t}^e$ in the *bottom-up* and *top-down* fusion (bottom rows); intermediate decisions for individual scales $\mathbf{t}^{(s)}$ are overlaid in red for each of the candidate maps $\mathbf{c}^{(s)}$ (top row).

tampered regions with significantly different size, this filtering step applies only to regions that can potentially be detected in the current scale of analysis. We consider applicable scales to have window size smaller than the minimum dimension of the CC's bounding box. In order to improve robustness, one may consider relaxation of this condition and permanently accept CCs that have been noticed sufficient number of times.

Analogously, the TD fusion successively updates the shape of the identified regions using smaller scale maps. For every CC in the current estimate $\mathbf{t}^e$, the algorithm attempts to replace its rough shape with a better estimate from a smaller scale map. The regions are matched by comparing the relative overlap between the rough and the small-scale estimates to a threshold $\theta_r$. If no match better than $\theta_{fp}$ can be found, the region is considered a false positive and discarded. In TD fusion it is necessary to take into account that a single shape from large-scale analysis might be replaced with multiple shapes in small-scale analysis. Additionally, detection of multiple tampered regions of different size requires explicit insertion of smaller regions once the current scale is small enough for their detection.

In both BU and TD fusion the last step of the iterative update procedure involves filling prospective holes in the smaller-scale map. Locations of holes are determined by comparing $\mathbf{t}^e$ with its morphologically closed version and the decisions are updated from the larger scale map. Successive updates in BU and TD fusion are illustrated in Fig. 10. The final tampering map in both strategies is obtained by weighted summation of the the candidate scores $c_i^{(s)}$ in the regions identified both in $\mathbf{t}^{(s)}$ and $\mathbf{t}^e$. If binary decisions are needed, the resulting map $\mathbf{t}$ is compared against a threshold $\tau$.

### C. Fusion Methods for Comparison

As a reference for performance comparison we consider 4 fusion strategies representing various approaches to the problem. The majority voting (*MV fusion*) strategy performs decision-level fusion. Each individual candidate map is separately compared to a threshold $\tau$, and the final tampering map is obtained by majority voting with prospective ties resolved in favor of the *tampered* decision. The averaging strategy (*AV fusion*) performs measurement-level fusion by comparing the average candidate map $\frac{1}{S}\sum_{s=1}^{S}\mathbf{c}^{(s)}$ to a threshold.

The supervised learning strategy (*SL fusion*) uses machine learning to make a decision for each of the authentication units separately. We used a SVM with the RBF kernel for classification based on a set of $2S + 6$ features including: $S$ candidate scores $c_i^{(s)}$; $S$ reliability scores $p_i^{(s)}$ (normalized to the [0,1] range by $f(x) = 1 - e^{-x}$); scores of four immediate neighbors $c_j^{(s)} : j \in \Xi_i$; and the average scores in the $3 \times 3$ and $5 \times 5$ neighborhoods. The neighborhood-related features were extracted from the candidate scale with best individual performance, i.e., 32 px. We used a Gaussian-like filter (with the central element excluded) to obtain the average scores.

The SVM parameters $\gamma$ and $C$ are determined by a grid-search with 3-fold cross-validation. The classifier is trained on randomly sub-sampled examples from 1,600 sets of candidate maps originating from all of the considered tampering patterns and from both tampering scenarios. The number of training examples was chosen as 40,000 which corresponds to nearly saturated classification accuracy (91.96%). In order to demonstrate the trade-off between classification accuracy and fusion time, we include a smaller and faster version of the classifier, trained on 2,500 examples (denoted as *SL'*; accuracy 91.04%).

The clustering analysis (*CA fusion*) strategy uses K-means to identify two separate clusters of authentication units, corresponding to authentic and tampered regions. Each unit is described by a $S$-dimensional feature vector corresponding to the candidate scores from all scales $c_i^{(s)}$. Depending on the tampering scenario, we group the authentication units into 3 (double inside) or 2 (single inside) clusters. The centroids are initialized as constant vectors of zeros (pristine regions), ones (tampered regions), and $1/4$ (uncertain boundaries around the

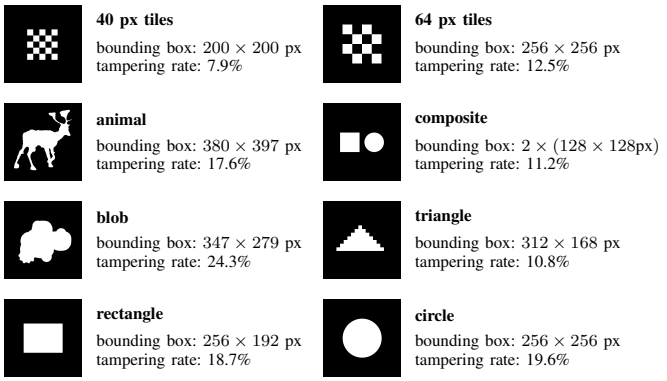



Fig. 11. Considered tampering patterns along with their basic metrics (relative to the original image size - $512 \times 512$ px); white regions represent tampered areas.

actual tampering). Prospective empty clusters are reinitialized with a new centroid in the most distant location.

## IV. Experimental Evaluation

This section presents the experimental evaluation results. We begin with a detailed description of our testing sets. Then, we discuss parameter selection for the proposed EM and TD/BU fusion strategies. Finally, we compare the performance of the proposed approaches. The primary performance measure is the commonly used $F_1$ score:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \ , \tag{10}$$

which is more representative than accuracy for tampering-localization. For the sake of discussion completeness, we also present the corresponding receiver operation characteristics.

### A. Data Sets

We consider two data sets: a large set of *synthetic forgeries*, and a small set with *realistic forgeries*. In order to obtain the synthetic forgery dataset, we generated 32,000 sets of multi-scale candidate maps based on synthetic splicing forgeries by replacing some regions of an image with the same content but with a different compression history. The affected regions were chosen according to 8 patterns (Fig. 11) that include both simple and complex shapes. For each pattern, two versions of the forgery were generated corresponding to the *double-inside* and *single-inside* scenarios. In order to increase data diversity the pattern was randomly placed within the image.

The tampered images were generated from 200 uncompressed natural images from the *BOSSbase* dataset. For each input image, we produced 10 versions of every tampering configuration, each with a different combination of the first and the second JPEG quality factors chosen uniformly from $Q_1 \in \{50, 51, \ldots, 99\}, Q_2 \in \{Q_1 + 1, \ldots, 100\}$. The candidate maps were obtained with window-based analysis (Section II).

Finally, the obtained multi-scale map sets were filtered to remove unreliable sets where none of the maps contained meaningful results. The selection was performed according to the algorithm described in Section II-E. For performance evaluation, we chose a subset of 24,000 candidate map sets (1,500 map sets for 16 tampering configurations). The remaining maps were used for training of the SVM classifier (SL fusion) and parameter search (EM fusion).

Our realistic forgery dataset was derived from an existing test set, originally prepared for evaluation of copy-and-move forgery detection methods [50]. The set contains 48 uncompressed color images, definitions of the copy and move attacks, and additional tools to perform the forgeries with optional pre- or post-processing. We resized the images to a unified width of 800 px, and compressed them to $Q_1 = 80$. After the forgery, the images were saved with $Q_2 = 90$. The ground truth maps were generated by comparing the difference of the images against a threshold of 4, followed by removal of small noise by morphological closing (with a disk-shaped structural element of size 4 px), and optional manual refinement. The multi-scale candidate maps were obtained as before.

### B. Parameter Selection

Candidate map filtering is controlled by a threshold on the KL distance $\theta_{\mathrm{KL}}$, chosen as the local maximum of the average $F_1$ score for the AV fusion on a subset of 4,096 candidate maps sets. The initial increase of the threshold $\theta_{\mathrm{KL}}$ causes an increase in the $F_1$ score which can be attributed to successive removal of the noise caused by unreliable candidate maps. After a certain point, the $F_1$ scores begin to deteriorate which indicates removal of maps with useful information.

While such a procedure is not expected to provide ideal separation between reliable and unreliable maps, visual inspection confirms satisfactory performance. Several example results of this procedure, and the corresponding graph of the $F_1(\theta_{\mathrm{KL}})$ dependency can be found in supplementary materials. Similar evaluation was also repeated with the EM and TD fusion, leading to the same value of the threshold $\theta_{\mathrm{KL}} = 0.05$.

*a) EM Fusion:* The EM fusion contains 3 important parameters: $\alpha$ controls the preference towards sparser tampering maps; $\beta$ controls the neighborhood interaction strength; $\delta$ controls the threshold drift. When $\alpha = \beta = \delta = 0$, the EM formulation is equivalent to naive candidate map averaging (AV fusion). We search for the best values over the following lattice: $\alpha \in \{0.0, 0.025, \ldots, 0.5\}$, $\beta \in \{0.0, 0.2, \ldots, 6.0\}$, and $\delta \in \{0.0, 0.025, \ldots, 0.5\}$. The search is driven by the $F_1$ score calculated on a sub-set of 6,400 candidate map sets (400 sets for each of the considered 16 configurations).

The results are collected in Table I and contour plots of selected $\alpha \times \beta$ grids are visualized in Fig. 12. Based on the collected results, we can observe that large, solid and regular shapes (black markers in Fig. 12) tend to reveal similar behavior, distinct from more irregular shapes (magenta markers). This can be motivated by two observations. Firstly, larger solid shapes are typically more reliably detected in large-scale analysis and significantly lower strength of the threshold drift is sufficient to accurately recover their boundary. This is well visible on the example of the *40 px tiles* that required the threshold drift over 3 times as strong as the *circle* pattern.

Secondly, it is clear that larger regular shapes prefer greater values of $\beta$, which essentially penalizes regions with longer

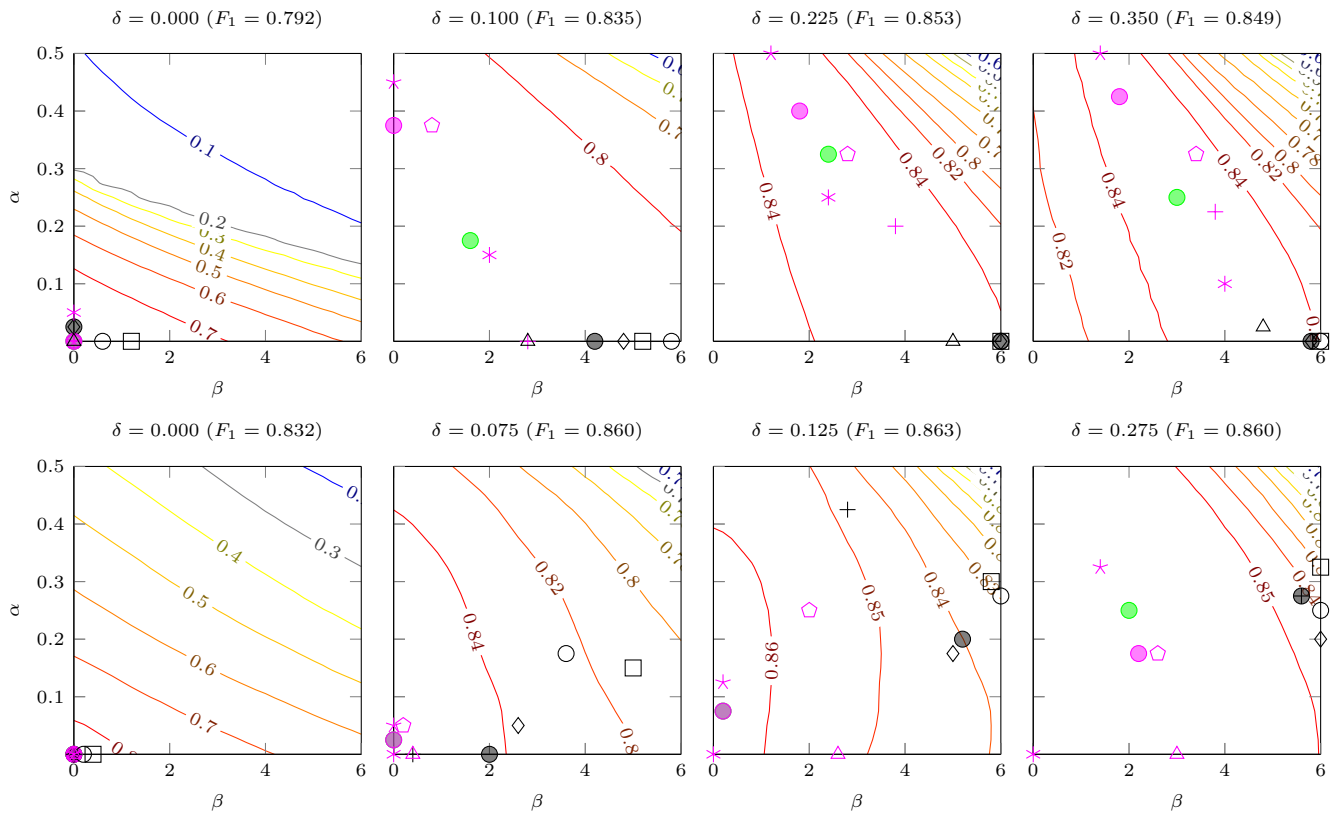



Fig. 12. Contour plots of the average $F_1$ scores over all tampering patterns along with the best observed combinations of $(\alpha, \beta)$ for individual patterns (plain markers), similar pattern groups (magenta and black-filled circles) and for all patterns altogether (green-filled circles); double-inside scenario (top); single-inside scenario (bottom); note that points may overlap - the overall best solution (green) tends to be dominated by more sensitive irregular patterns (magenta) - refer to hue differences in the color version of the figure.

TABLE I
PARAMETER SELECTION FOR THE EM FUSION.

| Pattern | | double-inside | | | single-inside | | |
|---|---|---|---|---|---|---|---|
| | | $\alpha$ | $\beta$ | $\delta$ | $\alpha$ | $\beta$ | $\delta$ |
| ⋆ | 40 px tiles | 0.500 | 1.4 | 0.350 | 0.325 | 1.4 | 0.275 |
| ⬠ | 64 px tiles | 0.325 | 2.8 | 0.225 | 0.175 | 2.6 | 0.275 |
| ∗ | animal | 0.250 | 2.4 | 0.225 | 0.000 | 0.0 | 0.125 |
| + | composite | 0.200 | 3.8 | 0.225 | 0.425 | 2.8 | 0.125 |
| △ | triangle | 0.000 | 2.8 | 0.100 | 0.000 | 2.6 | 0.125 |
| ◇ | blob | 0.100 | 4.8 | 0.100 | 0.175 | 5.0 | 0.125 |
| □ | rectangle | 0.000 | 5.2 | 0.100 | 0.300 | 5.8 | 0.125 |
| ○ | circle | 0.000 | 5.8 | 0.100 | 0.175 | 3.6 | 0.075 |
| Irregular shapes | | 0.425 | 1.8 | 0.250 | 0.000 | 0.2 | 0.175 |
| Solid shapes | | 0.000 | 5.0 | 0.125 | 0.200 | 5.2 | 0.125 |
| General | | 0.325 | 2.4 | 0.225 | 0.075 | 0.2 | 0.125 |

edges, and thus eliminates smaller parts and details of the detected shapes, e.g., legs in the *animal* pattern or corners in the *triangle* pattern. As a result, irregular shapes will typically benefit from the simpler preference towards sparser tampering maps, controlled by $\alpha$. This phenomenon can also be observed for $\delta = 0$ where most of the patterns prefer $\alpha, \beta \approx 0$. As the boundaries and fine details become more reliably represented in the energy of the data term along with increasing threshold drift, the best values of $\alpha$ and $\beta$ quickly increase.

Finally, in the following experimental evaluation we consider two parameter choice variants. Firstly, we choose a general set of parameters $\alpha, \beta, \delta$, regardless of the tampering pattern (separately for the double-inside, and single-inside scenarios). This variant is denoted as EM'. Secondly, we choose a separate set of parameters for large regular regions and for irregular regions with small details. This variant is denoted as EM. The chosen parameters are shown in Table I.

While $\tau$ is not explicitly used as a decision threshold, it serves the same purpose and can be used to control the trade-offs in the classification performance. In our evaluation we observed the best average results for $\tau \approx 0.5$ in the double-inside scenario, and $\tau \approx 0.45$ in the single-inside scenario.

*b) BU and TD fusion:* The following parameter values were chosen empirically by trial and error: the small connected component rejection threshold $\theta_c = 10$; the false positive rejection threshold $\theta_{fp}$ is 0.25 (BU) and 0.1 (TD); the minimum region overlap threshold for shape update in the TD strategy $\theta_u = 0.75$; the region overlap threshold in the final summation step $\theta_s = 0.5$. The weights for the final summation were set to $\mathbf{w} = \frac{1}{255}[128,\ 64,\ 32,\ 16,\ 8,\ 4,\ 2,\ 1]$. Experimental evaluation shows little sensitivity to the values of individual weights as long as the general preference towards smaller scales is maintained.

Both BU and TD fusion produce relatively sharp decision maps. False positive errors are eliminated during iterative map refinement. As a result, these strategies are less sensitive to the choice of the final decision threshold $\tau$, which tends to be lower than for other approaches. In our investigation, the best results were typically achieved for $\tau$ between 0.2 and 0.4.

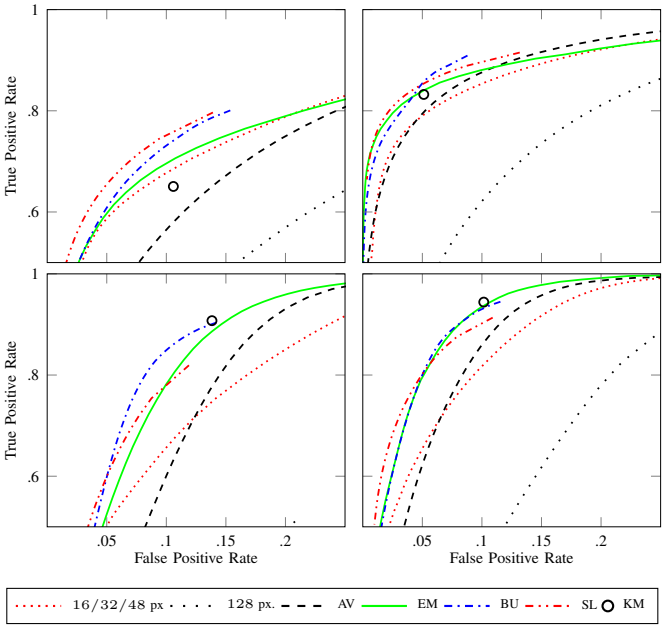



Fig. 13. Receiver operation characteristics for selected fusion methods for the *40 px tiles* (left) and *composite* (right) tampering patterns in the *double-inside* (top) and *single-inside* (bottom) scenarios; for reference the best and the worst individual candidate scales are also shown.

### C. Performance Comparison for Synthetic Forgeries

In this experiment we compare the performance of the considered fusion methods. The testing set consists of 24,000 candidate map sets, divided into 8 tampering patterns and two tampering scenarios (Section IV-A). We sweep the threshold $\tau$ over 50 possible values, unevenly distributed within $(0, 1)$ with higher concentration around 0.5. The receiver operation characteristic (ROC) curves for selected fusion methods and two example patterns are shown in Fig. 13. The best and the worst candidate maps are marked with dotted lines. Note that some methods may have a narrow range of achievable classification rates (TD/BU, SL) or have no control over the classification trade-off (CA).

In our evaluation we focus on the $F_1$ score. For every considered configuration we measure the highest achievable $F_1$ score over different thresholds $\tau$, which serves as an upper bound on the tampering localization capability. Fig. 14 shows the average $F_1$ scores for two selected patterns (*40 px tiles* and *composite*) and the average over all patterns for the *double-inside* (top row) and *single-inside* (bottom row) scenarios. The performance of the best individual candidate scale - which changes depending on the tampering pattern between 16 px and 64 px - is additionally marked with a dotted line. On average, the best performance is delivered by the 32 px scale.

The results clearly show that it is indeed possible to improve the performance of tampering localization by fusing the results from multi-scale analysis. However, simple strategies like MV or AV fusion are unsuitable for this purpose. While occasionally some improvement could be achieved (if the tampered regions were large and simple), they failed when the shapes were more complex. Hence, even though simple fusion strategies can roughly identify the tampered area despite meaningless small-scale maps, for precise tampering localization individual small-scale maps are still typically a better choice.

All of the proposed fusion methods (EM, TD/BU fusion) could always improve the tampering localization. The performance gap with respect to the considered reference fusion methods is particularly well visible for the most challenging *tiles* and *animal* patterns (e.g., *40 px tiles*; 1st column in Fig. 14). The BU fusion delivered the best results for 7 out of 8 patterns in the *double-inside* scenario, and only slightly worse than CA fusion for the remaining simple *circle* pattern. In the *single-inside* scenario the best results for 6 patterns were delivered by the EM fusion, and by the TD fusion for the remaining 2. Note that although specialized choice of parameters for the EM fusion gives certain performance improvement, the overall $F_1$ scores are similar.

Interestingly, simple clustering analysis provided surprisingly good performance. While it had problems with complex tampering patterns, it always outperformed the averaging and voting strategies. This observation seems to be in agreement with a recent proposition that detection of forensic feature inconsistencies could be a good approach to tampering localization [8]. Hence, more sophisticated clustering or anomaly detection might also be beneficial for multi-scale fusion. This issue might be an interesting direction for further research.

The SL strategy also delivered reasonable results. If trained on sufficiently large data set, it performs only slightly worse than the proposed dedicated fusion strategies. However, visual inspection of the resulting maps shows large amount of noise, including both false-positive and false-negative errors. It is also significantly slower than any other fusion strategy. Speed improvement is possible, but with a considerable accuracy penalty (see results for SL' fusion). This issue is discussed in detail in Section V.

Example fusion results are shown in Fig. 15 for diverse candidate maps with: disappearing large-scale candidates (2nd, 6th and 9th row); unreliable small-scale candidates (3rd, 5th, 7th, and 9th row); merging of separate regions in large-scale analysis (4th - 6th rows). More examples are included in supplementary materials. The decision thresholds were chosen for each image individually, hence the maps represent the best result each fusion method can offer ($F_1$ score wise).

### D. Performance Comparison for Realistic Forgeries

Evaluation of the realistic forgeries was performed analogously to the synthetic ones. In this case, however, we choose the best decision threshold $\tau$ for each image individually. The parameters were chosen accordingly to the previously distinguished *single-inside* scenario. We consider only one variant of the EM fusion with a single set of parameters, without any prior assumptions about the tampering pattern. We chose the default potential for saturated regions $c_{\text{sat}} = 0.4$.

The obtained average $F_1$ scores (Table II) again demonstrate superior performance of the proposed fusion methods. The choice of high and distinct JPEG quality levels guaranteed good performance (little noise) of individual single-scale detectors. Similarly to previous evaluation, the best individual scale varied from 16 px to 48 px windows, and the average

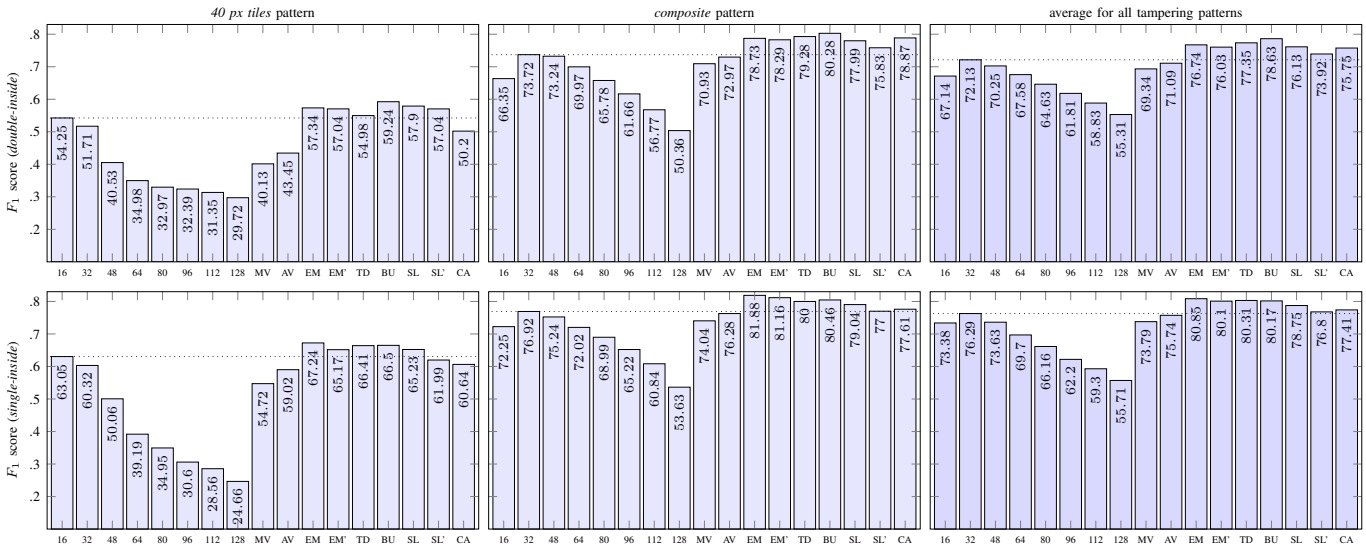



Fig. 14. $F_1$-based localization ranking for both individual candidate scales and the considered fusion methods; *40 px tiles* pattern (1st column); *composite* pattern (2nd column); average for all patterns (3rd column); for the sake of presentation clarity, the reported numbers are multiplied by 100.

TABLE II
AVERAGE $F_1$ SCORES, SLIDING-WINDOW ANALYSIS TIME AND DECISION FUSION TIME FOR REALISTIC FORGERIES

| | **Individual candidate scales [px]** | | | | | | | | **Oracle** | **Multi-scale fusion technique** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **16** | **32** | **48** | **64** | **80** | **96** | **112** | **128** | | **MV** | **AV** | **EM'** | **TD** | **BU** | **SL** | **SL'** | **CA** |
| $F_1$ | 84.7 | 85.5 | 80.1 | 73.3 | 53.9 | 45.9 | 43.4 | 37.2 | 87.0 | 84.2 | 85.4 | 87.6 | 86.3 | 87.5 | 85.6 | 82.9 | 74.3 |
| **Time [s]** | 6.72 | 6.16 | 4.33 | 3.02 | 3.36 | 2.73 | 2.17 | 1.97 | - | 0.004 | 0.004 | 0.015 | 0.05 | 0.10 | 3.03 | 0.22 | 0.01 |

best score ($F_1 = 0.855$) was obtained by the 32 px scale. A hypothetical oracle capable of choosing the best candidate scale for each image individually yielded the average $F_1$ score of 0.870. The proposed EM and BU fusion strategies were slightly better ($F_1$=0.876 for the EM' fusion, and $F_1$=0.875 for the BU fusion) than the oracle. This result shows that successful fusion techniques can properly exploit the information available in the best scale. A fixed choice of the analysis window size of 64 px, recommended by Amerini et al. [6], yielded the average $F_1$ score of only 0.733 - significantly worse than multi-scale fusion. Fig. 16 shows example fusion results for selected methods.

## V. DISCUSSION

In this section, we discusses practical implementation issues, limitations of the considered methods, and perspectives for future improvement.

### A. Computational Complexity

Computational complexity of the proposed multi-scale approach depends on the runtime of two principal components: decision fusion, and sliding-window analysis. The average time of decision fusion for 48 images (100 repetitions per image) in the realistic tampering dataset is collected in Table II. The results were obtained in Matlab (desktop PC with a 3.6 GHz Core i7-4790 processor) with potentially most time-consuming operations (graph cuts, and SVM classification) performed using MEX routines implemented in C/C++. Just as expected, the fastest methods were the MV and AV fusion, followed by the CA method. The proposed EM and BU/TD methods required longer, but still negligible runtime (<0.1 s).

The slowest of the considered methods was the SL fusion, which was very sensitive to the number of examples during SVM training. Due to poor separation of the classes, relatively large number of support vectors needs to be retained which negatively impacts the performance. The considered configuration with 40,000 training examples required on average 3 s to process a single image. By choosing a smaller training set, it is possible to trade-off the processing time with localization performance. In our experiments, we were able to reduce the processing time to 0.2 s (still the slowest fusion method) by using only 2,500 examples. While many fused maps appear similar, the numerical results show a considerable performance penalty (Table II and Fig. 14). Example fusion results for this scenario are included in supplementary materials.

The most time-consuming step is the sliding-window analysis. Our C++ implementation of the considered detector (window-based MBFDF, as described in Section II) needed on average 3.8 seconds for single-scale analysis of the images in the realistic tampering test set (measured on a desktop PC with a 3.6 GHz Core i7-4790 processor with 8 simultaneous threads). The average times for individual scales are collected in Table II. Although the problem is trivially parallel (windows can be evaluated independently) and thus well suited to contemporary computing architectures, it clearly indicates that decision fusion is not the bottleneck. Note also that other forensic features might require even more effort.

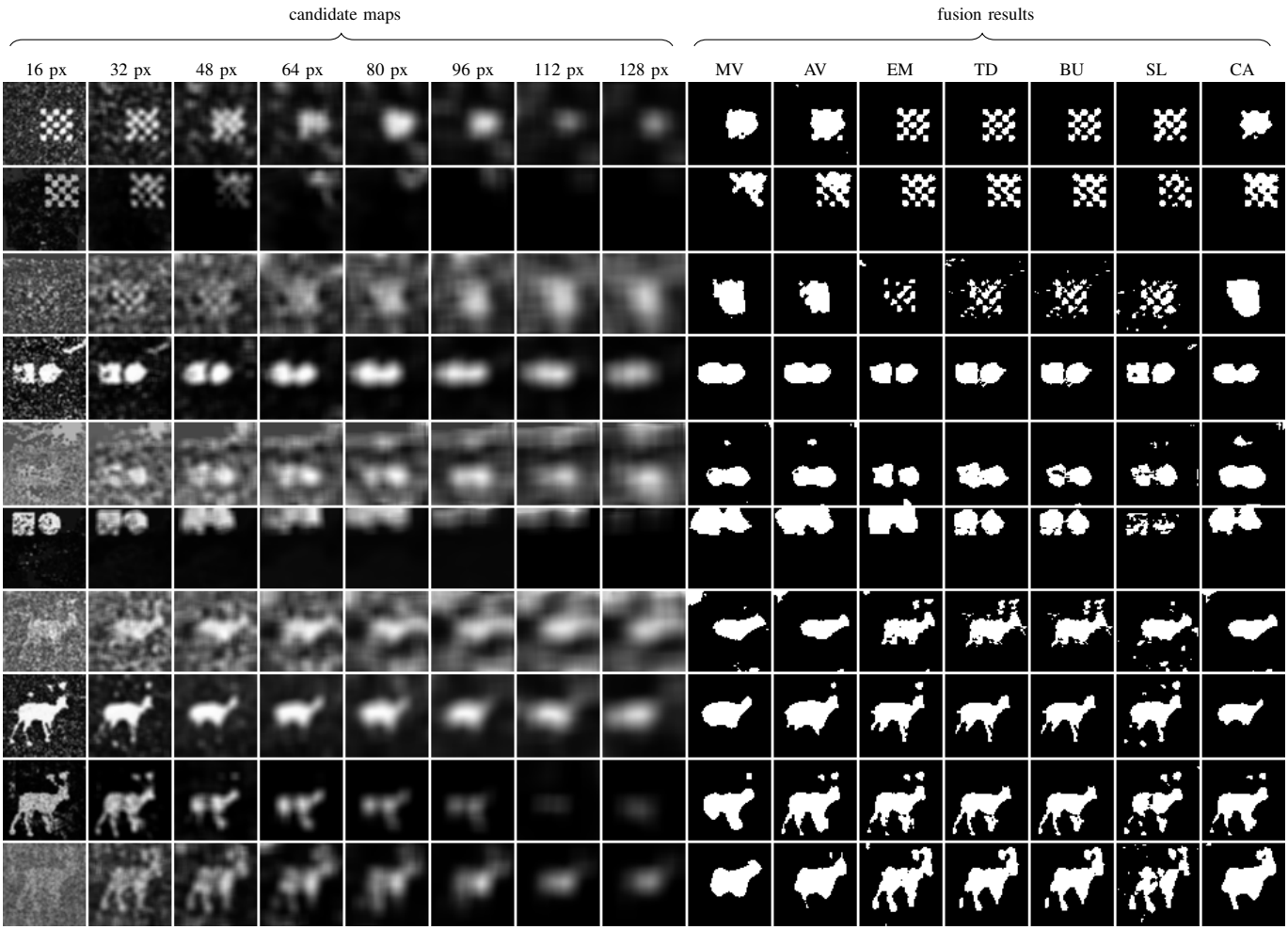



Fig. 15. Example results of multi-scale tampering map fusion along with the corresponding candidate maps for the *40 px tiles* (rows 1-3), *composite* (rows 4-6), and *moose* (rows 7-10) patterns; more examples are available in supplementary materials.

### *B. Limitations & Perspectives for Improvement*

Depending on the forensic feature at hand, different fusion methods might be appropriate. However, it remains critical to correctly exploit the dependencies between different scales of analysis. An attempt to learn them automatically, is successful to some extent, but is ultimately crippled by the lack of flexibility. While other fusion methods could easily reject unreliable candidate maps, the SVM would require separate training for every possible combination of valid inputs, the number of which grows very quickly (255 in this study). The noisy nature of the candidate maps makes it difficult to obtain reliable separation of the classes, and increasing the number of training examples may result in computational effort disproportional to performance improvement. Despite including neighborhood-related features, the decisions for every authentication unit are still independent. This issue could possibly be addressed by adopting Markov-like dependencies, e.g., like in discriminative random fields [51], but this issue requires separate future research and still does not address other shortcomings that we encountered in this study.

While evaluation on the realistic forgery dataset has shown that the proposed fusion methods can handle many tampered regions, further improvement in this respect is needed. If the tampered regions differ in size or shape considerably, it might be beneficial to adapt the parameters (e.g., neighborhood interactions in the EM fusion) to local image characteristics, or even to directly exploit image content (e.g., image segmentation results).

## VI. Conclusions and Future Work

In conclusion, the major contributions of our work include:

- a detailed analysis of the multi-scale fusion problem in the context of JPEG splicing forgeries; we have clearly shown that fusion of candidate maps obtained on multiple scales of analysis can improve the tampering localization performance of sliding window-based detectors by combining the benefits of small-scale and large-scale analysis;
- a novel multi-scale fusion technique based on energy minimization and threshold drift; the latter is a key component that allows to exploit the dependencies between different scales of analysis;
- two novel fusion techniques based on heuristic top-down and bottom-up refinement of an initial single-scale tampering map; the refinement follows simple rules cor-

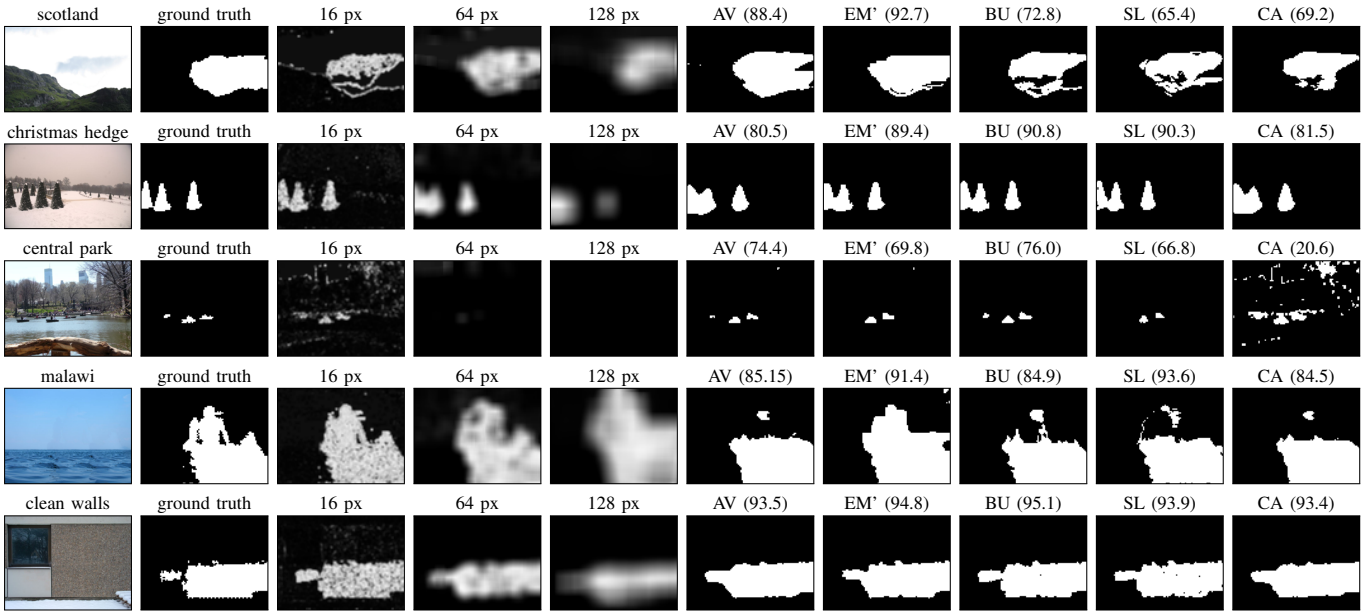



Fig. 16. Example fusion results for realistic forgeries; from left: tampered image, factual tampering locations, 3 selected candidate maps, 5 selected fusion results; numbers in brackets correspond to $F_1$ scores; more examples can be found in supplementary materials.

responding to the expected dependencies between small-scale and large-scale analysis.

In our future work, we will investigate the choice of the analysis windows for multi-scale fusion. Specifically, we will determine the preferred number of candidate maps, and the sliding window overlap that would guarantee good localization and computation performance. The latter will become particularly important when dealing with high-resolution images.

We will also investigate suitability of multi-scale fusion for other forensic features (e.g., PRNU or splicing detectors based on rich feature sets). Positive results would indicate feasibility of a combined multi-scale and multi-modal approach.

## REFERENCES


[1] P. Korus and A. Dziech, "Efficient method for content reconstruction with self-embedding," *IEEE Trans. on Image Processing*, vol. 22, no. 3, pp. 1134–1147, March 2013.

[2] P. Korus, J. Bialas, and A. Dziech, "Towards practical self-embedding for JPEG-compressed digital images," *IEEE Trans. on Multimedia*, vol. 17, no. 2, pp. 157–170, Feb 2015.

[3] S. Sarreshtedari and M.A. Akhaee, "A source-channel coding approach to digital image protection and self-recovery," *IEEE Trans. on Image Processing*, vol. 24, no. 7, pp. 2266–2277, July 2015.

[4] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, March 2009.

[5] M.C. Stamm, Min Wu, and K.J.R. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, 2013.

[6] I. Amerini, R. Becarelli, R. Caldelli, and A. Del Mastio, "Splicing forgeries localization through the use of first digit features," in *Proc. of IEEE Int. Workshop on Information Forensics and Security*, 2014.

[7] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "A Bayesian-MRF approach for PRNU-based image forgery detection," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 4, pp. 554–567, 2014.

[8] Y.-L. Chen and C.-T. Hsu, "What has been tampered? from a sparse manipulation perspective," in *Proc. of IEEE Int. Workshop on Multimedia Signal Processing*, Sept 2013, pp. 123–128.

[9] Bin Li, Tian-Tsong Ng, Xiaolong Li, Shunquan Tan, and Jiwu Huang, "Revealing the trace of high-quality jpeg compression through quantization noise analysis," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 3, pp. 558–573, March 2015.

[10] Z. Lin, J. He, X. Tang, and C.-K. Tang, "Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis," *Pattern Recognition*, vol. 42, no. 11, pp. 2492 – 2501, 2009.

[11] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Trans. on Information Forensics and Security,*, vol. 7, no. 3, pp. 1003–1017, June 2012.

[12] W. Wang, J. Dong, and T. Tan, "Exploring DCT coefficient quantization effects for local tampering detection," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 10, pp. 1653–1666, 2014.

[13] S. Duncan and S. Sameer, "Approaches to multisensor data fusion in target tracking: A survey," *IEEE Trans. on Knowledge and Data Engineering*, vol. 18, no. 12, pp. 1696–1710, 2006.

[14] E. F. Nakamura, A. F. Loureiro, and A. C. Frery, "Information fusion for wireless sensor networks: Methods, models, and classifications," *ACM Comput. Surv.*, vol. 39, no. 3, Sept. 2007.

[15] X. Wang, J.-H. Cho, K. Chan, M. Chang, A. Swami, and P. Mohapatra, "Trust and independence aware decision fusion in distributed networks," in *Proc. IEEE Int. Conf. on Pervasive Computing and Communications Workshops*, March 2013, pp. 481–486.

[16] Z. Chair and P.K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. on Aerospace and Electronic Systems*, vol. AES-22, no. 1, pp. 98–101, 1986.

[17] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[18] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[19] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, April 2012.

[20] L. Gaborini, P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, "Multi-clue image tampering localization," in *Proc. of IEEE Int. Workshop on Information Forensics and Security*, 2014, pp. 125–130.

[21] M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni, "A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence," *IEEE Trans. on Information Forensics and Security*, vol. 8, no. 4, pp. 593–607, 2013.

[22] M. Barni and A. Costanzo, "Dealing with uncertainty in image forensics: A fuzzy approach," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 1753–1756.

[23] D. Cozzolino, F. Gargiulo, C. Sansone, and L. Verdoliva, "Multiple classifier systems for image forgery detection," in *Image Analysis and Processing*, vol. 8157 of *Lecture Notes in Computer Science*, pp. 259–268. 2013.

[24] P. Ferrara, M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni, "Unsupervised fusion for forgery localization exploiting background

information," in *IEEE Int. Conf. on Multimedia & Expo Workshops*, 2015.
[25] P. Viola and M. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[26] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.
[27] M. Kraus and M. Strengert, "Depth-of-field rendering by pyramidal image processing.," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 645–654, 2007.
[28] C. Burger and S. Harmeling, "Improving denoising algorithms via a multi-scale meta-procedure," in *Pattern Recognition*, vol. 6835 of *Lecture Notes in Computer Science*, pp. 206–215. 2011.
[29] J. Sulam, B. Ophir, and M. Elad, "Image denoising through multi-scale learnt dictionaries," in *IEEE Int. Conf. on Image Processing*, 2014, pp. 808–812.
[30] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 79:1–79:11, July 2015.
[31] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, 2011.
[32] M. A. Garcia and D. Puig, "Supervised texture classification by integration of multiple texture methods and evaluation windows," *Image and Vision Computing*, vol. 25, no. 7, pp. 1091 – 1106, 2007.
[33] Y. Wang and C. Fan, "Single image defogging by multiscale depth fusion," *IEEE Trans. on Image Processing*, vol. 23, no. 11, pp. 4826–4837, 2014.
[34] C. Chamaret, J.C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in *IEEE Int. Conf. on Image Processing*, 2010, pp. 1077–1080.
[35] S.M. Muddamsetty, D. Sidibe, A. Tremeau, and F. Meriaudeau, "A performance evaluation of fusion techniques for spatio-temporal saliency detection in dynamic scenes," in *IEEE Int. Conf. on Image Processing*, Sept 2013, pp. 3924–3928.
[36] X. Cao, Z. Tao, B. Zhang, H. Fu, and X. Li, "Saliency map fusion based on rank-one constraint," in *Int. Conf. on Multimedia & Expo*, July 2013, pp. 1–6.
[37] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2012, pp. 438–445.
[38] B. Li, Y.Q. Shi, and J. Huang, "Detecting doubly compressed jpeg images by using mode based first digit features," in *Proc. IEEE Workshop on Multimedia Signal Processing*, Oct 2008, pp. 730–735.
[39] G. Schaefer and M. Stich, "Ucid - an uncompressed colour image database," in *Proc. SPIE Storage and Retrieval Methods and Applications for Multimedia*, 2004, pp. 472–480.
[40] "The dataset from the break our steganographic system contest," http://www.agents.cz/boss/index.php, 2010, Visited on 26 March 2015.
[41] H. H. Bauschke, C.H. Hamilton, M.S. Macklem, J.S. McMichael, and N.R. Swart, "Recompression of jpeg images by requantization," *IEEE Trans. on Image Processing*, vol. 12, no. 7, pp. 843–849, 2003.
[42] O. Gendler and M. Porat, "Toward optimal real-time transcoding using requantization in the DCT domain," in *IEEE Int. Conf. on Image Processing*, Nov 2009, pp. 3677–3680.
[43] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Trans. on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, 2008.
[44] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
[45] S. Z. Li, *Markov Random Field Modeling in Image Anlaysis*, Springer-Verlang, New York, 2001.
[46] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, Nov 2001.
[47] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.
[48] M. Schmidt, "UGM: A matlab toolbox for probabilistic undirected graphical models," http://www.cs.ubc.ca/~schmidtm/Software/UGM.html, 2011 version.
[49] G. Chierchia, D. Cozzolino, G. Poggi, C. Sansone, and L. Verdoliva, "Guided filtering for PRNU-based localization of small-size image forgeries," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2014, pp. 6231–6235.
[50] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An evaluation of popular copy-move forgery detection approaches," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 6, pp. 1841–1854, 2012.
[51] S. Kumar and M. Hebert, "Discriminative random fields: a discriminative framework for contextual interaction in classification," in *Proc. IEEE Int. Conf. on Computer Vision*, 2003, pp. 1150–1157 vol.2.

**Paweł Korus** (S'09-M'13) received his M.Sc. and Ph.D. degrees in telecommunications (both with honors) from the AGH University of Science and Technology in 2008, and in 2013, respectively. Since 2014 he has been an assistant professor with the Department of Telecommunications, AGH University of Science and Technology, Krakow, Poland. He is currently a postdoctoral researcher with the College of Information Engineering, Shenzhen University, Shenzhen, China.

His research interests include various aspects of multimedia security & image processing, with particular focus on digital image forensics, content authentication, digital watermarking & information hiding. In 2015 he received a scholarship for outstanding young scientists from the Polish Ministry of Science and Higher Education.

**Jiwu Huang** (M'98–SM'00-F'16) received the B.S. degree from Xidian University, Xi'an, China, in 1982, the M.S. degree from Tsinghua University, Beijing, China, in 1987, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 1998. He was with the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China. He is currently a Professor with the College of Information Engineering, Shenzhen University, Shenzhen, China.

His current research interests include multimedia forensics and security. He is also a member of the IEEE Circuits and Systems Society Multimedia Systems and Applications Technical Committee and the IEEE Signal Processing Society Information Forensics and Security Technical Committee. He served as an Associate Editor of the IEEE Transactions on Information Forensics and Security from 2010 to 2014. He was a General Co-Chair of the IEEE Workshop on Information Forensics and Security in 2013. He is a Fellow of IEEE.