Spatial Feature-based Fake Capsule Network Model for Deep fake Detection for Image and Video Data

B. N. Karthik
Research Scholar
Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Email: jayamkarthik85@gmail.com

Dr. P. Anbalagan
Assistant Professor
Dept. of Computer Science and Engineering
Annamalai University
Annamalainagar – 608002
Email: anbalagansamy@gmail.com

Dr. G. Pradeep
Professor / MCA
AVC College of Engineering
Mayiladuthurai – 609305
Email: pradeep.g8@gmail.com

Article Info
Page Number: 1481-1489
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract— The development in the image or video editing techniques paved the way for attackers to make fake videos and images. To overcome this problem, Convolutional Neural Network (CNN) techniques had been introduced and it had delivered substantial results. But the fake videos created using the Deepfake tool had been challenging to the existing CNN techniques. Also, CNN has drawbacks such as the network being significantly slow due to max pool operation and requiring a large dataset to train and process the neural network. The drawbacks of CNN can be overcome by Capsule Networks to detect Deepfake videos. Spatial Feature based Fake Capsule Network Model (FCNM) is proposed to detect fake news through images and video. The FCNM model comprises of Capsule structures, Exponential Linear Unit (ELU), LP Pooling layer and dynamic routing algorithm. The detection performance of the proposed Capsule Network over the attacks such as Deepfake, Face2face and FaceSwap had produced significant results.

Keywords: - Deepfake, Capsule Networks, CNN, Fake News, VGG-19

1. INTRODUCTION

The manipulation of images and videos with advanced techniques paved the way for the creation of forged images/videos. Initially, the manipulation of images and videos had been made for enhancement, but nowadays manipulations had led to changes in the identity of the person. With the advancement of the internet and social media, manipulation tools form a threat to the authenticity of images/videos. High-quality manipulated images and videos were created using deep learning techniques. Hence, with the use of these manipulation techniques, people may create fake videos most often and share them over social media leading to security problems. An effective way of communicating is visual media such as images and videos, as they can provide information effectively. There are various tools available that can be used to manipulate images and videos. By using these techniques, people may hide crime incidents, and defame a person and reputed organizations [1]. The quality of the manipulated images/videos had improved significantly with the advancement of machine learning and deep learning techniques. Nowadays, the possibility of creating fake videos in a shorter duration has become quite easy [2].

There are several measures designed to detect fake images and videos such as Convolutional Neural networks (CNNs), Long Short-Term Memory (LSTM), Recurrent Neural networks (RNN), etc. [3]. There were methods applicable for the images and video separately. Singular

Value Decomposition (SVD) is used in classifying the image as fake or real [4].Malicious people manipulate the photos of the people to produce the court for fake evidence creation [5]. In general, images will be edited using Photoshop software, which may appear like real images due to pixelization [6]. People were able to manipulate the videos of shorter duration, modifying their facial expressions [2]. In video-based methods, manipulations were made on face swapping, lip synchronization, and head/eye movement synchronization. Video-based methods may perform better than image-based methods but they are restricted to particular attacks. Video detection techniques may fail due to the proper manipulations made in facial expression, eye movement, and audio-video synchronization. Hence, an image-based approach has been taken in the proposed work to detect forged images.

Generally, manipulations were done manually using photo editing tools such as Photoshop, making it a tedious process and change can be identified by naked eyes. With the introduction of machine learning tools, these manipulations had been achieved efficiently without any manual work. Convolutional Neural Networks (CNN) had been used in the detection of fake images and videos. In order to improve the performance of the CNN, more connections may be added. But with the increase in the number of connections, the size of the network grows leading to computation complexity and also requires more training data. To overcome this problem Capsule Network had been introduced to detect fake images and videos. Capsule Networks had been efficient in detecting the manipulated videos and images [7]. Improvements in the capsule network shall be made to achieve better performance. The designed network will aim to reduce the complexity and maps the spatial features. The proposed FCNM model includes;

(i) Exponential Linear Unit (ELU) had been used for the activation function for lowering the computational complexity.

(ii) LP Pooling layer used with the statistical layer for the mapping of statistical and spatial feature maps.

2. RELATED WORKS

The advancement in technology can make people create manipulated images with a high level of accuracy. Video is the collection of images known as frames. The fake detection process begins with the frame extraction and saves it in image form. After extracting the frame, the process of fake detection remains the same for both video and image. Detecting Deepfake has been challenging and different methods were proposed to detect their changes. There had been several techniques proposed for Deepfake detection such as face swapping, face2face, etc. Neural Networks were trained initially to detect fake and real images [8].

Pre-trained CNN is used for fake detection by extracting the whole image instead of the face region [9]. Support Vector Machine (SVM) is used in classifying the features extracted by CNN [10]. People were defamed by modifying their faces using visualization and Deepfake techniques [11]. Manipulation techniques such as automatic face swap were made using Video Face Replacement methods [12]. The facial expression had been altered in a realistic manner mapping with their mouth movements using VDub [13]. In general face image synthesis approaches had been proposed using deep learning techniques such as Generative

Adversarial Networks (GANs), for altering the face looks and attributes like skin color, shape, etc. [14].

CNN architectures with the pooling layer compute statistical functions such as mean, and variance to obtain improved performance in the network for fake detection [15]. CNN had been used with InceptionNet for detecting face tampering in the videos [16]. CNN-based models had been efficient in facial features extraction compared to the traditional methods [17]. In order to improve the performance of CNN, a signal enhancement layer had been used in the CNN structure to detect the fake image [18]. A new convolution layer had been included in CNN for making it learn the detection features [19]. The limitations of the CNN include more memory requirements and a large amount of training data.

Capsule Networks had been used for Deepfake detection to overcome the drawbacks of CNN. It has a more robust architecture comprising many capsules. The Capsule network had achieved better performance in object classification compared to the CNN [20]. The proposed work aims to make Deepfake detection using the capsule network. The routing process enables each capsule to maintain information obtained from earlier capsules and makes classification by information comparison. Hence, Capsule Network can save the orientation and location of components in an image.

3. FAKE CAPSULE NETWORK MODEL FOR DEEPFAKE DETECTION

Detection of Deepfake videos had received a lot of attention in recent years and the research on efficient detection of misleading videos is still lacking. In most cases, CNN had been used for the detection of fake videos and images. A Capsule network had been efficient which can store the information in vector values rather than scalar values. These capsule vectors represent the richer information in the architecture. It has an equivariance property that reduces the training data required. The proposed Fake Capsule Network Model is capsule network-based Deepfake detection as described in figure 3 comprises three steps: In the first step, the given input video sequence is extracted into frames and saved in image form for processing. Facial areas are cropped and sent to the capsule network for classification in the second step. In the third step, after attaining the detection results average scores of frames were taken for the final output.

The Capsule network for Deepfake detection comprises of VGG-19 feature extractor, capsules, and output capsule. Random weight initialization had been made; hence they will have different behaviors after training. The number of primary capsules in the network is taken as 10.VGG-19 is used for the feature extraction using pre-trained CNN, which is trained on ImageNet. Pre-trained CNN is used in extracting meaningful features from the new samples presented.

Primary Capsule

The primary capsule comprises 2D convolution which slides over the given 2D input data. Batch normalization is applied to make a layer to learn independently. Exponential Linear Unit (ELU) activation function had been applied, that has negative values to push the mean unit activation function near to zero with lower computational complexity.

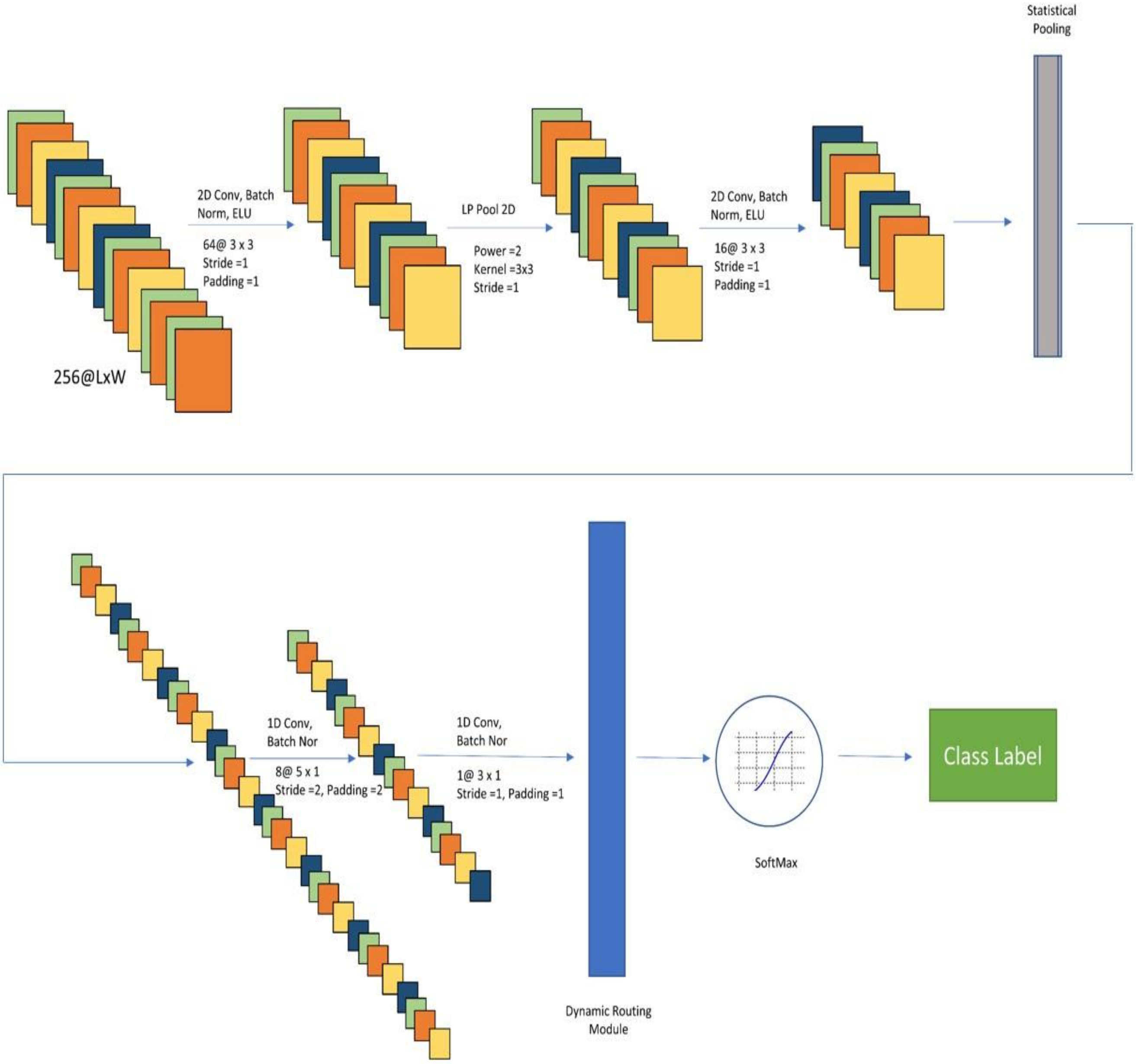

Figure 1 Proposed Fake Capsule Network Model (FCNM) for Improved Deepfake Detection

The Exponential Linear Unit activation function E(x) is defined by the expression (1) as follows,

$$E(x) = \begin{Bmatrix} x & x > 0 \\ \alpha.(e^{x} - 1) & x < 0 \end{Bmatrix} \quad (1)$$

LP Pooling layer is added which takes power average pooling over the input signal composed of several input planes.

$$f(Z) = \sqrt[a]{\sum_{z \in Z} Z^a} \tag{2}$$

At a = ∞, Max pooling and a=1, average pooling will be obtained. LP pooling will be a fine tuning between max pooling and average pooling. LP pooling layer had been integrated with the statistical pooling layer. Statistical pooling layer is used for

detecting the computer-generated images by using the statistical differences between real and fake images. Mean and variance of the statistic function is used in each filter for differentiating the fake and real videos.

$$\text{Mean (x):} \quad x_k = \frac{1}{Y \times Z} \sum_{i=1}^{Y} \sum_{j=1}^{Z} T_{kij} \tag{3}$$

$$\text{Variance } (\alpha)\text{: } \alpha_k^2 = \frac{1}{Y \times Z - 1} \sum_{i=1}^{Y} \sum_{j=1}^{Z} (T_{kij} - x_k)^2 \tag{4}$$

In the expression (3) and (4), k denotes the layer index, T - denotes two-dimensional kernel, Y and Z are length and width of the filters.

After differentiating the real and fake video sequences by the statistical pooling layer, features are passed to 1D convolution layer and subsequent batch normalization is applied. Dynamic routing algorithm is used to fuse the features obtained from different primary capsules. The weights of the capsule are initialized differently; hence each capsule will learn different features for the given same input. Features from different capsules need to be fused together for prediction of fake and real videos. Dynamic routing algorithm makes this fusion dynamically. Then, it is passed to the output capsule for binary classification. After the binary classification of fake and real capsule, softmax function had been used for finding the probability distribution. Then, mean of the softmax function is taken to determine the final output.

$$S = \frac{1}{n} \sum_{i=1}^{n} softmax\ (b_i^{(1)}, b_i^{(2)}, b_i^{(3)}, \ldots \ldots . b_i^{(n)}) \tag{5}$$

S is the predicted probability of the softmax function.

4. RESULTS AND DISCUSSION

FaceForensics++ dataset and Google Deepfake Detection (DFD) datasets were used to test the performance of the proposed FCNM Model. For training, FaceForensics++ datasets were used that includes original and manipulated videos. Google DFD dataset was used in testing. C23 compression form in the dataset was used in the Deepfake detection.

Table 1 FaceForensics++ Dataset for Training and Testing

Method	Training set (Videos)	Validation set (Videos)	Test set (Videos)
Real	720	140	140
Deepfake	720	140	140

Face2face	720	140	140
FaceSwap	720	140	140

Table 1 comprises number of videos taken from FaceForensics++ dataset for training, testing and validation. 720 videos for training, 140 videos for validation and 140 videos for testing was taken from different manipulation techniques such as Deepfake, Face2face and FaceSwap, where Real indicate the videos that are original. The results of the proposed FCNM for Deepfake detection had shown improved performance in comparison to the existing approaches. By including the Exponential Linear Unit (ELU) activation function in the capsule network, the mean activation had been obtained with lower computation.

Table 2Performance Accuracy of the Feature Extraction

Feature Extractor	Training Accuracy (%)	Validation Loss (%)	Number of Parameters
Proposed Model			
VGG-19 + Proposed FCNM	99.86	3.1461	15,70,430
Existing Model			
VGG-19 + Existing Capsule Network	99.78	3.1940	23,25,500
Resnet-50 + Existing Capsule Network	99.50	4.152	2,40,357

Performance accuracy of the feature extractor had improved significantly as described in the table 2. Compared to the existing approaches such as Resnet-50 and VGG-19, proposed method had improved performance with lesser number of parameters. Pre-trained CNN and proposed capsule structures are used for the feature extraction in the proposed approach which resulted in better performance.VGG-19 feature extractor have higher accuracy with lesser number of parameters. By varying the number of primary capsules, performance improvement can be achieved.

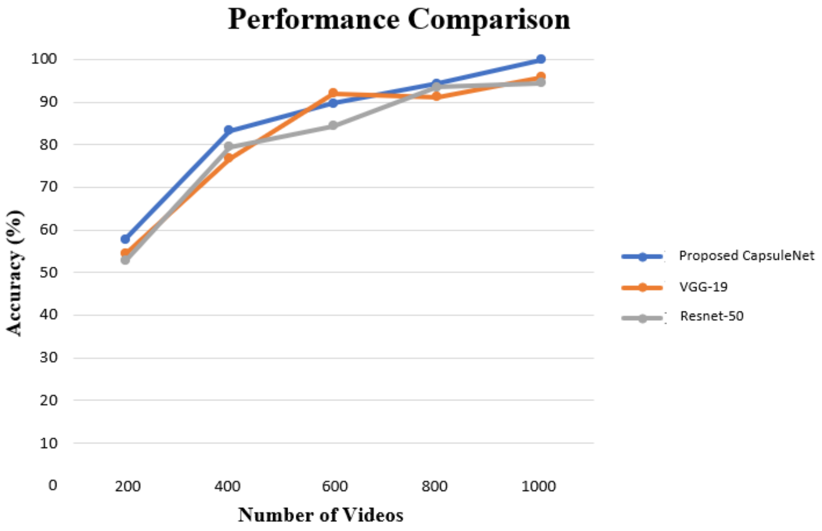

Figure 2 Performance Comparison of FCNM with Existing Approaches

In the figure 2, performance accuracy in comparison to the existing approaches was made. The proposed FCNM model outperforms the existing approaches in detecting the fake videos. With the varying number of videos, fake video detection accuracy had been examined. In the training phase, all the existing models were efficient in detecting the fake videos.

Table 3 Performance Accuracy Comparison in Primary Capsule with Statistical Pooling Layer and LP Pooling layer

Settings	Test Accuracy (%)	Error Rate (%)	Number of Parameters
Statistical Pooling Layer	92.00	10.64	15,71,070
LP Pooling Layer and Statistical Pooling Layer	95.36	9.75	15,70,430

Performance accuracy comparison in the primary capsule had been made with the statistical pooling layer and statistical pooling layer with LP Pooling layer as shown in the table 3. Initially statistical pooling layer used in differentiating the fake and real video sequence, but the integration of LP Pooling layer with statistical pooling layer used to make power average pooling for the given input signal.

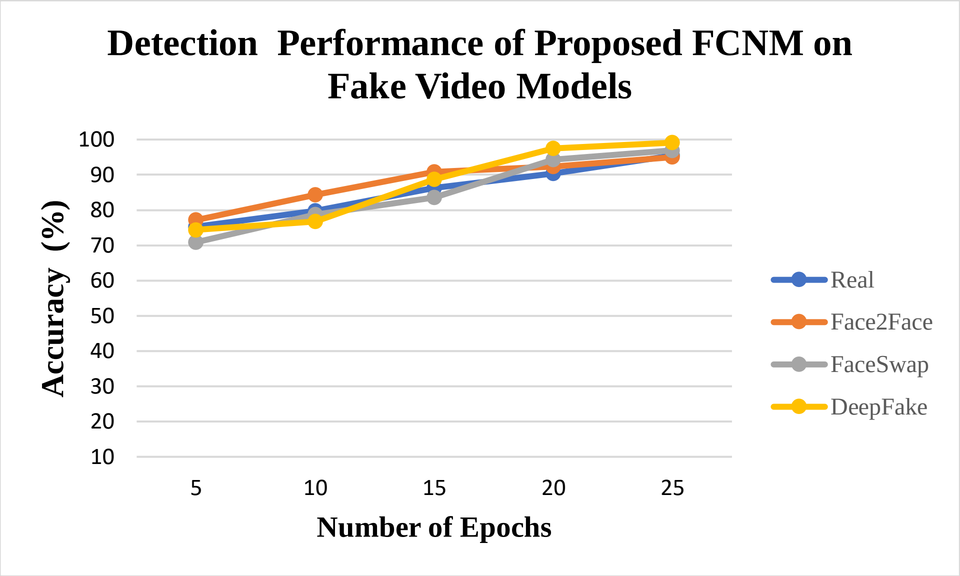

Figure 3 Performance Accuracy of FCNM on Fake Video Models

The graph shows the detection performance of FCNM models on fake videos and real videos as described in figure 3. The result shows the model detection accuracy against the videos that are created by fake video creation models such as Deepfake, Face2face and FaceSwap, and it also shows the accuracy of detection for real videos. The model has higher accuracy of detection of fake videos.

5. CONCLUSION

A modified capsule network FCNM was proposed and found to outperform other state-of-the-art techniques. By using the proposed capsule network, memory and computation power has been saved. Even in the compressed videos, the Deepfake detection performance of capsule network had been high compared to existing models and CNN. Also, the improved performance is achieved with lesser number of parameters. Further the study can be extended to multi-face Deepfake detection from the crowd videos utilizing the flexibility and robustness of vision transformers.

REFERENCES

1. Nabi, S.T., Kumar, M., Singh, P. *et al.* (2022). A comprehensive survey of image and video forgery techniques: variants, challenges, and future directions. *Multimedia Systems* 28, 939–992. https://doi.org/10.1007/s00530-021-00873-8.
2. Thies, J, Zollhofer, M, Stamminger, M, Theobalt, Cand Niebner, M. (2016) Face2Face: real-timeface capture and reenactment of RGB videos. In: Conference on computer vision and pattern recognition (CVPR). IEEE.
3. Guera, D. and Delp, E.J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, 27-30 November 2018, 1-6.https://doi.org/10.1109/AVSS.2018.8639163.

4. El Abbadi, Nidhal AL-Rammahi, AdilMudher, et al.(2013). Blind Fake Image detection. *IJCSI International Journal of Computer Science Issues*, Vol. 10.
5. D. Strigl, K. Koflerand S. Podlipnig, (2010). Performance and scalability of GPU-based convolutional neural networks. In 18th Euromicro Conference on Parallel, Distributed, and Network-Based Processing.
6. M. D. Ansari, S. P. Ghrera and V. Tyagi. (2014). Pixel-based image forgery detection: A Review. *IETE Journal of Education*, 55(1), 40–46.
7. H. H. Nguyen, J. Yamagishi and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307-2311, doi: 10.1109/ICASSP.2019.8682602.
8. M. Villan, A. Kuruvilla, K. J. Paul and E. P. Elias. (2017). Fake Image Detection Using Machine Learning. *IRACST International Journal of Computer Science and Information Technology & Security (IJCSITS).*
9. Davide Cozzolino, Giovanni Poggi and Luisa Verdoliva, (2017).Recasting residual-based local descriptors as Convolutional neural networks: an application to image forgery detection," in *IH&MMSEC. ACM*.
10. Jianwei Yang, Zhen Lei and Stan Z Li (2014). Learn Convolutional neural network for face anti-spoofing," arXiv preprint arXiv:1408.5601.
11. Deepfakes github. https://github.com/deepfakes/faceswap.
12. Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, et al. (2011). Video face replacement. ACM Trans. Graph., 30(6):130:1–130:10.
13. Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, et al.(2015).Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Computer Graphics Forum,* 34(2):193–204.
14. GrigoryAntipov, MoezBaccouche and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In IEEE International Conference on Image Processing,2017.
15. Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, andIsao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In IEEE Workshop on Information Forensics and Security, pages 1–6, 2017.
16. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. 2017.
17. Sun K, Xiao B, Liu D, and Wang J. (2019). Deep high-resolution representation learning for human pose estimation. In: CVPR.
18. Yang, P., Ni, R. and Zhao, Y. (2017). Recapture image forensics based on Laplacian convolutional neural networks. In International Workshop on Digital Watermarking. 119–128 (Springer).
19. Bayar B, and Stamm MC. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Workshop on information hiding and multimedia security (IH&MMSEC). ACM.
20. Xiang C, Zhang L, Tang Y, at al. (2018).Ms-capsnet: a novel multi-scale capsule network. IEEE Signal Process Lett 25(12):1850–1854.