Exploring Multi-Modal Fusion for Image Manipulation Detection and Localization

Konstantinos Triaridis * and Vasileios Mezaris

Information Technologies Institute, Centre for Research & Technology Hellas,
Thessaloniki, Greece
{triaridis,bmezaris}@iti.gr

Abstract. Recent image manipulation localization and detection techniques usually leverage forensic artifacts and traces that are produced by a noise-sensitive filter, such as SRM and Bayar convolution. In this paper, we showcase that different filters commonly used in such approaches excel at unveiling different types of manipulations and provide complementary forensic traces. Thus, we explore ways of merging the outputs of such filters and aim to leverage the complementary nature of the artifacts produced to perform image manipulation localization and detection (IMLD). We propose two distinct methods: one that produces independent features from each forensic filter and then fuses them (this is referred to as late fusion) and one that performs early mixing of different modal outputs and produces early combined features (this is referred to as early fusion). We demonstrate that both approaches achieve competitive performance for both image manipulation localization and detection, outperforming state-of-the-art models across several datasets[1].

Keywords: Image forensics · Image manipulation localization · Image manipulation detection · Multimodal fusion

1 Introduction

Editing and manipulating digital media has gotten increasingly easier and more accessible in recent years. Recent advances in image editing software, as well as deep generative models such as Generative Adversarial Networks (GANs) [15,36] and diffusion models [20, 32], facilitate producing manipulations that are often imperceptible to the human eye and are widely available, even to potentially malicious users. The widespread use of smartphones and social networks also enables the spread of such manipulated media at a rapid pace. As a result, such edited images can cause social problems when used as evidence to support disinformation campaigns and stories or mislead the public by obfuscating important content from news, resulting in diminished trust. Therefore, techniques for image manipulation detection and localization, as part of complete toolboxes for media verification such as [24], are now needed more than ever.

[*] Corresponding author

[1] Code is publicly available at https://github.com/IDT-ITI/MMFusion-IML

Image forgery localization and detection are tasks the media forensics field has been working on for many years. Early works typically focused on a specific type of manipulation such as splicing [23], copy-move [3] or removal/inpainting [17]. More recently, deep-learning-based solutions of increasing robustness are proposed that are able to recognize multiple different types of manipulations [2,9,11,14,18,26,29,33]. In order to be able to perform manipulation localization in a semantic-agnostic manner these models need to suppress image contect to reveal forensic artifacts. Most approaches achieve this by applying a high-pass filter to extract noise maps [2, 11, 26, 29, 33]. The most popular high-pass filters used are the ones proposed in the Steganalysis Rich Model (SRM) [8], utilized in a wide variety of works [11, 16, 19, 29, 37], while the Bayar convolution [1] is also used in a multitude of approaches [2, 11, 29, 33] and NoisePrint++ is used in a more recent model [9].

We hypothesize that those different forensic filters actually produce artifacts of complementary forensic capabilities. NoisePrint [4], and its successor NoisePrint++ [9] produce artifacts that relate to camera model and editing history, thus displaying limited performance for copy-move images (Section 4.3). On the other hand, SRM [8] filters can identify edges and boundaries without relying on camera or compression/editing artifacts, but their predetermined nature makes them vulnerable to adversarial attacks, whereas the bayar convolution [1] learns the manipulation traces directly from data, proving more robust against malicious attacks. In this work we explore ways to expand existing state-of-the-art IMLD approaches to support multiple auxiliary forensic filters as inputs. We start with TruFor [9] as our baseline and propose utilizing NoisePrint++, SRM, and Bayar convolution as inputs auxiliary to the RGB image. We propose two different approaches: a late-fusion paradigm that extracts features from each modality separately, and an early-fusion paradigm that mixes the multi-modal features by early convolutional blocks. Our main contributions in this paper are summarized as follows:

- We compare the efficacy of different forensic filters, namely SRM, Bayar conv and NoisePrint, as inputs for deep networks performing forgery localization.
- We propose two distinct approaches for combining the outputs of different forensic filters for the purpose of image manipulation localization and detection.
- Both methods achieve state-of-the-art performance across five datasets by effectively leveraging and combining forensic cues from various input modalities.

2 Related Work

Image forensics methods have been based for a long time on detecting inconsistensies on low-level semantic-agnostic artifacts such as compression or internal camera filter artifacts. These artifacts are usually revealed by suppressing the image content through high-pass filtering, producing a noise-sensitive view. In recent times, various filters for noise extraction have been integrated into deep learning models to address the challenge of image manipulation localization. RGB-N [37] is a model that uses both RGB images and SRM filters together with a faster R-CNN [22] module to perform forgery detection with a bounding box, while Constrained R-CNN [33] uses a trainable noise extractor, namely bayar convolution, to perform the same task. ManTraNet [29] integrates both SRM filters and bayar convolution within a VGG-based architecture, while SPAN [11] enhances this approach by modeling relationships between image patches through a pyramid of local self-attention blocks. Chen et al. [2] also use the bayar convolution together with the RGB image in a late fusion paradigm that is trained through multi-scale supervision. NoisePrint [4] is a noise extractor proposed by Cozzolino et al. that is trained in a self-supervised manner to extract camera-specific artifacts and is expanded in TruFor [9], where it is used jointly with RGB images in a dual-branch CMX [34] architecture.

Our approach innovatively explores various strategies for combining the outputs of diverse noise extractors, leveraging their complementary capabilities to develop a robust end-to-end image forgery detection localization model.

3 Methods

3.1 Encoder-Decoder Architecture

Our goal is to extend existing encoder-decoder-based architectures to be able to use multiple forensic filters (SRM [8], bayar convolution [1], NoisePrint++ [9]) in tandem, so as to produce more robust representations for the task of Image Manipulation Localization and Detection. To this end we adopt the TruFor [9] architecture, showcased in Fig. 1, that consists of an encoder, an anomaly decoder, a confidence decoder, and a forgery detector; and we follow its two-phase training regime for anomaly localization and detection, respectively. The encoder follows the dual-branch architecture proposed in [34], comprising of 4 stages of Multi-Head Self Attention (MHSA) blocks [31] that produce feature maps f_r^i of different scales: $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$, where H and W are the spatial dimensions of the input and C_i is the channel dimension of the output at scale i. The two MHSA blocks' outputs in each stage are rectified through a Cross-Modal Feature Rectification Module (FRM) [34] that exploits the interactions between the two input modalities (RGB and NoisePrint in the case of TruFor). The FRM uses features from both modalities to produce weighted channel- and spatial-wise feature maps that are residually added for both modalities to perform channel- and spatial-wise rectification. The two sets of feature maps are then combined using a Feature Fusion Module (FFM) [34] whose outputs f^i at each scale are combined to produce the encoder output f. The FFM consists of an information exchange stage, where a cross-attention mechanism exchanges information between modalities and produces two sets of mixed feature maps, and a fusion stage where the feature maps are merged into a single output through a residual MLP module that uses 1×1 convolutions. The Decoders are simple MLP decoders proposed in [31].

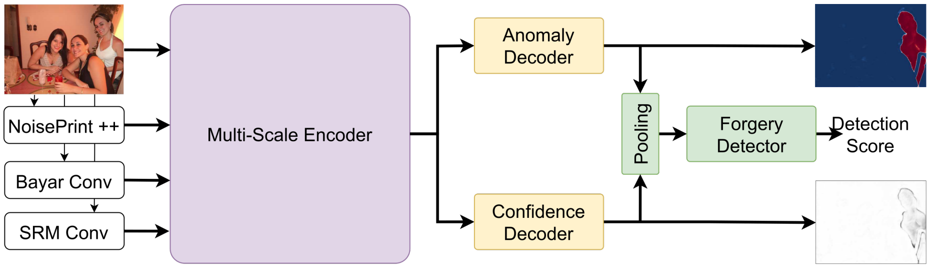

Fig. 1. Full encoder-decoder architecture

Utilizing this architecture one can combine RGB images with an auxiliary forensic modality to perform Image Manipulation Localization. In [9] Guillaro et al. use their own feature extractor NoisePrint++, however a multitude of other forensic filters' outputs like bayar convolution [1] or SRM [8] can be utilized. All those filters are analyzed in Section 3.2. We propose two different ways of extending the encoder architecture to multiple auxiliary modal inputs: a late fusion paradigm, where each auxiliary modality is combined with RGB inputs separately using a dual-branch architecture [34] (Section 3.3), and an early fusion paradigm where auxiliary modalities are combined early before being utilized as input to the dual-branch encoder together with the RGB inputs (Section 3.4).

3.2 Auxiliary modalities

For both approaches, we use the outputs of three forensic filters, namely NoisePrint++, SRM, and Bayar convolution as inputs together with RGB images.

NoisePrint++ In [4] Cozzolino et al. propose Noiseprint, a CNN-based model designed to extract camera-model-based artifacts from RGB images while suppressing image content. In [9] they expand their approach, namely NoisePrint++, to be able to recognize and extract artifacts related to the editing history of an image (e.g. compression, resizing, gamma correction). NoisePrint++ is trained in a supervised contrastive manner [12]: a batch of images is provided, from which patches are extracted from different locations. Then the patches go through different editing pipelines. Patches extracted from the same source image, the same location, and with the same editing history are considered positive samples, while others are considered negative. For our approach we use NoisePrint++ as a pretrained feature extractor.

SRM Another way to suppress the image content and highlight forensic traces and noise is through static high-pass filters, the most common of which are the ones proposed for producing residual maps for the Steganalysis Rich Model (SRM) [8]. Out of the 30 high-pass filters proposed, we used the 3 most commonly used in literature [11, 29, 37] which will be referred to as SRM filters.

Bayar Convolution In contrast to using static high-pass filters for noise extraction Bayar et al. [1] propose the constrained convolutional layer as a noise extractor that adaptively learns manipulation traces from data. We use the constrained convolutional layer as an extra noise feature extractor and refer to it as Bayar convolution. For both modal fusion approaches the Bayar convolutional layer is pretrained alone in a dual branch CMX encoder (Section 4.3) and then used with its weights frozen.

3.3 Late Fusion

For the late fusion method first we extract the auxiliary representations $r_{noiseprint}$, r_{srm}, r_{bayar} of the RGB image x from the NoisePrint++, SRM and bayar filters respectively. Then the output of each auxiliary filter is fed together with the original RGB input into a dual-branch CMX encoder to produce 4-scale feature maps $f^i_{mod} = \mathcal{E}_{mod}(x, r_{mod}), mod \in \{noiseprint, srm, bayar\}, i \in \{1, 2, 3, 4\}$ as shown in Fig. 2.

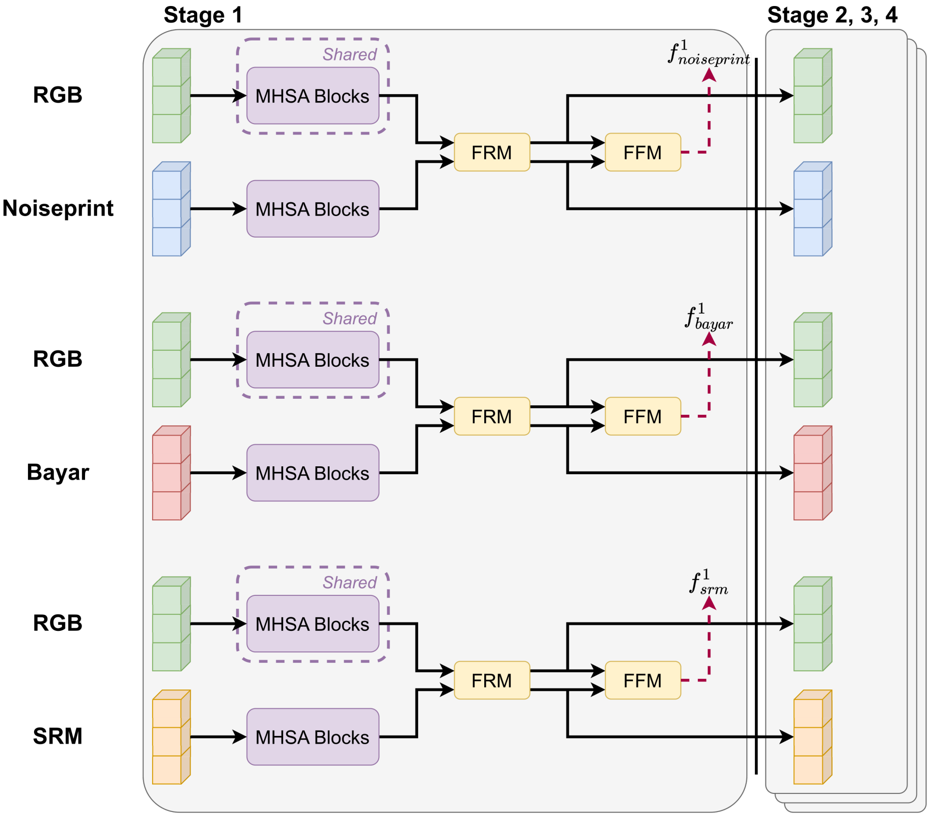

Fig. 2. Late Fusion with weight sharing

At each scale the outputs of the 3 encoders are concatenated to produce the final output f of the encoder. We use the same decoder architecture as in TruFor for the anomaly and confidence decoders. Like other multi-modal approaches this approach is prone to overfitting and the "modality imbalance" problem [7, 27], where different modalities converge and overfit at different rates, thus hindering joint optimization. To tackle this we make the weights of the modules along the RGB branch shared across all 3 encoders to increase regularization. We also employ Dropout before the anomaly decoder as the complete encoder is rather large and the simple MLP decoder is prone to overfitting.

3.4 Fusion by early convolutions

For the early fusion method we again extract the auxiliary representations $r_{noiseprint}$, r_{srm}, r_{bayar} of the RGB image x. Then each input is passed through a convolutional block $\mathcal{C}$ to produce early features $f^{ef}_{mod} = \mathcal{C}(r_{mod}), mod \in \{noiseprint, srm, bayar\}$. The 3 sets of feature maps are then concatenated to produce the complete set of early features f^{ef}. Those features then pass through another convolutional block $\mathcal{C}$ to produce mixed features $f^{mf} = \mathcal{C}(f^{ef})$. The convolutional blocks are good at early visual processing, resulting in a more stable optimization [30], thus aiding in mixing the features from different modalities smoothly. The mixed features f^{mf} and RGB image x are used as input for a dual-branch CMX encoder [34], in the same manner as in TruFor. This is a particularly lightweight approach to expanding the TruFor architecture to handle multiple auxiliary modalities as it does not increase the number of parameters significantly (68.9M params compared to TruFor's 68.7M).

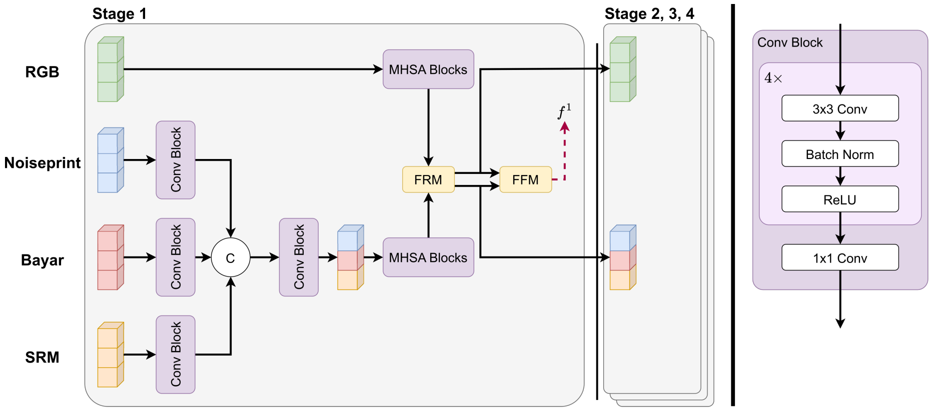

Fig. 3. Fusion by early convolutions

Convolutional Block The convolutional block consists of four 3×3 convolutions followed by a 1×1 convolutional layer to resize the output to 3 channels.

There is a batch normalization (BN) and a ReLU layer after each 3×3 convolutional layer. The output channels for the 3×3 convolutional layers are [24, 48, 96, 192].

4 Experiments

4.1 Experimental Setup

Training We follow the training procedure proposed by Guillaro et al. [9]: first, we jointly train the encoder and anomaly decoder and finally, we train the confidence decoder and the forgery detector, while the encoder and anomaly decoder are kept frozen. For both training phases we use the datasets used by Kwon et al. [14], and sample an equal number of images from each one for every epoch. Training datasets are summarized in Table 1.

	Number of Images	
Dataset	Real	Fake
Casiav2 [6]	7,491	5,105
IMD2020 [21]	414	2,010
FantasticReality [13]	16,592	19,423
cm_coco [14]	-	200,000
bcm_coco [14]	-	200,000
bcmc_coco [14]	-	200,000
sp_coco [14]	-	200,000

Table 1. Details for training datasets

	Number of Images	
Dataset	Real	Fake
Coverage [28]	100	100
Columbia [10]	183	180
Casiav1+ [6]	800	921
DSO-1 [5]	100	100
CocoGlide [9]	512	512

Table 2. Details for testing datasets

Testing For testing, we evaluate our model on five datasets: Coverage [28], Columbia [10], Casiav1+[2] [6] and DSO-1 [5], which are widely used in the relevant literature, and CocoGlide, a diffusion-based manipulation dataset proposed recently by Guillaro et al [9].

Metrics For localization performance we follow most previous work and report average pixel-level performance using the F1 metric. We use a fixed threshold of 0.5, as setting a best threshold per test dataset [14] or even per image [9] like some previous works is not realistic in practical scenarios where the ground truth is not available, thus leading in exaggerated performance estimates. For detection we use image-level Area Under Curve (AUC), which is a metric that does not require selecting a threshold, and balanced accuracy, the arithmetic mean of sensitivity and specificity, with a threshold once again set to 0.5.

[2] Casiav1+ is a modification of the Casiav1 dataset proposed by Chen et al. [2] that replaces authentic images that also exist in Casiav2 with images from the COREL [25] dataset to avoid data contamination.

Implementation All models are implemented in PyTorch and trained on an NVIDIA RTX 4090 GPU, using an effective batch size of 24 for 100 epochs. Physical batch size ranged from 4 to 8 depending on the model and an effective batch size of 24 was reached by utilizing gradient accumulation. The MHSA modules were initialized with ImageNet-pretrained weights as proposed in [34, 35]. We utilized an SGD optimizer with an initial learning rate of 0.005, momentum of 0.9, weight decay of 0.0005 and a polynomial learning rate schedule. For training augmentations we followed Guillaro et al. [9] and resized the images in the [0.5-1.5] range, performed random cropping of size 512×512 and JPEG compression with a random Quality Factor QF∈[30,100].

4.2 Comparisons

We compare our methods with recent approaches for Image Manipulation Localization. Following Guillaro et al. we consider methods with open source models provided and we exclude models that use part of our testing datasets for training to avoid bias. Overall we compare with TruFor [9], CAT-Netv2 [14], ManTraNet [29], PSCC-Net [18], SPAN [11], Constrained R-CNN [33], MVSS-Net [2]. Results are presented in Table 3.

Model	Coverage	Columbia	Casiav1+	CocoGlide	DSO-1	AVG
TruFor	.600	.859	.737	.523	**.930**	.729
CAT-Netv2	.381	.859	.752	.434	.584	.602
ManTraNet	.317	.508	.180	.516	.412	.387
PSCC-Net	.473	.604	.520	.515	.458	.514
SPAN	.235	.759	.112	.298	.233	.327
CR-CNN	.391	.631	.481	.447	.289	.448
MVSS-Net	.514	.729	.528	.486	.358	.523
Early Fusion	**.663**	**.888**	**.784**	.553	.863	.750
Late Fusion	.641	.864	.775	**.574**	.899	**.751**

Table 3. Comparison for localization performance. The metric is average pixel-level F1. The best and second-best results for each dataset are presented in bold and underlined respectively. Results for all models except for the proposed ones are taken from [9].

During our experiments, we replicated the training of TruFor for the purposes of our ablation study and we discovered a large variance in Localization results between training runs. For this purpose, we train our networks 4 times and report average localization performance in terms of average pixel F1 in Table 4. Both our multi-modal fusion approaches showcase state-of-the-art performance, being either the best or second-best model for every dataset. Especially for the Coverage dataset that contains only copy-move forgeries, our best approach surpasses the previous best, TruFor, by 6.3%. The only dataset where we can't achieve state-of-the-art performance is DSO-1 where our best method is 3% behind TruFor.

Model	Coverage	Columbia	Casiav1+	CocoGlide	DSO-1	AVG
TruFor (retrained)	.577(±.019)	.884(±.019)	.761(±.011)	.516(±.008)	.895(±.017)	.726(±.008)
Early Fusion	**.663**(±.011)	**.888**(±.014)	**.784**(±.001)	.553(±.015)	.863(±.025)	.750(±.005)
Late Fusion	.641(±.014)	.864(±.023)	.775(±.008)	**.574**(±.020)	**.899**(±.010)	**.751**(±.003)

Table 4. Comparison for localization performance for models with multiple training runs. Metric is average pixel-level F1 (± standard deviation)

We also compare across models in terms of detection performance and present the results in Table 5. Notably, our early fusion method demonstrates exceptional performance, surpassing the state-of-the-art on average. Particularly noteworthy is its outstanding performance on the Coverage dataset, where it achieves a remarkable improvement of nearly 7% in terms of the Area Under the Curve (AUC) and 9% in terms of balanced accuracy (bAcc) compared to the prior leading method. Our late fusion approach also exhibits competitive AUC performance, but falls slightly behind the TruFor model in terms of bAcc. This disparity in bAcc performance could potentially be attributed to the size of our late fusion model, which may be susceptible to overfitting. Further investigation and experimentation are warranted to explore the possibility of requiring additional regularization techniques to optimize its performance for the detection task.

	Coverage		Columbia		Casiav1+		CocoGlide		DSO-1		AVG	
Model	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc	AUC	bAcc
TruFor	.770	.680	.996	**.984**	.916	.813	.752	.639	**.984**	.930	.884	.809
CAT-Netv2	.680	.635	.977	.803	**.942**	.838	.667	.580	.747	.525	.803	.676
ManTraNet	.760	.500	.810	.500	.644	.500	**.778**	.500	.874	.500	.773	.500
PSCC-Net	.657	.473	.300	.604	.869	.520	.777	.515	.650	.458	.651	.514
SPAN	.670	.235	**.999**	.759	.480	.112	.475	.298	.669	.233	.659	.327
CR-CNN	.553	.391	.755	.631	.670	.481	.589	.447	.576	.289	.629	.448
MVSS-Net	.733	.514	.984	.729	.932	.528	.654	.117	.552	.358	.771	.449
Early Fusion	**.839**	**.770**	.996	.962	.929	.845	.755	.660	.966	**.935**	**.897**	**.834**
Late Fusion	.792	.720	.977	.822	.930	**.860**	.760	**.677**	.958	.830	.884	.782

Table 5. Comparison for detection performance. Metrics are Area Under Curve (AUC) and balanced accuracy (bAcc).

4.3 Ablation Study

In this section, for the purpose of contrasting various forensic filters (SRM, Bayar conv, NoisePrint++), we employ a dual-branch CMX architecture where each filter serves as an auxiliary input alongside the RGB image. The outcomes are presented in Table 6. During this training the bayar convolutional layer is trainable, while SRM and NoisePrint are kept frozen. We can see that NoisePrint++'s editing history based training helps achieve the best performance on DSO-1, where manipulations are covered using post-processing operations, while SRM and bayar perform better in CocoGlide and Coverage. Coverage contains only copy-move manipulations for which NoisePrint's camera model identification might not provide robust enough forensic traces, whereas CocoGlide's manipulations are diffusion-based inpaintings potentially resulting in distinct artifacts that diverge from conventional editing histories. Consequently, NoisePrint encounters difficulties in effectively handling such cases.

Version	Coverage	Columbia	Casiav1+	CocoGlide	DSO-1	AVG
CMX (RGB+NP++)	.577	.884	.761	.516	.895	.726
CMX (RGB+Bayar)	.592	.872	.774	.566	.776	.716
CMX (RGB+SRM)	.630	.834	**.791**	**.585**	.792	.726
Late Fusion (No weight sharing)	.611	**.912**	.760	.566	.785	.727
Early Fusion	**.663**	.888	.784	.553	.863	.750
Late Fusion	.641	.864	.775	.574	**.899**	**.751**

Table 6. Ablation study. Localization results in avg pixel F1.

We compare all methods to our multi-modal fusion approaches and we can see that both the early- and late-fusion paradigms effectively combine the forensic traces provided by the filters, resulting in increased performance. To substantiate our rationale for introducing shared weights between RGB branches in order to enhance regularization within the late fusion paradigm, we also evaluate a method that does not employ weight sharing and observe a substantial improvement in performance for the weight-sharing approach.

4.4 Robustness Analysis

In this section, we include experiments performed on images with varying quality degradations to demonstrate the robustness of our approaches. We use the Casiav1+ dataset and perform Gaussian blurring with different kernel sizes and JPEG compression with varying quality factors and compare to our baseline, TruFor. The findings depicted in Figure 4 demonstrate that both of our fusion approaches exhibit good robustness across a broad spectrum of degradations, maintaining a consistent advantage over TruFor across all degradation levels employed.

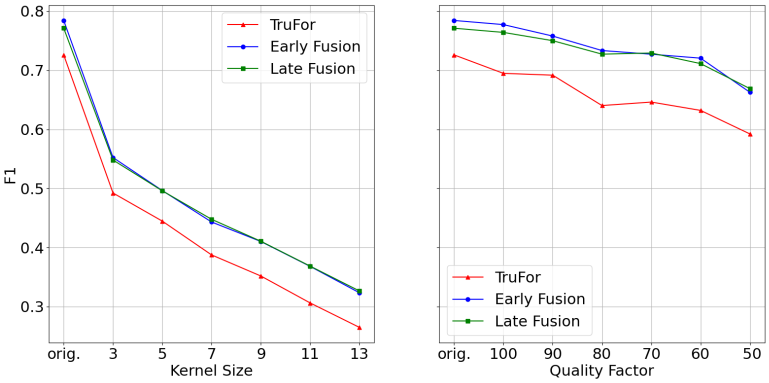

Fig. 4. Robustness analysis with regards to Gaussian blur (left) and JPEG compression (right)

5 Conclusion

In this work, we explore approaches toward expanding existing encoder-decoder architectures for IMLD to support multiple forensic filters as inputs. We compare the performance of approaches using Bayar conv, SRM filters, and NoisePrint++ and discover that they indeed showcase complementary forensic capabilities as was hypothesized. We propose two different modal-fusion paradigms and conduct extensive experiments to demonstrate that both approaches reach state-of-the-art across several datasets, showcasing good generalization abilities, and are effective at leveraging and combining diverse forensic artifacts from different filters. In future work, we would like to explore the performance limitations of models reliant on forensic filters against directed adversarial attacks.

Acknowledgements This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreement H2020-101021866 CRiTERIA.

References

1. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. p. 5–10 (2016). https://doi.org/10.1145/2909827.2930786
2. Chen, X., Dong, C., Ji, J., Cao, J., Li, X.: Image manipulation detection by multi-view multi-scale supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14185–14193 (2021)

3. Cozzolino, D., Poggi, G., Verdoliva, L.: Copy-move forgery detection based on patchmatch. In: 2014 IEEE international conference on image processing (ICIP). pp. 5312–5316. IEEE (2014)
4. Cozzolino, D., Verdoliva, L.: Noiseprint: A cnn-based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2019)
5. De Carvalho, T.J., Riess, C., Angelopoulou, E., Pedrini, H., de Rezende Rocha, A.: Exposing digital image forgeries by illumination color classification. IEEE Transactions on Information Forensics and Security **8**(7), 1182–1194 (2013)
6. 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. pp. 422–426. IEEE (2013)
7. Fan, Y., Xu, W., Wang, H., Wang, J., Guo, S.: Pmr: Prototypical modal rebalance for multimodal learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20029–20038 (2023)
8. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security **7**(3), 868–882 (2012). https://doi.org/10.1109/TIFS.2012.2190402
9. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20606–20615 (June 2023)
10. Hsu, Y.F., Chang, S.F.: Detecting image splicing using geometry invariants and camera characteristics consistency. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 549–552. IEEE (2006)
11. Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., Nevatia, R.: Span: Spatial pyramid attention network for image manipulation localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 312–328. Springer (2020)
12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
13. Kniaz, V.V., Knyaz, V., Remondino, F.: The point where reality meets fantasy: Mixed adversarial generators for image splice detection. Advances in neural information processing systems **32** (2019)
14. Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C.: Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision **130**(8), 1875–1895 (2022)
15. Lahiri, A., Jain, A.K., Agrawal, S., Mitra, P., Biswas, P.K.: Prior guided gan based semantic inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13696–13705 (2020)
16. Li, D., Zhu, J., Wang, M., Liu, J., Fu, X., Zha, Z.J.: Edge-aware regional message passing controller for image forgery localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8222–8232 (2023)
17. Li, H., Luo, W., Huang, J.: Localization of diffusion-based inpainting in digital images. IEEE transactions on information forensics and security **12**(12), 3050–3064 (2017)
18. Liu, X., Liu, Y., Chen, J., Liu, X.: Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 7505–7517 (2022)

19. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16317–16326 (2021)
20. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)
21. Novozamsky, A., Mahdian, B., Saic, S.: Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 71–80 (2020)
22. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
23. Salloum, R., Ren, Y., Kuo, C.C.J.: Image splicing localization using a multi-task fully convolutional network (mfcn). Journal of Visual Communication and Image Representation **51**, 201–209 (2018)
24. Teyssou, D., Leung, J.M., Apostolidis, E., Apostolidis, K., Papadopoulos, S., Zampoglou, M., Papadopoulou, O., Mezaris, V.: The invid plug-in: web video verification on the browser. In: Proceedings of the first international workshop on multimedia verification. pp. 23–30 (2017)
25. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on pattern analysis and machine intelligence **23**(9), 947–963 (2001)
26. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2364–2373 (2022)
27. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12695–12705 (2020)
28. Wen, B., Zhu, Y., Subramanian, R., Ng, T.T., Shen, X., Winkler, S.: Coverage—a novel database for copy-move forgery detection. In: 2016 IEEE international conference on image processing (ICIP). pp. 161–165. IEEE (2016)
29. 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. pp. 9543–9552 (2019)
30. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. Advances in neural information processing systems **34**, 30392–30400 (2021)
31. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
32. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)
33. Yang, C., Li, H., Lin, F., Jiang, B., Zhao, H.: Constrained r-cnn: A general image manipulation detection model. In: 2020 IEEE International conference on multimedia and expo (ICME). pp. 1–6. IEEE (2020)

34. Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R.: Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems pp. 1–16 (2023). https://doi.org/10.1109/TITS.2023.3300537
35. Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R.: Delivering arbitrary-modal semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1136–1147 (2023)
36. Zhang, X., Wang, X., Shi, C., Yan, Z., Li, X., Kong, B., Lyu, S., Zhu, B., Lv, J., Yin, Y., et al.: De-gan: Domain embedded gan for high quality face image inpainting. Pattern Recognition **124**, 108415 (2022)
37. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1053–1061 (2018)