A Multi-Stream Fusion Network for Image Splicing Localization

Maria Siopi, Giorgos Kordopatis-Zilos(✉), Polychronis Charitidis, Ioannis Kompatsiaris, and Symeon Papadopoulos

Information Technologies Institute, CERTH, Thessaloniki 60361, Greece
{siopi,georgekordopatis,charitidis,ikom,papadop}@iti.gr

Abstract. In this paper, we address the problem of image splicing localization with a multi-stream network architecture that processes the raw RGB image in parallel with other handcrafted forensic signals. Unlike previous methods that either use only the RGB images or stack several signals in a channel-wise manner, we propose an encoder-decoder architecture that consists of multiple encoder streams. Each stream is fed with either the tampered image or handcrafted signals and processes them separately to capture relevant information from each one independently. Finally, the extracted features from the multiple streams are fused in the bottleneck of the architecture and propagated to the decoder network that generates the output localization map. We experiment with two handcrafted algorithms, i.e., DCT and Splicebuster. Our proposed approach is benchmarked on three public forensics datasets, demonstrating competitive performance against several competing methods and achieving state-of-the-art results, e.g., 0.898 AUC on CASIA.

Keywords: image splicing localization · image forensics · multi-stream fusion network · late fusion deep learning

1 Introduction

Images have long been considered reliable evidence when corroborating facts. However, the latest advancements in the field of image editing and the wide availability of easy-to-use software create very big risks of image tampering by malicious actors. Moreover, the ability to easily alter the content and context of images especially in the context of social media applications further increases the potential use of images for disinformation. This is especially problematic as it has become almost impossible to distinguish between an authentic and tampered image by manual inspection.

To address the problem, researchers have put a lot of effort on the development of image forensics techniques that can automatically verify the authenticity of multimedia. Nevertheless, capturing discriminative features of tampered regions with multiple forgery types (e.g., splicing, copy-move, removal) is still an open challenge [14], especially in cases that the image in question is sourced from the Internet. Internet images have typically undergone many transformations (e.g. resizing,

D.-T. Dang-Nguyen et al. (Eds.): MMM 2023, LNCS 13834, pp. 611–622, 2023.
https://doi.org/10.1007/978-3-031-27818-1_50

recompression), which result in the loss of precious forensics traces that could lead to the localization of tampered areas [23]. This work focuses on the localization of splicing forgeries in images, where a foreign object from a different image is inserted in an original untampered one.

Several approaches have been proposed in the literature, both *handcrafted* and *deep-learning*, that attempt to tackle the problem of image splicing localization. Handcrafted approaches [3,16] aim at detecting the manipulations by applying carefully-designed filters that highlight traces in the frequency domain or by capturing odd noise patterns in images. On the other hand, the more recent deep learning approaches [1,9,22] leverage the advancements in the field and build deep networks, usually adapting encoder-decoder architectures, trained to detect the tampered areas in images based on large collections of forged images.

Although most splicing localization methods rely on handcrafted or deep learning schemes and work directly with the RGB images [1,3,9,16,18,25], there are only few works that combine the two solutions using a single network to fuse the information extracted from the raw image and/or several handcrafted signals [2,6,22]. The latter methods usually stack/concatenate all signals together in a multi-channel fashion and process them simultaneously by a single network stream that combines evidence from all signals to generate the output. This can be viewed as a kind of early fusion. Yet, some traces can be missed by the network when all signals are processed together. Instead, a late fusion approach could be employed, where each handcrafted signal along with the raw image is fed to a different network stream and then fused within the network to derive the output localization map. Each input signal is processed independently, and the network is able to capture the relevant information with different streams focused on a specific signal.

Motivated by the above, in this paper, we propose an approach that leverages the information captured from extracted handcrafted signals and the tampered images themselves. We develop a multi-stream deep learning architecture for late fusion following the encoder-decoder scheme. More specifically, the RGB images along with several handcrafted forensic signals, i.e., DCT [16] for a frequency-based and Splicebuster [3] for noise-based representation, which are robust to the localization of the tampered areas with splicing manipulation [2]. All signals are fed to different encoder streams that generate feature maps for each input signal. All extracted feature maps are then concatenated and propagated to a decoder network that fuses the extracted information and generates an output pixel-level map, indicating the spliced areas. By leveraging separate network streams for each input, we are able to extract richer features and, therefore, have more informative representations for the localization.

Our contributions can be summarized in the following:

- We address the splicing localization problem with a multi-stream fusion approach that combines handcrafted signals with the RGB images.
- We build an encoder-decoder architecture that processes each signal in a different encoder stream and fuses them during the decoding.
- We provide a comprehensive study on three public datasets, where the proposed approach achieves state-of-the-art performance.

2 Related Work

Image splicing localization has attracted the interest of many researchers in the last few decades; hence, several solutions have been proposed in the literature. The proposed methods can be roughly classified into two broad categories, i.e., handcrafted and deep learning.

Early image manipulation detection methods were designed to tackle the splicing localization problem using handcrafted methods, consisting of a simple feature extraction algorithm and can be categorized according to the signals they use and the compression type of the images they are applied on. For example, there are noise-based methods that analyse the noise patterns within images, e.g., Splicebuster [3] and WAVELET [17], methods that work with raw images analyzing them using different JPEG compression parameters and detect artifact inconsistencies, e.g., GHOST [5] and BLOCK [15], and double quantization-based algorithms operating in the frequency domain, e.g., Discrete Cosine Transform (DCT) [16]. An extensive review of such methods can be found in [23]. Nevertheless, the forensic traces captured by these algorithms can be easily erased by simple resizing and re-encoding operations. Also, these methods are often outperformed by their deep learning-based counterparts.

Later works use deep learning to localize splicing forgeries based on a neural network, extracting features only from the raw images. A seminal work in the field is ManTra-Net [22], which consists of two parts, a feature extractor and an anomaly detection network. The feature extractor computes features of the image by combining constrained CNNs, SRM features, and classical convolutional features concatenating them in a multi-channel fashion so as to be processed by the rest of the network. The detection network applies deep learning operations (LSTMs and CNNs) to the features extracted and exports the final localization map. SPAN [9] advanced ManTra-Net and modeled the spatial correlation between image regions via local self-attention and pyramid propagation. In [1], the model utilizes both the information from the frequency and the spatial domain of images. A CNN extracts the features in the spatial domain, while a Long Short-Term Memory (LSTM) layer receives the resampled features extracted from the image patches as input. The outputs of the two streams are fused into a decoder network that generates the final localization map. Mazaheri [18] added a skip connection to the above architecture, which exploits low-level features of the CNN and combines them with high-level ones in the decoder. In [21], the authors followed an encoder-decoder architecture and introduced a bidirectional LSTM layer and gram blocks. The method proposed in [25] combines top-down detection methods with a bottom-up segmentation-based model. In [8], the authors proposed a multi-scale network architecture based on Transformers, exploiting self-attention, positional embeddings, and a dense correction module. However, the above architectures utilize only the information that can be extracted from the raw images, which can be boosted with the use of handcrafted signals. Only ManTra-Net leverages some handcrafted features, which, however, are early fused with image features in a multi-channeled manner.

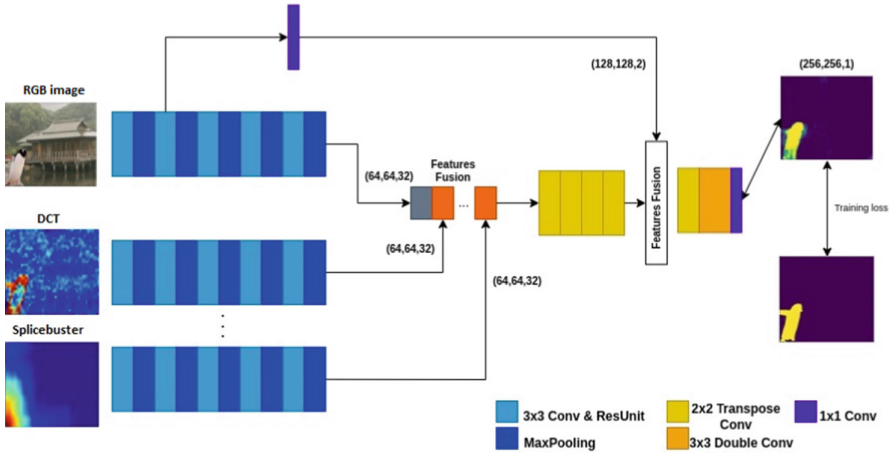

Fig. 1. Overview of the proposed network architecture (best viewed in colour). The inputs are the RGB image and the handcrafted signals, i.e., DCT and Splicebuster, processed by a different encoder stream. The encoding outputs are fused and propagated to the decoder that outputs the predicted localization map, which is compared to the ground-truth mask to compute the loss.

Finally, there are fusion approaches that leverage several handcrafted forensic signals [2,6,10] aiming to increase the robustness of the models. In [6], the authors employed the Dempster-Shafer theory of evidence [7] that allows handling uncertain predictions provided by several image forensics algorithms. In [10], the authors proposed a handcrafted approach that extracts several handcrafted signals that are further refined to generate the output map. In [2], the authors proposed a deep learning-based architecture based on an encoder-decoder scheme that receives several maps from handcrafted signals concatenated in a multi-channel way and processed by a single-stream network. The latter two works do not exploit the information from the raw images into the fusion process and do not rely on multi-stream processing of the signals.

3 Approach Overview

The main objective of our work is to develop a model that fuses different handcrafted forensic signals - in this work, we explore DCT [16] and Splicebuster [3] - along with the raw manipulated image. The model follows an encoder-decoder architecture. In the decoder, we fuse the outputs of multiple encoding streams, which extract features of the input signals through convolutional operations. Figure 1 illustrates an overview of the proposed architecture.

3.1 Multi-stream Architecture

Our architecture has two parts, i.e., an encoder and a decoder. For the encoder, we build multiple streams that process either the RGB images or the employed handcrafted signals. For each stream, we employ a network architecture similar to the one proposed in [18]. More specifically, each encoder stream comprises five stages consisting of a 3×3 convolutional layer, a residual block, and a max-pooling layer. The residual blocks are composed of two 3×3 convolutional layers with batch normalization [12] and a ReLU activation in the output. The number of channels of each stage output are [32, 64, 128, 256, 512]. At the end of each stream, we apply a 3×3 convolution with 32 output channels. Finally, we have two kinds of encoder streams, with and without the skip connection, as proposed in [18]. We use an encoder stream with the skip connection for the RGB image, while the remaining streams of the handcrafted signals do not have that skip connection. We empirically found that this setup yields the best performance.

For the decoder part, we first concatenate the feature maps of the encoder streams following a late-fusion approach, and we then process them by the main decoder network. The size of the decoder input depends on the number of encoder streams in the system. Unlike prior works [1,18], we build a more sophisticated architecture for our decoder, which performs upsampling in a learnable way by employing trainable transpose convolutional layers. We use as many transpose convolutional layers as the number of stages in our encoder streams, i.e., five layers with output channels [64, 32, 16, 2, 2]. Following the practice of [18], We add a skip connection from the second stage of the image encoder stream and concatenate it to the feature map of the fourth decoder layer. At the end of the decoder, we apply two 3×3 convolutions with a number of channels equal to 2. The output aggregated pixel-level predictions derive from the application of a final 1×1 convolution with a single output channel, followed by a sigmoid activation that maps the values to the $[0, 1]$ range.

In that way, we build a multi-stream architecture that encodes several signals independently, i.e., RGB images and handcrafted signals, and performs a late-fusion in the model's bottleneck. The extracted features are then processed altogether by a decoder network that outputs a binary mask with per-pixel predictions without the need for further post-processing.

3.2 Handcrafted Signals

In our approach, the number of the encoding streams equals the number of the handcrafted forensics signals used for the prediction, plus one for the manipulated image. Previous works tried to utilize the information from the frequency and the spatial domain of the processed images, combining schemes based on CNN and LSTM layers [1,18]. In contrast, to capture information from the frequency domain, we employ the DCT [16] handcrafted signal, which is a Fourier-based transform and can represent the JPEG images in the frequency domain. It divides the image into segments based on its resolution and applies the discrete

cosine transform whose coefficients contain information related to the frequencies in the segments. DCT has been widely used in many different applications, including forgery detection. In that way, the input images are represented in the frequency domain, and with the application of convolutional filters, we capture spatial information from the frequency domain.

Other works, e.g., ManTra-Net [22], extract noise-based handcrafted signals, combined with the RGB image in a multi-channel way for image tampering localization. To this end, in this work, we employ the Splicebuster [3] as a noise-based handcrafted signal, which is among the top-performing handcrafted algorithms. Splicebuster extracts a feature map for the whole image in three steps: (i) compute the residuals of the image through a high-pass filter, (ii) quantize the output residuals, and (iii) finally generate a histogram of co-occurrences to derive a single feature map. Similar to DCT, we generate noise-based representations for the input images, and with the application of convolutional filters, we extract spatial information from these representations.

3.3 Training Process

The signals are extracted using the publicly available service in [24]. During training, the RGB images and the extracted handcrafted signals are fed to the model, each in a different stream, and it outputs a binary map as a pixel-level prediction. The loss function used for the end-to-end training of the network is the binary cross-entropy loss computed based on the ground-truth masks and the generated outputs.

4 Experimental Setup

This section describes the datasets used for training and evaluation of our models, the implementation details, and the evaluation metrics used in our experiments to measure the splicing localization performance.

4.1 Datasets

For the training of our model, we use the Synthetic image manipulation dataset [1], and we extract the maps of our handcrafted signals for each image in the dataset. The synthetic dataset contains images with tampered areas from splicing techniques.

For evaluation, we used three image manipulation datasets, i.e., CASIA [4], IFS-TC [11] and Columbia [19]. The models are further fine-tuned to evaluation datasets. For CASIA, for the fine-tuning of our models we use CASIA2, which includes 5,123 tampered images, and for evaluation CASIA1, which includes 921 tampered images, i.e. we use only the subset with spliced images for evaluation. Regarding IFS-TC, we split the dataset to training and test set, and for fine-tuning we use the training set, which includes 264 tampered images, and for evaluation we use the test set, which includes 110 tampered images. Columbia [19] is a very small dataset (180 tampered images); hence, we use it in its entirety for evaluation without fine-tuning our model.

4.2 Implementation Details

All of the models have been implemented using PyTorch [20]. For the training of our model, we use Adam [13] as the optimization function with a 10^{-4} learning rate. The network is trained for 20 epochs, and we save the model parameters with the lowest loss in a validation set. Each batch contains 16 images along with the maps of the handcrafted signals and their ground-truth masks. We run our experiments on a Linux server with an Intel Xeon E5-2620v2 CPU, 128GB RAM, and an Nvidia GTX 1080 GPU.

4.3 Evaluation Metrics

We use the pixel-level Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) as our primary metric to capture the model's performance and for comparison against the state-of-the-art splicing localization methods.

5 Experiments and Results

In this section, we provide an ablation study for our proposed method under various configurations (Sect. 5.1), the comparison against state-of-the-art methods (Sect. 5.2), and some qualitative results (Sect. 5.3).

5.1 Ablation Study

Impact of Each Handcrafted Signal. First, we examine the impact of each handcrafted signal, separately and combined, fused with our Multi-Stream (MS) scheme and baseline Multi-Channel (MC) approach, where the signals are concatenated along the image channels. For the MS runs, we have two streams where one handcrafted signal is used and three streams when all inputs are combined. For the MC runs, the network has a single stream in all cases. Table 1 illustrates the results on the three evaluation datasets for several handcrafted signal combinations using different fusion schemes. In general, using the MS fusion scheme leads to better results than using MC for the majority of the handcrafted signals and datasets. RGB+SB with MS consistently achieves very high performance, being among the top ranks in all datasets. It outperforms its MC counterpart, achieving significantly better results on the IFS-TC dataset. Additionally, RGB+DCT+SB with MS outperforms the corresponding run with MC in all datasets, highlighting that fusing multiple signals using MS leads to better accuracy. MS-DCT reports improved performance compared to the MC-DCT on two datasets, but it is worse than the other two configurations. Additionally, combining handcrafted signals with the RGB images improves performance in general, especially in the IFS-TC dataset. Finally, DCT and SB achieve competitive performance in two datasets. This indicates that they capture useful information, which our MS architecture exploits to further improve results.

Table 1. Performance of our method with three signal and two fusion schemes on CASIA, IFS-TC, and Columbia. MS and MC stand for Multi-Stream and Multi-Channel processing, respectively.

Signals	Fus.	CASIA	IFS-TC	Columbia
DCT	-	0.743	0.646	0.640
SB	-	0.689	0.750	0.830
RGB	-	0.877	0.614	0.818
RGB+DCT	MC	0.866	0.732	0.688
	MS	0.873	0.689	0.777
RGB+SB	MC	0.869	0.679	0.855
	MS	0.898	0.776	0.836
RGB+DCT+SB	MC	0.851	0.721	0.717
	MS	0.873	0.759	0.782

Table 2. Performance of our method with three signals with and without fine-tuning on CASIA, IFS-TC, and Columbia. Note that we do not fine-tune our model on Columbia due to its small size.

Signals	FT	CASIA	IFS-TC	Columbia
RGB	✗	0.765	0.470	0.818
	✓	0.877	0.614	-
RGB+DCT	✗	0.868	0.507	0.777
	✓	0.873	0.689	-
RGB+SB	✗	0.753	0.460	0.836
	✓	0.898	0.776	-
RGB+DCT+SB	✗	0.887	0.497	0.782
	✓	0.873	0.759	-

Impact of Fine-Tuning. Furthermore, we benchmark the performance of the proposed multi-stream approach with and without fine-tuning on the evaluation datasets. Table 2 displays the results of our method on the three evaluation datasets when using the pre-trained and fine-tuned versions. Keep in mind that we do not fine-tune for the Columbia dataset. It is noteworthy that there is substantial performance gain in almost all cases where fine-tuning is applied. A reasonable explanation is that, with the fine-tuning on the evaluation datasets, the network learns to capture the information from the handcrafted features based on the specific domain expressed by each dataset. Therefore, the extracted cues from the employed handcrafted features might not be generalizable across different datasets. We might improve the performance further on the Columbia dataset if we could fine-tune our model on a dataset from a similar domain.

Impact of Skip Connections. Additionally, we benchmark the performance of the proposed multi-stream approach with different configurations for the skip

Table 3. Performance of our method with three signals and three configurations for the skip connection on CASIA, IFS-TC, and Columbia. *No* indicates that no skip connections are used. *Img* indicates that skip connection is used only for the image stream. *All* indicates that skip connections are used only for all streams.

Signals	Skip	CASIA	IFS-TC	Columbia
RGB+DCT	***No***	0.857	0.732	0.623
	Img	0.873	0.689	0.777
	All	0.840	0.616	0.742
RGB+SB	***No***	0.879	0.773	0.741
	Img	0.898	0.776	0.836
	All	0.882	0.718	0.826
RGB+DCT+SB	***No***	0.797	0.763	0.566
	Img	0.873	0.759	0.782
	All	0.871	0.808	0.762

connection. Table 3 displays the results of our method on the three evaluation datasets using no, image-only and all-streams skip connections. It is noteworthy that the methods perform very robustly when a skip connection is used in the image stream only. It achieves the best AUC in all cases, except for RGB+DCT+SB in the IFS-TC dataset. Finally, the experiments with no use of skip connections lead to the worst results, indicating that, thanks to the skip connections, the network learns to successfully propagate useful information from the encoder streams to the decoder. Yet, skip connections from the handcrafted signals do not always help.

5.2 Comparison with the State-of-the-Art

In Table 4, we present our evaluation in comparison to four state-of-the-art approaches. We select our networks with MS fusion and with skip connection only to the image stream, denoted as MS-DCT, MS-SB, and MS-DCT+SB for the three signal combinations. As state-of-the-art approaches, we have re-implemented three methods, LSTMEnDec [1], LSTMEnDecSkip [18], and OwAF [2], using the same training pipeline as the one for the development of our networks for fair comparison. These are closely related methods to the proposed one. Also, we benchmark against the publicly available PyTorch implementation of ManTra-Net [22][1] without fine-tuning it on the evaluation datasets. All methods are benchmarked on the same evaluation sets. In general, all three variants of our method achieve competitive performance on all evaluation datasets, outperforming the state-of-the-art approaches in several cases with a significant margin. Our MS-SB leads to the best results with 0.898 AUC, respectively, with the second-best LSTMEnDecSkip approach achieving 0.810. Similar results are reported on the IFS-TC dataset. Our MS-SB achieves the best AUC with 0.776,

[1] https://github.com/RonyAbecidan/ManTraNet-pytorch.

Table 4. Performance comparison against the state-of-the-art on CASIA, IFS-TC, and Columbia.

Method	CASIA	IFS-TC	Columbia
ManTra-Net [22]	0.665	0.547	0.660
LSTMEnDec [1]	0.628	0.648	0.809
LSTMEnDecSkip [18]	0.810	0.670	0.207
OwAF [2]	0.754	0.680	0.551
MS-DCT (Ours)	0.873	0.689	0.777
MS-SB (Ours)	**0.898**	**0.776**	**0.836**
MS-DCT+SB (Ours)	0.873	0.759	0.782

followed by the OwAF method with 0.680. Finally, our MS-SB achieves the best results in the Columbia dataset with 0.836. Notably, the LSTMEnDec is the second-best approach, outperforming our two other variants, MS-DCT and MS-DCT+SB; however, this method performs poorly on the other two datasets.

5.3 Qualitative Results

Figure 2 illustrates some example results from the IFS-TC dataset. The first three columns contain the network inputs, i.e., the RGB image, DCT, and Splicebuster. The third column presents the ground truth masks, and the last ones depict the network predictions. In the first example, Splicebuster provides a useful lead to the network, which is able to detect the tampered area with high accuracy, especially the MS-SB run. In the second case, all of our networks detect

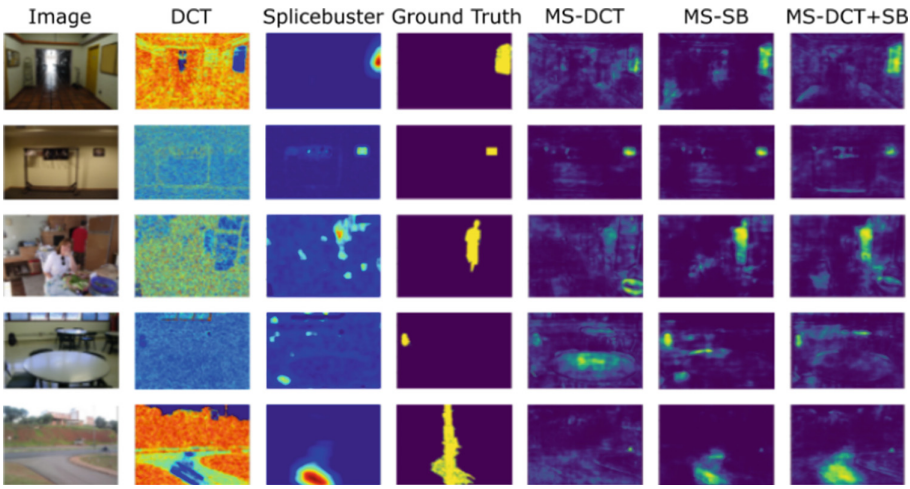

Fig. 2. Visual examples of our multi-stream network with three signal combinations from the IFS-TC dataset.

the tampered area, even though DCT does not seem to be helpful. In the next two examples, none of the handcrafted signals precisely localize splicing, but our MS-SB and MS-DCT+SB are able to detect it partially. Finally, in the last case, our networks failed to localize the forged areas in the image, although the two handcrafted signals highlight the correct area only in a small part. In general, the qualitative results here align with the quantitative of the previous sections, with MS-DCT providing the worst predictions among our three settings, while MS-SB detects the tampered areas with significantly higher accuracy.

6 Conclusion

In this work, we proposed a deep learning method that localizes spliced regions in images by fusing features extracted from the RGB images with ones extracted from handcrafted signals based on a multi-stream fusion pipeline. We experimented with two popular handcrafted signals based on DCT and Splicebuster algorithms. Through an ablation study on three datasets, we demonstrated that our multi-stream fusion approach yields competitive performance consistently. Also, we compared our approach to four state-of-the-art methods, achieving the best performance on all three datasets. In the future, we plan to investigate more architectural choices that improve the effectiveness of signal fusion and employ more robust handcrafted signals.

Acknowledgments:. This research has been supported by the H2020 MediaVerse and Horizon Europe vera.ai projects, which are funded by the European Union under contract numbers 957252 and 101070093.

References

1. Bappy, J.H., Simons, C., Nataraj, L., Manjunath, B., Roy-Chowdhury, A.K.: Hybrid LSTM and encoder-decoder architecture for detection of image forgeries. IEEE Trans. Image Process. (2019)
2. Charitidis, P., Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, I.: Operation-wise attention network for tampering localization fusion. In: International Conference on Content-based Multimedia Indexing (2021)
3. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: a new blind image splicing detector. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS) (2015)
4. Dong, J., Wang, W., Tan, T.: CASIA image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing (2013)
5. Farid, H.: Exposing digital forgeries from jpeg ghosts. IEEE Trans. Inf. Forensics Secur. **4**(1), 154–160 (2009)
6. Fontani, M., Bianchi, T., De Rosa, A., Piva, A., Barni, M.: A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence. IEEE Trans. Inf. Forensics Secur. **8**(4), 593–607 (2013)
7. Gordon, J., Shortliffe, E.H.: The Dempster-Shafer theory of evidence. Rule-Based Exp. Syst.: MYCIN Exp. Stanford Heuristic Program. Proj. **3**, 832–838 (1984)

8. Hao, J., Zhang, Z., Yang, S., Xie, D., Pu, S.: Transforensics: image forgery localization with dense self-attention. In: IEEE/CVF International Conference on Computer Vision (2021)
9. Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., Nevatia, R.: SPAN: spatial pyramid attention network for image manipulation localization. In: European Conference on Computer Vision (2020)
10. Iakovidou, C., Papadopoulos, S., Kompatsiaris, Y.: Knowledge-based fusion for image tampering localization. In: IFIP International Conference on Artificial Intelligence Applications and Innovations (2020)
11. IFS-TC: Report on the IEEE-IFS challenge (2016). http://ifc.recod.ic.unicamp.br/
12. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (2015)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Korus, P.: Digital image integrity-a survey of protection and verification techniques. Digit. Signal Proces. **71**, 1–26 (2017)
15. Li, W., Yuan, Y., Yu, N.: Passive detection of doctored jpeg image via block artifact grid extraction. Signal Process. **89**(9), 1821–1829 (2009)
16. Lin, Z., He, J., Tang, X., Tang, C.K.: Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. Pattern Recogn. **42**(11), 2492–2501 (2009)
17. Mahdian, B., Saic, S.: Using noise inconsistencies for blind image forensics. Image Vis. Comput. **27**(10), 1497–1503 (2009)
18. Mazaheri, G., Mithun, N.C., Bappy, J.H., Roy-Chowdhury, A.K.: A skip connection architecture for localization of image manipulations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 119–129 (2019)
19. Ng, T.T., Hsu, J., Chang, S.F.: Columbia image splicing detection evaluation dataset. Columbia Univ CalPhotos Digit Libr, DVMM lab (2009)
20. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the International Conference on Neural Information Processing Systems (2019)
21. Shi, Z., Shen, X., Chen, H., Lyu, Y.: Global semantic consistency network for image manipulation detection. IEEE Signal Process. Lett. **27**, 1755–1759 (2020)
22. Wu, Y., AbdAlmageed, W., Natarajan, P.: ManTra-Net: manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
23. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Large-scale evaluation of splicing localization algorithms for web images. Multimed. Tools Appl. **76**(4), 4801–4834 (2016). https://doi.org/10.1007/s11042-016-3795-2
24. Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y., Bouwmeester, R., Spangenberg, J.: Web and social media image forensics for news professionals. In: Proceedings of the International AAAI Conference on Web and Social Media (2016)
25. Zhang, Y., Zhang, J., Xu, S.: A hybrid convolutional architecture for accurate image manipulation localization at the pixel-level. Multimed. Tools Appl. **80**(15), 23377–23392 (2021). https://doi.org/10.1007/s11042-020-10211-1