IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2022, accepted April 12, 2022, date of publication April 18, 2022, date of current version April 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3167714

MSRD-CNN: Multi-Scale Residual Deep CNN for General-Purpose Image Manipulation Detection

KAPIL RANA, GURINDER SINGH, AND PUNEET GOYAL, (Member, IEEE)
Indian Institute of Technology Ropar, Rupnagar, Punjab 140001, India

Corresponding author: Kapil Rana (2018csz0007@iitrpr.ac.in)

This work was supported in part by the Indian Institute of Technology Ropar under Institute Scheme for Innovative Research and Development (ISIRD) under Grant 9-231/2016/IIT-RPR/1395, and in part by the Department Of Science & Technology (DST) under the Cognitive Science Research Initiative (CSRI) under Grant DST/CSRI/2018/234.

ABSTRACT The authenticity of digital images is a major concern in multimedia forensics due to the availability of advanced photo editing tools/devices. In the literature, several image forensic methods are available to detect specific image processing or editing operations. However, it remains a challenging task to design a universal forensic method that can detect multiple image editing operations. In this paper, a novel Multi-Scale Residual Deep CNN (MSRD-CNN) is designed to learn the image manipulation features adaptively for multiple image manipulation detection. Our network comprises of three stages: pre-processing, hierarchical high-level feature extraction, and classification. Firstly, a multi-scale residual module is employed in pre-processing stage to extract the prediction error or noise features adaptively. Afterwards, the obtained noise features are processed by feature extraction network having multiple Feature Extraction Blocks (FEBs) for the extraction of high-level image tampering features. Lastly, the resultant feature map is provided to the fully-connected dense layer for classification. The experiment results show that our model surpasses the existing schemes even under anti-forensic attacks, when evaluated on large-scale datasets by considering multiple image processing operations. The proposed network provides overall classification accuracies of 97.07% and 97.48% for BOSSBase and Dresden datasets, respectively.

INDEX TERMS Multiple image manipulation detection, anti-forensic attacks, convolutional neural networks, multi-scale residual module.

I. INTRODUCTION

The digital information can be shared in the form of audio, image, and video using various social media platforms such as Facebook, Instagram, Snapchat, etc. The advent of powerful editing software results in a significant increase in the number of tampered images on social media related to political, individual attacks, publicity, etc. Therefore, the authenticity of digital images is very crucial. Moreover, the investigation of digital images can play important role in many fields related to medical, news media, scientific exploration, law and crime [1]–[3]. Thus, it is a concern of great importance in multimedia forensics.

The detection of different image processing operations has a great relevance to the forensic community due to the fact that these operations may be used by the counterfeiter in the creation of an image forgery. It is perceived that different image processing operations embed special artifacts or footprints in the processed image. Several forensic algorithms have been designed to detect the particular image processing operation by analyzing the corresponding artifacts. Some image processing operations considered are resampling [4]–[8], JPEG compression [9]–[12], median filtering [13]–[16], contrast enhancement [17]–[20], etc. Also, many anti-forensic approaches related to different image processing operations such as JPEG compression [21], [22], median filtering [23], and contrast enhancement [24] have also been proposed to mislead the forensic techniques by concealing the footprints of corresponding image processing operations.

The researchers have also developed general-purpose image manipulation detection schemes to detect different image processing operations [25]–[30]. Moreover, it is observed that recent works on multi-purpose image tampering detection are based on deep learning techniques, for instance, Convolutional Neural Networks (CNNs).

The associate editor coordinating the review of this manuscript and approving it for publication was Sudipta Roy.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

These CNNs have demonstrated the ability to automatically learn the image manipulation features from data. A novel constrained convolutional layer based CNN is proposed in [25] to detect the multiple image processing operations by suppressing the image content information and the authors further optimized their constrained neural network in [28] for better performance. In [26], a densely connected CNN based on isotropic constraint is proposed for general-purpose image forensics by considering the anti-forensic attacks. The isotropic convolutional layer works as a high-pass filter to highlight the image processing operations artifacts by suppressing the image content information. Moreover, an image manipulation detection approach built upon [25] and combined with a deep Siamese CNN network is presented in [27]. However, their work was not to identify the specific image manipulation but to classify the input patch pair (two images) whether they are identically processed or not. In [29], Xception architecture is employed to classify multiple image processing operations by considering small-sized images. Most of the existing general-purpose forensic techniques can be easily circumvented by using some anti-forensic attacks. Recently, a universal image manipulation detection approach based on densely-connected CNN is proposed in [30] and it has also considered most of the image processing operations including various anti-forensic techniques for evaluation. However, the proposed CNN is significantly different from the existing approach [30] in terms of network architecture as well as used image manipulation datasets.

Overall, designing a unified forensic scheme capable of detecting different image manipulations under different attacks is still a challenging task for the researchers. Also, to the best of our knowledge, the existing works have not performed any cross dataset testing to evaluate the generalization of their models. In this work, we present a novel and effective image manipulation detection approach capable of detecting multiple editing operations including anti-forensic methods. The main contributions of our work are as follows:

- We propose a novel method: MSRD-CNN for general-purpose image manipulation detection.
- Inspired by Res2Net [31], we propose a multi-scale residual module to obtain efficient noise features adaptively. Further, the obtained noise features are processed by using FEBs to extract the high-level image manipulation features.
- In this paper, we have considered several image processing operations including anti-forensic schemes and with arbitrary parameters to evaluate our network. The extensive experiment results show that our MSRD-CNN provides better accuracy in comparison to the existing methods, even in cross-dataset settings.

The remaining part of the paper includes a detailed description of the proposed network in Section II and the experiment results are discussed in Section III. Finally, we conclude our work in Section IV.

II. PROPOSED MSRD-CNN ARCHITECTURE

In this section, we propose a novel MSRD-CNN architecture capable of detecting the traces of multiple image processing operations and anti-forensic techniques. The architecture of MSRD-CNN, as shown in Fig. 1, includes three different stages i.e., extraction of noise features using a multi-scale residual module, feature extraction network to extract high-level features related to image tampering artifacts, and classification.

A. MULTI-SCALE RESIDUAL MODULE

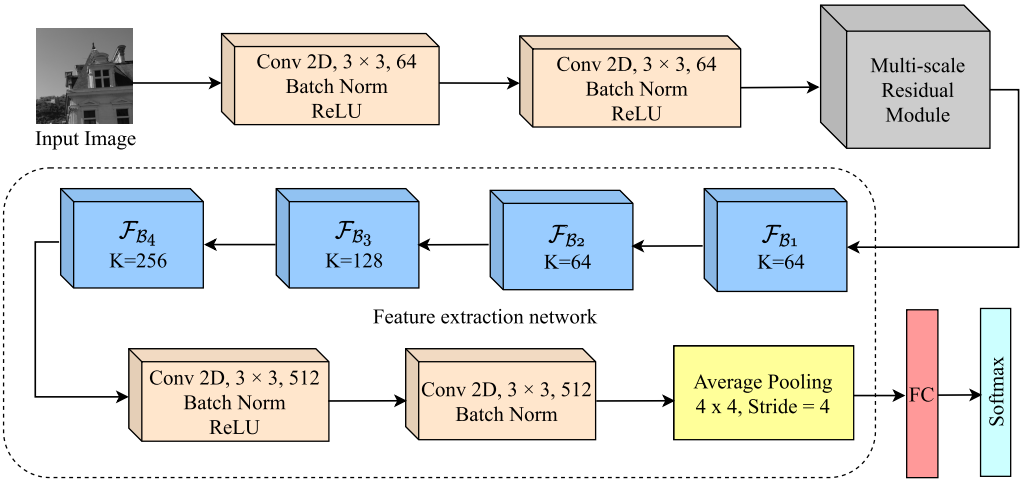
Most of the image manipulation detection schemes use the idea of suppressing the content information of an input image to highlight the image manipulation artifacts. Compared to applying fixed filters to the input image prior to CNN for the extraction of prediction error features, it is preferred to employ a trainable filtering scheme for pre-processing to potentially learn more appropriate image manipulation features adaptively for image forensic tasks. In our approach, we use a data-driven pre-processing scheme that consists of a two-layer CNN and a multi-scale residual module. Each convolution layer in the two-layer CNN contains 64 filters of 3×3 followed by batch normalization and the ReLU layer. This two-layer CNN is employed to obtain better input features for the multi-scale residual module. Let us denote the functions of these two convolution layers by $C_1(\cdot)$ and $C_2(\cdot)$, respectively. For a given input image I of size 256×256, the output of this two-layer CNN is formulated as:

$$I_{C_1C_2} = C_2(C_1(I)), \tag{1}$$

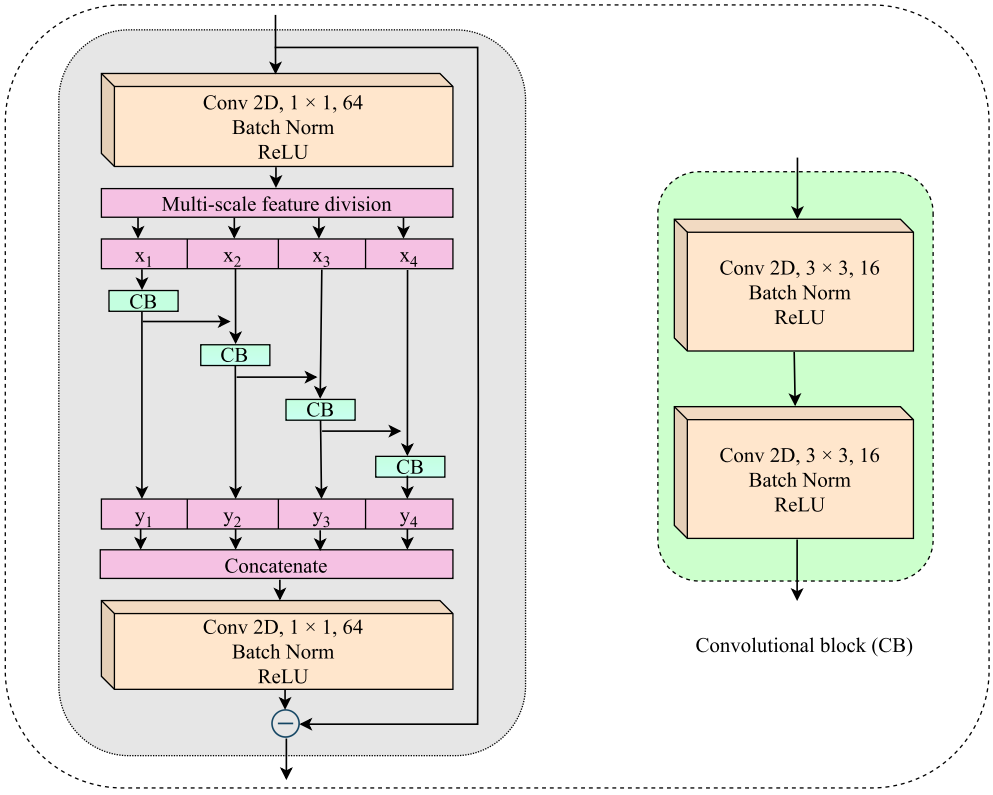
This output $I_{C_1C_2}$, having size of $256 \times 256 \times 64$, is then passed to the multi-scale residual module which is inspired from Res2Net [31] and designed to learn the suitable noise features. The proposed multi-scale residual module explores the multi-scale feature representation by dividing the input features of size $256 \times 256 \times 64$ along the channel axis, which results in four different groups of size $256 \times 256 \times 16$. These groups are then interconnected in a hierarchical residual-like style as shown Fig. 1(b). Each group is further processed by a Convolutional Block (CB) having two convolution layers with 16 filters of 3×3 followed by batch normalization and ReLU layers. The output feature maps of the first CB is added to the second group before passing to the second CB as shown in Fig. 1(b). Let x_i represents the feature maps of i^{th} group, where $i \in \{1, 2, 3, 4\}$, and $H_i(\cdot)$ is the function performed by the convolutional block of i^{th} group. The output of $H_i(\cdot)$ which is y_i will be added to x_{i+1} group and passed to $(i+1)^{th}$ convolutional block (H_{i+1}) as provided in Eq. (2).

$$y_i = \begin{cases} H_i(x_i) & i = 1 \\ H_i(x_i + y_{i-1}) & i = 2, 3, 4 \end{cases} \tag{2}$$

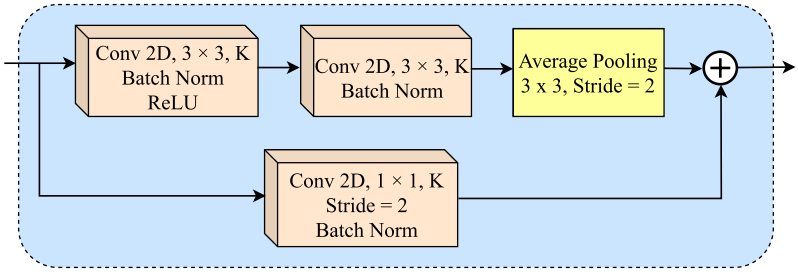
The outputs of all the convolutional blocks are concatenated and passed to a convolution layer having 64 filters of size 1×1. The output of this convolutional layer is subtracted from the input of the multi-scale residual module to obtain the



(a) Overall Architecture



(b) Multi-scale residual module



(c) Feature extraction block ($\mathcal{F}_{\mathcal{B}}$), K = number of filters

FIGURE 1. **Proposed MSRD-CNN architecture.**

final noise features as:

$$I_{MSRM} = MSRM(I_{C_1C_2}) - I_{C_1C_2} \tag{3}$$

where, $MSRM(\cdot)$ denotes the function performed by the multi-scale residual module. The feature extraction blocks further process these noise features to extract the high-level image manipulation features. Note that the features size i.e., height and width remains same during the pre-processing stage except the channel size.

B. FEATURE EXTRACTION NETWORK

The noise features obtained from the multi-scale residual module are passed to the feature extraction network to extract the high-level image manipulation features. This feature extraction network has four FEBs and each FEB ($\mathcal{F}_{\mathcal{B}}$) is based on a residual skip connection containing two regular convolution layers of size 3×3 and a 1×1 convolution layer. The input of a FEB is added to the output of the second convolution layer followed by the average pooling operation as shown in Fig. 1(c). Note that we have not used the pooling layer in the multi-scale residual module of pre-processing stage because pooling layer strengths the image content and reduces noise signal by averaging. The purpose of the pooling layer is to down-sample the features for learning high-level image manipulation features. Number of filters in the four FEBs i.e $\mathcal{F}_{\mathcal{B}1}, \mathcal{F}_{\mathcal{B}2}, \mathcal{F}_{\mathcal{B}3}$, and $\mathcal{F}_{\mathcal{B}4}$ are 64, 64, 128 and 256 respectively. The resultant features obtained from this feature extraction network can be formulated as:

$$I'_{\mathcal{F}_{\mathcal{B}}} = \mathcal{F}_{\mathcal{B}4}(\mathcal{F}_{\mathcal{B}3}(\mathcal{F}_{\mathcal{B}2}(\mathcal{F}_{\mathcal{B}1}(I_{MSRM})))) \tag{4}$$

The output of this feature extraction network i.e. $I'_{\mathcal{F}_{\mathcal{B}}}$ is further processed by two convolution layers each having 64 filters of size 3×3 to obtain the more relevant image manipulation features. First convolution layer is followed by batch normalization and ReLU and the second convolutional layer is followed by batch normalization. Afterward, the average pooling layer with filter size 4×4 and stride 4 is applied to reduce the feature dimension.

Lastly, the global features obtained after the average pooling layer is fed to a fully-connected (FC) layer with 11 neurons corresponding to image processing operations used for classification. We use the softmax function to get the probability of predicted classes and the cross-entropy function to calculate the overall network loss.

C. COMPARISON WITH GIMD-NET

The proposed CNN is significantly different from the existing GIMD-Net approach [30] in terms of network architecture as well as used image manipulation datasets. The model proposed in [30] is inspired from DenseNet [32] employing the concept of local and global residual learning for the extraction of high-level image manipulation features using residual dense blocks (RDBs). On the contrary, the proposed MSRD-CNN is inspired from Res2Net [31] that learns prediction error features adaptively to highlight the image manipulation artifacts and then extract high-level hierarchical image tampering features by using feature extraction network. In [30], there is no preprocessing used to extract the noise features, whereas we propose a preprocessing stage (multi-scale residual module) to extract the noise features adaptively. The RDBs used in [30] are fussed globally and the convolutional layers used in each RDB are densely connected to comfort the training and optimization. But, instead of using global fusion, the FEBs in proposed method are connected sequentially to extract the high-level features. Also, the convolutional layers are employed without dense connectivity in each FEB. Further, the image processing operations used in [30] are based on fixed parameters to create image manipulation datasets. On the other hand, we have created image manipulation datasets based on arbitrary parameters as shown in Table 1. Therefore, we have considered a more challenging dataset in this work to evaluate the model performance as compared to [30].

III. EXPERIMENTAL RESULTS

We conducted extensive experiments to evaluate the performance of the proposed model in the detection of multiple image processing operations and various anti-forensic attacks. Firstly, to confirm the multi-purpose nature of our MSRD-CNN, we considered 10 image processing operations along with corresponding parameters listed in Table 1. The image processing parameters are selected randomly to create more challenging image manipulation datasets. For instance, in JPEG compression, we compress the original images by randomly selecting the Quality Factor (QF) ranging from 60 to 90.

TABLE 1. Different image processing operations used for the generation of manipulation datasets with arbitrary parameters.

Image editing operations	Parameters
JPEG compression (JPEG)	$QF = 60, 61, 62, \ldots, 90$
Gaussian Blurring (GB)	$\sigma = 0.7, 0.9, 1.1, 1.3$
Adaptive White Gaussian Noise (AWGN)	$\sigma = 1.4, 1.6, 1.8, 2$
Resampling (RS) using bilinear interpolation	Scaling $= 1.2, 1.4, 1.6, 1.8, 2$
Median Filtering (MF)	Kernel $= 3, 5, 7, 9$
Contrast Enhancement (CE)	$\gamma = 0.6, 0.8, 1.2, 1.4$
JPEG anti-forensics (JPEGAF) [21]	$QF = 60, 61, 62, \ldots, 90$
JPEG anti-forensics (JPEGAF) [22]	$QF = 60, 61, 62, \ldots, 90$
Median filtering anti-forensics (MFAF) [23]	Kernel $= 3, 5, 7, 9$
Contrast enhancement anti-forensics (CEAF) [24]	$\gamma = 0.6, 0.8, 1.2, 1.4$

We consider BOSSBase [33] and Dresden image dataset [34] for the evaluation of different image tampering detection approaches. The standard BOSSBase dataset comprises of 10,000 grayscale images of resolution 512×512 in PGM format. We have transformed these PGM images into PNG format for evaluation purposes. The standard Dresden dataset contains 3008×2000 size 1491 raw images in NEF format. We converted these raw images into PNG format for evaluation. Our model is implemented by using PyTorch 1.8 deep learning framework and all the experiments are performed using Tesla V100 GPU with 32GB RAM. We compared our network with recent multi-purpose image tampering detection methods [26], [28]–[30] in terms

of detection accuracy. We also assessed our model's robustness and generalization by performing cross-dataset testing. The experimental results exhibit the efficacy of the proposed model in comparison to the existing image manipulation detection methods. All the relevant codes are available on request for reproducibility and research advancement.

A. MULTIPLE IMAGE MANIPULATION DETECTION

In this subsection, we evaluate our MSRD-CNN performance in the detection of multiple image processing operations including anti-forensic techniques using BOSSBase and Dresden datasets. We created one original image (OR) and 10 tampered image datasets using the image processing operations as listed in Table 1 by considering 4,167 and 1,333 images sequentially from the BOSSBase dataset for training and testing, respectively. We extracted 4 patches of size 256 $\times$ 256 from each of these images, which results in 16,668 training and 5,332 testing images for each of the image processing operations. Therefore, we obtained a dataset having 2,42,000 grayscale images. We used 1,83,348 images (including 16,668 original images) for training, and remaining 58,652 images (including 5,332 original images) for testing purposes. Note that we follow the strategy used by the existing works [28] to create image manipulation datasets corresponding to different image manipulation operations to make the comparison feasible. Therefore, we have used only 4167 and 1333 images from the BOSSBase dataset for training and testing, respectively. This may also be noted that the complete BOSSBase dataset images are not used in consideration to the limited computational facilities availability, as we are considering 10 image manipulation methods including anti-forensic approaches which are highly compute-intensive and time-consuming.

We also evaluated our network ability using 881 images from the Dresden dataset. We follow the same strategy as used for the BOSSBase dataset in preparing image manipulation datasets using different image processing operations. We considered 667 images for training and 214 images for testing the considered neural networks. All of these images are cropped from the center to obtain a sub-image region of size 1280 $\times$ 1280. Afterward, each sub-image region is processed to extract 25 patches of size 256 $\times$ 256 and then converted into grayscale format. Therefore, we obtained 16,668 (approx.) images for training and 5,332 (approx.) images for testing corresponding to image processing operations provided in Table 1. The training of our network is performed by using the Adam optimizer with a learning rate of 0.001 and we trained our network for 100 epochs in each experiment.

We evaluated confusion matrices for our model based on multiple image processing operations for BOSSBase and Dresden datasets as shown in Tables 2 and 3. Our MSRD-CNN provides average accuracies of 97.07% and 97.48% for BOSSBase and Dresden datasets, respectively, when evaluated on multiple image processing operations. Table 2 reveals that the proposed network gives an accuracy of greater than 97% for each image processing operation except for the original and CE images on the BOSSBase dataset. The accuracy of original and contrast-enhanced images is 87.92% and 90.15%, respectively for the BOSSBase dataset. Table 3 demonstrates that our proposed approach identifies each image processing operation with an accuracy of greater than 97% except for the original and contrast-enhanced images with 92.22% and 85.03% respectively on the Dresden dataset. Moreover, the robustness of our model is confirmed by the fact that it provides high accuracies against different anti-forensic approaches on both the datasets.

We also conducted an experiment by combining both the training sets of BOSSBase and Dresden datasets. It is observed that combining both the training datasets increases the model accuracy further, likely because of the increase of training dataset size and/or more diversity. The testing accuracy increases from 97.07% to 97.38% on the BOSSBase test dataset. Similarly, model testing accuracy increases from 97.48% to 98.11% on the Dresden test dataset. However, the training time increases significantly due to the large training data.

B. COMPARATIVE ANALYSIS WITH EXISTING APPROACHES

We compared our MSRD-CNN with existing multi-purpose forensic schemes [26], [28]–[30] by considering multiple images processing operations including anti-forensic techniques using the same training and testing datasets as defined in Section III-A. We provide the diagonal entries of confusion matrices in Table 4 for different methods for ease of comparison. The proposed model provides better detection as compared to the existing approaches for all the considered image manipulations except GB, JPEGAF [22], and CEAF [24] operations, when tested on the BOSSBase dataset as shown in Table 4. Similarly, our network achieves better detection accuracy for all image manipulations except JPEG, GB, and CE operations for the Dresden dataset. However, it may be noted that for GB and CEAF [24] operations in the BOSSBase dataset, our model is second best and is around 0.2% lower than the best performing method. Also, for the JPEG and GB operations in Dresden dataset, our method is 0.02% and 0.17% lower than the best performing method, respectively. Moreover, Table 4 shows that our model outperforms the recent deep learning based scheme [30] with average accuracy improvements of 1.04% and 1.48% for the BOSSBase and Dresden datasets, respectively.

C. PERFORMANCE EVALUATION BASED ON CROSS DATASET IMAGES

In this subsection, we evaluate the performance of our network by considering cross dataset testing images. In the first experiment, the considered models, trained on the BOSSBase training dataset images, are applied on the Dresden test set images. Similarly, we also perform the experiments considering Dresden training dataset images and BOSSBase test dataset images. The average accuracy results of these cross

TABLE 2. Confusion matrix for the proposed model evaluated on multiple image processing operations as given in Table 1 on BOSSbase dataset (1,83,348 training and 58,652 testing images. Overall test accuracy = 97.07%).

		Predicted Class										
		OR	JPEG	GB	WN	RS	MF	CE	JPEGAF [21]	JPEGAF [22]	MFAF [23]	CEAF [24]
True Class	OR	**87.92**	0.00	0.47	2.06	0.11	0.00	7.95	0.02	0.00	0.00	1.46
	JPEG	0.00	**99.89**	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.02
	GB	0.00	0.00	**99.70**	0.00	0.17	0.00	0.08	0.00	0.00	0.00	0.06
	WN	0.47	0.00	0.00	**99.34**	0.00	0.02	0.15	0.00	0.02	0.00	0.00
	RS	0.21	0.00	1.26	0.00	**97.81**	0.00	0.39	0.00	0.00	0.00	0.34
	MF	0.00	0.00	0.02	0.00	0.00	**99.76**	0.17	0.00	0.00	0.06	0.00
	CE	6.15	0.02	0.21	1.29	0.06	0.08	**90.15**	0.02	0.00	0.00	2.03
	JPEGAF [21]	0.02	0.00	0.00	0.06	0.00	0.00	0.08	**98.42**	1.43	0.00	0.00
	JPEGAF [22]	0.08	0.00	0.00	0.13	0.00	0.00	0.02	1.91	**97.86**	0.00	0.00
	MFAF [23]	0.00	0.00	0.00	0.00	0.00	0.11	0.13	0.00	0.00	**99.76**	0.00
	CEAF [24]	0.32	0.00	0.83	0.00	0.08	0.08	1.54	0.00	0.00	0.00	**97.17**

TABLE 3. Confusion matrix for the proposed model evaluated on multiple image processing operations as given in Table 1 on Dresden dataset (1,83,348 training and 58,652 testing images. Overall test accuracy = 97.48%).

		Predicted Class										
		OR	JPEG	GB	WN	RS	MF	CE	JPEGAF [21]	JPEGAF [22]	MFAF [23]	CEAF [24]
True Class	OR	**92.22**	0.00	0.00	0.02	0.02	0.02	7.63	0.00	0.00	0.00	0.09
	JPEG	0.00	**99.98**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
	GB	0.00	0.00	**99.74**	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.13
	WN	0.00	0.00	0.00	**99.94**	0.00	0.00	0.00	0.02	0.04	0.00	0.00
	RS	0.00	0.00	0.11	0.00	**99.81**	0.00	0.02	0.00	0.04	0.00	0.02
	MF	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00
	CE	11.89	0.00	0.00	0.09	0.02	0.00	**85.03**	0.00	0.04	0.00	2.93
	JPEGAF [21]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**99.49**	0.51	0.00	0.00
	JPEGAF [22]	0.00	0.00	0.00	0.00	0.02	0.00	0.00	2.55	**97.43**	0.00	0.00
	MFAF [23]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00
	CEAF [24]	0.36	0.00	0.09	0.00	0.02	0.00	0.84	0.00	0.06	0.00	**98.63**

TABLE 4. Performance comparison of different multi-purpose forensic schemes by considering multiple image processing operations.

	BOSSBase Dataset					Dresden Dataset				
	Chen [26]	Bayar [28]	Yang [29]	Singh [30]	Ours	Chen [26]	Bayar [28]	Yang [29]	Singh [30]	Ours
OR	23.65	54.86	79.76	82.50	**87.92**	39.07	23.48	35.54	81.40	**92.22**
JPEG	96.66	99.72	99.83	99.76	**99.89**	99.36	99.72	**100.00**	**100.00**	99.98
GB	98.52	99.36	**99.93**	99.61	99.70	99.91	94.90	99.79	**99.91**	99.74
AWGN	80.65	93.12	98.54	98.33	**99.34**	98.95	96.02	98.91	**99.94**	**99.94**
RS	62.04	90.72	97.51	96.31	**97.81**	82.2	84.26	98.33	99.27	**99.81**
MF	88.77	97.32	97.60	99.40	**99.76**	93.55	97.39	99.81	99.96	**100.00**
CE	28.84	50.21	65.56	86.53	**90.15**	53.06	74.79	78.94	**88.32**	85.03
JPEGAF [21]	56.77	87.45	96.85	97.99	**98.42**	51.03	79.95	97.85	99.01	**99.49**
JPEGAF [22]	63.93	93.57	97.68	**98.87**	97.86	59.75	70.93	95.99	95.72	**97.43**
MFAF [23]	95.16	99.29	99.42	99.64	**99.76**	95.37	99.08	99.87	99.94	**100.00**
CEAF [24]	76.44	93.34	95.35	**97.37**	97.17	58.55	66.84	89.45	92.55	**98.63**
Overall Avg.	70.13	87.18	93.45	96.03	**97.07**	75.53	80.67	90.41	96.00	**97.48**

dataset testing experiments are presented in Table 5 and it is observed that our MSRD-CNN architecture outperforms the recent multi-purpose forensic schemes by providing higher detection accuracies of 86.49% and 81.40% for BOSSTrain-DRESTest and DRESTrain-BOSSTest, respectively. It is also noted from Table 5 that all the considered forensic methods do not perform well for the original images because the proposed model focuses on the artifacts introduced by the image manipulation operations in the image. But, the original images do not have any manipulation artifacts except the camera fingerprint-related features. Moreover, the original images of these two datasets are acquired from different camera models/devices. Therefore, we also provided the overall average accuracies excluding the original images as shown in Table 5. These results are also in favour of proposed MSRD-CNN, with 95.1% and 87.7% accuracies in two settings considered. This highlights the overall best generalization ability of the proposed approach.

D. ABLATION STUDIES

The performance of our MSRD-CNN is examined considering the different architectural design choices to achieve an optimal design for the proposed model. Initially, we evaluate our MSRD-CNN model with different number of initial

TABLE 5. Performance comparison of different general-purpose image manipulation schemes by considering cross dataset testing.

	Models trained on BOSSBase and tested on Dresden dataset (BOSSTrain-DRESTest)					Models trained on Dresden and tested on BOSSBase dataset (DRESTrain-BOSSTest)				
	Chen [26]	Bayar [28]	Yang [29]	Singh [30]	**Ours**	Chen [26]	Bayar [28]	Yang [29]	Singh [30]	**Ours**
OR	**9.96**	1.74	2.03	0.04	0.17	25.84	3.04	**28.88**	1.88	18.55
JPEG	97.09	99.59	99.42	99.59	**99.96**	94.71	98.57	97.21	98.91	**99.83**
GB	**99.94**	99.06	99.27	99.83	99.76	93.98	88.47	**98.24**	97.81	91.65
AWGN	94.28	78.17	93.45	89.89	**95.09**	71.92	79.41	**94.34**	92.16	85.60
RS	74.72	41.64	69.47	82.37	**96.08**	59.92	73.37	**87.30**	77.44	85.07
MF	74.47	95.09	92.74	99.34	**99.98**	81.73	94.35	96.19	**98.54**	98.35
CE	33.36	29.24	31.83	76.03	**92.12**	18.06	32.24	36.37	27.89	**46.40**
JPEGAF [21]	35.90	71.57	88.24	92.24	**92.78**	48.82	71.27	86.42	92.76	**95.57**
JPEGAF [22]	69.47	80.95	88.62	**91.13**	89.20	51.11	78.96	91.37	91.92	**93.23**
MFAF [23]	96.98	99.62	98.67	**99.74**	**99.74**	84.47	98.69	**99.51**	99.01	98.26
CEAF [24]	58.78	66.32	67.07	79.22	**86.52**	63.92	37.55	72.69	66.04	**82.93**
Overall Avg.	67.72	69.36	75.53	82.67	**86.49**	63.14	68.72	80.77	76.76	**81.40**
Overall Avg. excluding OR	73.50	76.13	82.88	90.94	**95.12**	66.86	75.29	85.96	84.25	**87.69**

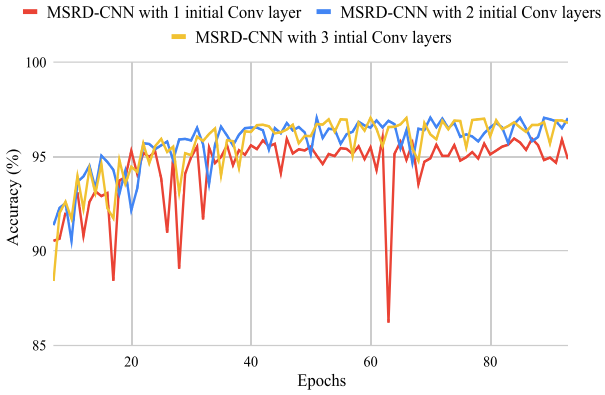


FIGURE 2. Testing accuracies versus number of epochs curves for our model based on the different choices of initial convolutional layers on BOSSBase dataset.

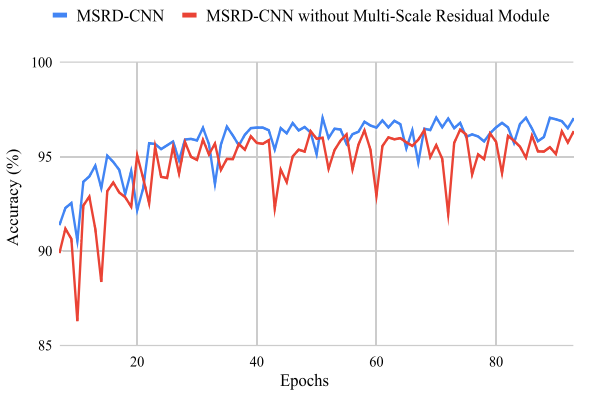


FIGURE 3. Testing accuracies versus number of epochs curves for our model based on the different structural design choices related to multi-scale residual module on BOSSBase dataset.

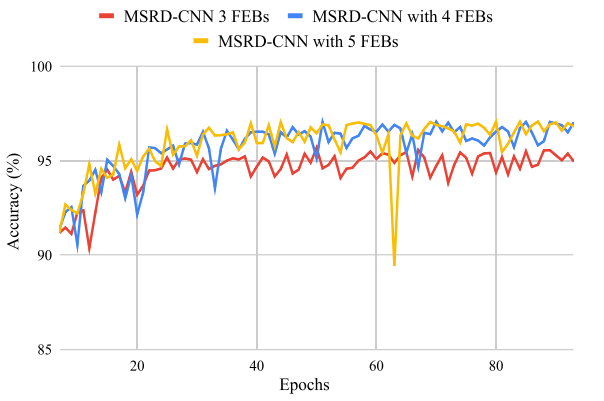


FIGURE 4. Testing accuracies versus number of epochs curves for our model based on the different number of FEBs on BOSSBase dataset.

convolution layers in pre-processing stage. Then, we examine the influence of multi-scale residual module on the model performance. Moreover, we also conducted experiments to evaluate the effect of number of FEBs on the model performance. We also perform experiments related to the choice of activation function used in the proposed model. All of these experiments based on different structural design choices are performed by considering multiple image processing operations on BOSSBase dataset. We have also plotted testing accuracy versus number of epochs for these experiments, as shown in Figs. 2 to 5.

In the first ablation study i.e., when different number of initial convolutional layers are considered, the overall classification accuracy of 95.99%, 97.07%, and 97.06% is achieved with one, two, and three convolutional layers, respectively. It is observed that accuracy is around 1.07% less when using only one convolutional layer and the accuracy in the case of two and three initial convolution layers is almost same. But training time increases significantly in the case of three initial convolution layers. This is because the pre-processing stage does not contain any pooling layer and perform convolution operations with full sized image. This results in the increase in the number of training parameters and the training time with the addition of each initial convolution layer. It is clear from the Fig. 2 that our MSRD-CNN with two initial convolution layers consistently perform better by providing higher

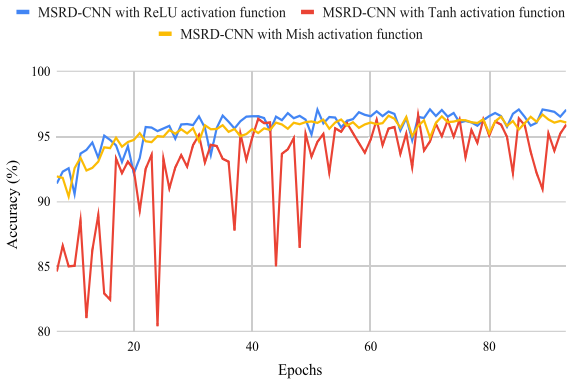


FIGURE 5. Testing accuracies versus number of epochs curves for our model based on the different design choices of activation on BOSSBase dataset.

classification accuracy for most of the epochs as compared to the other design choices. Moreover, we also evaluated our model performance without multi-scale residual module to reveal its importance. It is observed from Fig. 3 that our model with multi-scale residual module consistently performs better as compared to MSRD-CNN without multi-scale residual module by providing higher accuracy for most of the epochs. Therefore, these results reveal the importance of the multi-scale residual module in our proposed network.

In another experiment, we perform ablation study on number of FEBs. The classification accuracy with three, four, and five FEBs is 95.69%, 97.07%, and 97.08%, respectively. It is observed from Fig. 4 that there is a significant improvement in accuracy for all epochs, when number of FEBs are increased from three to four. But, when we evaluated our model by considering five FEBs, there is not much improvement in classification accuracy. However, adding FEBs to the model also increases the computation cost by increasing the total number of model parameters. Therefore, we choose four FEBs in our proposed model.

We also perform experiments by considering Tanh and recent Mish [35] activation functions to evaluate the model performance. Again, it is observed that the proposed MSRD-CNN (with ReLU activation function) provides better performance than the Tanh and Mish activation functions as shown in Fig. 5.

IV. CONCLUSION

In this paper, a novel general-purpose forensic approach is proposed for image manipulation detection. Our MSRD-CNN employs a multi-scale residual module to learn the prediction error features adaptively by suppressing the image content information. A feature extraction network further processes these low-level forensic features to provide high-level image manipulation features for better classification. A series of experiments were performed using two large-scale datasets. The results consistently show that our model can effectively classify different image processing operations, including anti-forensic attacks. Our model provides overall accuracy improvements of 1.04% and 1.48% as compared to the recent forensic method [30] on BOSSBase and Dresden datasets, respectively. Even in cross dataset testing settings, our model outperforms other approaches and exhibits good generalization ability. In the future, we further plan to evaluate the robustness of our network against adversarial attacks and image manipulation chain detection scenarios.

REFERENCES

[1] S. Battiato, O. Giudice, and A. Paratore, "Multimedia forensics: Discovering the history of multimedia contents," in *Proc. 17th Int. Conf. Comput. Syst. Technol.*, 2016, pp. 5–16.
[2] C. Pasquini, G. Boato, and R. Bohme, "Teaching digital signal processing with a challenge on image forensics [SP Education]," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 101–109, Mar. 2019.
[3] M. C. Stamm, M. Wu, and K. J. R. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, 2013.
[4] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, Feb. 2005.
[5] M. Kirchner, "Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue," in *Proc. 10th ACM Workshop Multimedia Secur. (MM)*, 2008, pp. 11–20.
[6] N. Dalgaard, C. Mosquera, and F. Perez-Gonzalez, "On the role of differentiation for resampling detection," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 1753–1756.
[7] X. Feng, I. J. Cox, and G. Doerr, "Normalized energy density-based forensic detection of resampled images," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 536–545, Jun. 2012.
[8] B. Mahdian and S. Saic, "Blind authentication using periodic properties of interpolation," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 529–538, Sep. 2008.
[9] T. Bianchi and A. Piva, "Detection of non-aligned double JPEG compression with estimation of primary compression parameters," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 1929–1932.
[10] R. Neelamani, R. de Queiroz, Z. Fan, S. Dash, and R. G. Baraniuk, "JPEG compression history estimation for color images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1365–1378, Jun. 2006.
[11] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.
[12] Z. Qu, W. Luo, and J. Huang, "A convolutive mixing model for shifted double JPEG compression with application to passive image authentication," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 1661–1664.
[13] M. Kirchner and J. Fridrich, "On detection of median filtering in digital images," *Proc. SPIE*, vol. 7541, Jan. 2010, Art. no. 754110.
[14] X. Kang, M. C. Stamm, A. Peng, and K. J. R. Liu, "Robust median filtering forensics using an autoregressive model," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 9, pp. 1456–1468, Sep. 2013.
[15] G. Cao, Y. Zhao, R. Ni, L. Yu, and H. Tian, "Forensic detection of median filtering in digital images," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 89–94.
[16] C. Chen and J. Ni, "Median filtering detection using edge based prediction matrix," in *Proc. Int. Workshop Digit. Watermarking*. Berlin, Germany: Springer, 2011, pp. 361–375.
[17] M. C. Stamm and K. J. R. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 492–506, Sep. 2010.
[18] H. Yao, S. Wang, and X. Zhang, "Detect piecewise linear contrast enhancement and estimate parameters using spectral analysis of image histogram," in *Proc. IET Int. Commun. Conf. Wireless Mobile Comput. (CCWMC)*, 2009, pp. 94–97.
[19] M. Stamm and K. J. R. Liu, "Blind forensics of contrast enhancement in digital images," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 3112–3115.
[20] M. C. Stamm and K. J. R. Liu, "Forensic estimation and reconstruction of a contrast enhancement mapping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 1698–1701.

[21] W. Fan, K. Wang, F. Cayre, and Z. Xiong, "A variational approach to JPEG anti-forensics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3058–3062.
[22] W. Fan, K. Wang, F. Cayre, and Z. Xiong, "JPEG anti-forensics with improved tradeoff between forensic undetectability and image quality," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 8, pp. 1211–1226, Aug. 2014.
[23] W. Fan, K. Wang, C. François, and Z. Xiong, "Median filtered image quality enhancement and anti-forensics via variational deconvolution," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 1076–1091, May 2015.
[24] H. Ravi, A. V. Subramanyam, and S. Emmanuel, "ACE–An effective anti-forensic contrast enhancement technique," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 212–216, Feb. 2016.
[25] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2016, pp. 5–10.
[26] Y. Chen, X. Kang, Z. J. Wang, and Q. Zhang, "Densely connected convolutional neural network for multi-purpose image forensics under anti-forensic attacks," in *Proc. 6th ACM Workshop Inf. Hiding Multimedia Secur.*, 2018, pp. 91–96.
[27] A. Mazumdar, J. Singh, Y. Singh Tomar, and P. Kumar Bora, "Universal image manipulation detection using deep Siamese convolutional neural network," 2018, *arXiv:1808.06323*.
[28] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2691–2706, Nov. 2018.
[29] L. Yang, P. Yang, R. Ni, and Y. Zhao, "Xception-based general forensic method on small-size images," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*. Singapore: Springer, 2020, pp. 361–369.
[30] G. Singh and P. Goyal, "GIMD-net: An effective general-purpose image manipulation detection network, even under anti-forensic attacks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.
[31] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.
[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[33] P. Bas, T. Filler, and T. Pevná, "'Break our steganographic system': The ins and outs of organizing BOSS," in *Proc. Int. Workshop Inf. Hiding*. Berlin, Germany: Springer, 2011, pp. 59–70.
[34] T. Gloe and R. Böhme, "The 'dresden image database' for benchmarking digital image forensics," in *Proc. ACM Symp. Appl. Comput.*, 2010, pp. 1584–1590.
[35] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*.

KAPIL RANA received the B.Tech. degree in computer science and engineering from Kurukshetra University, Haryana, India, in 2013, and the M.Tech. degree in computer science and engineering from the National Institute of Technology Calicut, Kerala, India, in 2017. He is currently pursuing the Ph.D. degree in computer science and engineering with the Indian Institute of Technology Ropar, Punjab, India. His research interests include image forensics and deep learning.

GURINDER SINGH received the Ph.D. degree in electronics and communication engineering from the Thapar Institute of Engineering and Technology, Punjab, India, in 2019. He is currently working as a Postdoctoral Researcher with the Computer Science and Engineering Department, Indian Institute of Technology Ropar, Punjab. He was awarded with the Visvesvaraya Ph.D. Scheme for Electronics and IT Fellowship, MeitY, Government of India. His research interests include multimedia forensics, anti-forensics, computer vision, and deep learning.

PUNEET GOYAL (Member, IEEE) received the dual B.Tech. and M.Tech. degrees in computer science and engineering from IIT Delhi, in 2006, and the Ph.D. degree in electrical and computer engineering from Purdue University, USA, in 2010. He is currently working as an Associate Professor with the Department of Computer Science and Engineering, Indian Institute of Technology Ropar, India. His current research interests include image processing, computer vision, image forensics, applied deep learning, and assistive technologies.