FvFc-Net: Forged Video Frame Classification Network

Santanu Das[1], Sourav Dey Roy[1], Priya Saha[2] and Mrinal Kanti Bhowmik[1]

[1] Tripura University (A Central University), Suryamaninagar-799022, India
[2] National Forensic Sciences University (Tripura Campus), Agartala-799006, India
santanud803@gmail.com, souravdeyroy49@gmail.com,
priyasaha.cse@gmail.com, mrinalkantibhowmik@tripurauniv.ac.in

Abstract. In the era of rampant digital manipulation, the need for robust forgery detection techniques for verifying the authenticity of the digital contents is imperative. Even though forgery detection in the images has received significant attention, the success in case of video de-forging (i.e., video manipulation detection techniques) has been less explored. This is because there exist a certain ghost phenomenon in the manipulated video that occurs when videos are altered and this makes the anti-forensic measures unsuitable for verifying their authenticity. Ghost phenomena in video manipulation refers to the artifacts created by multiple compression and decompression processes, resulting in faint, residual images or trails of previous frames appearing in subsequent frames. These artifacts manifest as translucent duplicates or blurs, degrading the quality and fidelity of the video. Despite the lack of appropriate video forensic techniques, the paper highlights and analyses the prevalence of ghost phenomenon associated with single video de-forging. Our objective is to provide insights into the ghost phenomenon associated with single video de-forging and thereafter propose a novel approach for forgery classification leveraging Deep Convolutional Neural Networks (DCNN). Our proposed method utilizes automatically learn discriminative features from manipulated images, enabling accurate classification of authentic and forged content. Evaluation on our own created dataset and available REWIND dataset showcases the superior performance of our approach, achieving an impressive accuracy of 96.73% and 95.31% respectively. This indicates the effectiveness of our method in discerning subtle manipulations, making it a valuable contribution to the field of multimedia forensics.
Keywords: Multimedia Forensics, Video Forgery, Splicing, Ghost Phenomenon Analysis, Convolutional Neural Network, Classification, Performance Evaluation.

1 Introduction

In the digital age, the proliferation of image editing software has made it increasingly challenging to distinguish between authentic and forged images [1]. This issue has significant implications across various domains, including journalism, forensics, and cyber security [2]. The emergence of Deep Convolutional Neural Networks (DCNNs) has provided a promising avenue for addressing this challenge by enabling automated image classification with remarkable accuracy and efficiency. Consequently, the ability to discern between authentic and manipulated images is crucial in numerous contexts. In journalism, the dissemination of fake news or manipulated images can mislead the public and erode trust in media outlets. Forensic investigations rely on accurate image analysis to establish evidence authenticity Thus, developing reliable techniques for forged image classification holds immense societal and technological significance. In recent years, significant advancements have been made in forged image classification using DCNNs [3]. Researchers have explored innovative architectures, optimization techniques, and training methodologies to enhance model performance and generalization capabilities. Beyond technical advancements, the ethical and societal implications of forged image classification must also be carefully considered. The potential misuse of such technology for censorship, surveillance, or ma-

nipulation underscores the importance of responsible deployment and regulatory frameworks to safeguard against misuse and protect individual privacy and freedom of expression.

Even though, the success of the verifying authenticity of the images using conventional detection methods [4] [5] [6] [7] are being praised in forensic communities but success of these methods has been observed to be limited for video de-forging techniques (i.e., forensic methods for video manipulation detection). This is because the democratization of video manipulation, while empowering content creators, has also given rise to an alarming trend – the emergence of ghost phenomena within manipulated videos. Ghost phenomena refer to subtle yet perceptible artifacts and inconsistencies that manifest when videos are altered or manipulated. These anomalies, often overlooked by the untrained eye, can range from unnatural lighting and shadows to distortions in motion and background elements. Such imperfections pose a significant challenge, not only for forensic experts and digital media analysts but also for society at large, as they have the potential to mislead, deceive, and manipulate the truth. Depending upon this phenomenon, the primary contribution of this paper are:

1. Firstly, the paper investigates and analyses the intricate realm of ghost phenomena during the creation of spliced videos. By addressing these challenges, our study contributes to the ongoing efforts in advancing the field of digital forensics and bolstering the reliability of multimedia content in an age where misinformation and manipulated media proliferate.
2. Secondly, to cope up with such challenging conditions of video based forgery classification task, the paper proposes a novel framework for classification of forged and authentic frames pertaining to the manipulated videos. The proposed framework combines the CNN based features from the holistic frames and ELA (Error Level Analysis) based CNN features for effective classification of forged and authentic frames present in the video. For an effective forgery classification task, the paper further proposed a novel Forged Video Frame Classification Network (FvFc-Net) using the building blocks of convolutional neural networks (CNNs).
3. Thirdly, the experimental results on our own created dataset and publicly available REWIND dataset has been investigated and showcases the superior performance of our proposed framework for forgery classification task.

Paper Outline. In Section 2, review on the existing methods for forgery classification task was elaborated. Section 3 describes and analyses the observable ghost phenomena of the manipulated videos which deteriorates the performance of the conventional methods for forgery classification. In Section 4 our proposed network for forged frames classification from video has been described. The evaluation results of our proposed network is reported in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

During the last few decades, various methodologies have been explored to accurately detect forgery within images. Here, we discuss the existing literature and approaches in this domain. Shen et al. (2017) introduced the Textural Features Based on the Grey Level Co-occurrence Matrices (TF-GLCM) method, focusing on texture features and utilizing the CASIA v2.0 dataset for evaluation. Similarly, Jaiswal et al. (2019) proposed a technique employing Resnet-50 with three distinct classifiers (Naïve Bayes, K-nearest neighbour, and Multi-Class Model using Support Vector Machine (SVM) Learner), demonstrating their effectiveness on the CASIA v2.0 dataset. Tiwari et al.

Table 1. Literature Review on State-of-the-Art Authentic/Forged Classification Methods

Author	Published in, Years	Proposed Method	Dataset Used	Performance & Discussion
X. Shen, et al. [8]	IET Image Processing, 2017	TF-GLCM	CASIA v2.0	The technique demonstrated an impressive classification accuracy of 98.54%.
A. K Jaiswal, et al. [9]	International Conference on Advanced Computing and Software Engineering (ICACSE), 2019	Resnet-50 with three different classifiers Naïve Bayes, K-nearest neighbour and Multi-Class Model using SVM Learner	CASIA v2.0	The proposed approach has attained classification accuracy across three distinct classifiers: Naïve Bayes (0.5991), Multiclass model using SVM Learner (0.7026), and K-Nearest Neighbour (0.5991).
S. K. Tiwari, et al. [10]	International Conference, PReMI 2019	CNN with hierarchical agglomerative clustering (HAC)	Dresden	The proposed approach has attained an impressive accuracy rate of 94.90% in classifying forged images.
N. Y. Hussien, et al. [11]	International Journal of Sociotechnology and Knowledge Development, 2020	DBN-DNN with PCA	Columbia Dataset	The proposed classifier is capable of determining whether an image is authentic or forged with an impressive classification accuracy of 98.20%.
D. Baleanu, et al. [12]	Computers, Materials & Continua, 2023	GKSWF with SVM Classifier	CASIA 2.0	The proposed method aims to retain image information in smooth areas while detecting signs of tampering in textural details. The method achieved an impressive accuracy rate of 98.60%.

(2019) adopted a CNN approach coupled with hierarchical agglomerative clustering (HAC) for forged frame classification, with experimentation conducted on the Dresden dataset. Furthermore, Hussien et al. (2020) introduced a method utilizing Deep belief network with Deep Neural Network (DBN-DNN) with Principal component analysis (PCA) on the Columbia Dataset, showcasing an alternative approach to tackle the classification task. In a recent study, Baleanu et al. (2023) proposed the generalized k- symbol Whittaker function (GKSWF) technique integrated with an SVM Classifier for forged frame classification, demonstrating promising results on the CASIA 2.0 dataset. While these existing methods have contributed significantly to the field, there remains room for improvement and exploration of novel techniques. However, the literature review presented in Table 1 indicates a notable deficiency in forgery detection methods, particularly in the domain of video, specifically within the context of forged frame classification. In this paper, to address the gap in the literature we present the Forged Frame Classification using Deep Convolution Neural Network (FvFc-Net), offering a unique approach to effectively detect forged frames within videos.

3 Ghost Phenomena Analysis of Manipulated Video

In the realm of literature, numerous conventional forensic techniques [4] [5] [6] [7] exist for detecting forgery in images. However, when it comes to videos, the task of forgery detection, especially identifying manipulated areas within video frames, poses a significant challenge. In this section, analysis of the phenomena (referred as ghost phenomena) mostly responsible for deteriorating the performance of the splicing detection methods. The **"Ghost phenomena"** refers to the imperceptible loss or degradation of data that occurs during video editing and video-to-frame conversion processes. This phenomena is characterized by the subtle disappearance of manipulated information within the multimedia contents, rendering it invisible to the naked eye due to certain level of editing. Despite the apparent visual fidelity, the compression algorithm selectively discards certain details, creating a spectral effect where the lost data becomes elusive or **'ghost-like.'**

There are numerous ways to manipulate the visual content, and new tools and methods are proposed by the day. When we are trying to splice some objects in a video first

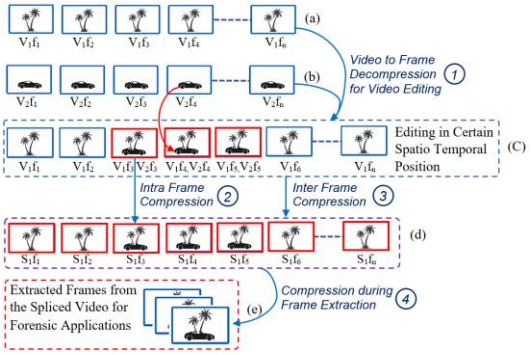


Fig 1. Overall Schematic Diagram of Creation of Forged Video (object spliced Video); (a) and (b) are the two different authentic videos; (C) Spliced the object of the second video into the first video; (d) Forged/Spliced Video; (e) Extracted Frames of Spliced video

we have to import two videos into a timeline of video editing tools where all the frames are situated by maintaining a sequence. The jagged diagrammatic representation of the creation of the manipulated video (i.e., in terms of object splicing) is demonstrated in Fig. 1. Here, V_1 and V_2 represents the two authentic videos and f_1, f_2 ...f_n represents their corresponding frame sequences. Therefore, when manipulation of the video is performed, firstly two videos (V_1 and V_2) need to be imported into the timeline of the video editing tools where all the frames are located by maintaining the sequences as shown in Fig. 1. Then the object of interest from the video frames of V_2 is extracted and placed in the corresponding video frames of V_1. Thereby maintaining a frame sequence that helps to make forged/spliced video (S_1) more realistic. Secondly, to make the manipulated video more realistic, we have also adjusted the color balance of the spliced objects according to the background video where the spliced object has been pasted. After the completion of the all manipulation process, the spliced/forged video (S_1) is exported in a widely used common video format (i.e., .MP4, .AVI, etc.). Therefore, video compression in the context of video editing and export involves two primary types i.e., Intra-frame compression and Inter-frame compression [13]. However, to make the applicability of the image based conventional blind detection methods on the manipulated video, the video is converted into frames. Therefore, multiple times of data loss happens in four different phases to make a spliced/tempered video as shown in Fig. 1 i.e., (1) During video to frames decompression in timeline of video editing tools before performing splicing, (2) Intra Frame Compression during exporting the manipulated video, (3) Intra Frame Compression during exporting the manipulated video, and (4) During extracting the frames from the manipulated video for performing the conventional forensic methods.

Due to these multiple compression as shown in Fig. 1, all the artifacts of the image sequences or frames are suppressed and thereafter the digital forensic methods especially designed for image splicing detection fails to quantify the authenticity of a video. Generally, conventional forensic techniques depend on defining appropriate features that aid in distinguishing between unaltered and altered images and then a classifier is trained on a large number of images of both the categories. Typically, within the JPEG compression algorithm, the quality factor is denoted by a numerical value between 1 and 100, where 1 indicates the poorest quality and 100 signifies the highest quality [14]. The quality factor of JPEG compression depends on the compression level, typically ranging from 2 to 5. A compression level of 2 indicates nearly lossless compression, maintaining the original image quality. On the other hand, a compression level of 5 implies lossy compression, resulting in a lower quality factor for the JPEG images. For effective understanding of the existence of the ghost phenomena in the manipulated video, Fig. 2 analyses the DCT coefficients based anomaly mask for different quality

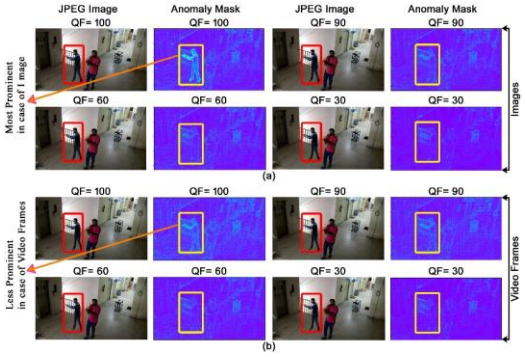


Fig 2. Analysis of DCT Coefficients Based Anomaly Mask; (a) Object Spliced in Image; (b) Object Spliced in Video Frames

factors (QF) of manipulated JPEG image and a JPEG frame of a manipulated video thereby maintaining a certain timestamp (i.e., QF= 30, 60, 90, 100). It can be observed from Fig. 2 (a) that in the case of the JPEG image with QF=100, the anomaly mask (surrounded by respective yellow bounding box) of manipulated region within the holistic JPEG image is prominent and visible thereafter applying DCT. Also, it can be observed from Fig. 2 (a) that as the quality factor of the JPEG compressed image decreases, the visibility of the manipulated region also gradually decreases. Therefore from the analysis it can be observed that when the compression is less with high quality factor, the manipulated part of the image can be easily detected and therefore can generate a mask for spliced/ forged objects. Consequently, when the compression is high with low quality factor, the manipulated part of the image cannot be well detected in the DCT coefficients based anomaly mask. This is because multiple compression occurred at the frame level during creation of the manipulated video at the timeline of the video editing tools. Conversely, for a JPEG frame extracted from manipulated video with different quality factors as displayed in Fig. 2 (b), the anomaly mask for the manipulated regions (i.e., surrounded by yellow bound bounding boxes) are not prominent for QF=100 and the same is observed for all the other considered quality factors. This is because of multiple compression factors occurred in the spliced frames so as mentioned in Fig. 1. To mitigate this challenge, a novel forgery detection method is needed that can adapt to the semantic features representation of the spliced regions within the frames of the video clips. Depending upon this phenomena, in this paper, a novel Forged Video Frame Classification Network (FvFc-Net) is proposed to classify the forged frames in the video so as elaborately described in the below sections.

4 Methodology

In this section, our proposed framework incorporating "FvFc-Net (Forged Video Frame Classification Network)" for forgery classification among frames present in the video is elaborately described. The overall diagrammatic representation of the proposed architecture for forgery classification in video is displayed in Fig. 3. The proposed architecture consists of five building blocks: Frame Extraction Layer, Error Level Analysis (ELA) [17], Feature Extraction Layer, Feature Fusion Module and Classification Layer.

4.1 Frame Extraction Layer

In this layer we have extracted frames I from the input video V to feed it into Deep Convolutional Network (DCNN). Here, I is the set of video Frames where $I = I_1, I_2, I_3, \ldots\ldots\ldots I_n$.

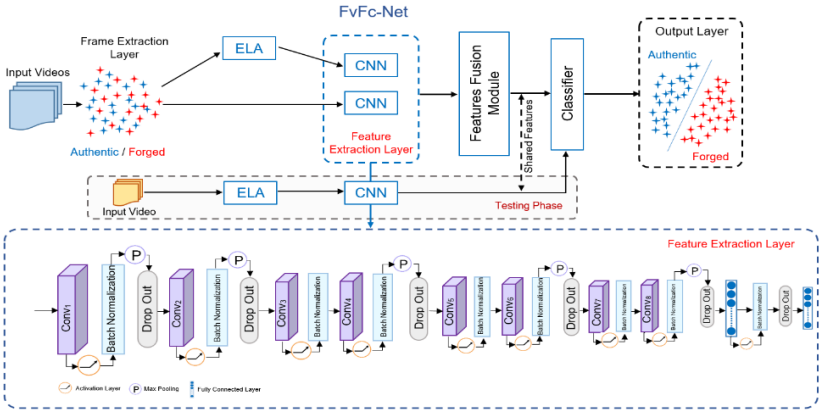


Fig. 3: Proposed Framework for Forged Video Frame Classification Network (FvFc-Net)

4.2 Error Level Analysis (ELA)

Error Level Analysis (ELA) is a valuable technique for detecting digital image forgeries, particularly in the context of forged frame classification using Convolutional Neural Networks (CNNs). In the proposed work, we leverage the ELA (Error Level Analysis) technique [17], which has traditionally been utilized to identify noise in authentic images concerning their background. According to the literature, original images captured by digital cameras should have high ELA values. However, in the case of each resave (i.e., recompression) of the same image, the potential error rate will decrease. But in the case of tampered/modified (in the case of object/region splicing) images, the modified/spliced region will exhibit higher ELA values compared to the background or authentic regions [18]. When combined with CNNs, it can significantly enhance the accuracy of forgery detection. In the realm of forged frame classification, ELA serves as a preprocessing step to highlight potential areas of manipulation within video frames. By analyzing the error levels across different regions of a frame, ELA provides a visual cue to CNNs, guiding them towards areas that are likely to be tampered with. This approach enhances the CNN's ability to distinguish between authentic and forged frames by focusing its attention on regions where manipulation is more prevalent. Moreover, integrating ELA with CNNs in forged frame classification offers a robust solution capable of detecting various forms of forgery, especially in splicing. The synergy between ELA and CNNs empowers forensic investigators with a powerful toolset for accurately identifying forged frames within digital video content. In our proposed work, the ELA takes the input of I and generates I_e.

4.3 Error Feature Fusion Module (FFM)

In this module, both the feature maps $(F_M I_r, F_M I_e)$ has been concatenated i.e., $\prod FF_M = F_M I_r \oplus F_M I_e$. The feature level fusion facilitates the detection of correlated feature values extracted by the feature extraction module, thereby identifying a concise set of salient features that enhance classification accuracy.

4.4 Classification layer

Finally, the $\prod FF_M$ is propagated through this layer, where the sigmoid function plays a crucial role in classifying each frame into one of two distinct classes: Authentic (Label 0) or Forged (Label 1).

Table 2: Description of the FvFc-NET Architecture for Classifying Forged Frames in Videos, Presented Layer by Layer

Layer Type	Input	Kernel Size	No. of Filters	Stride	No. of LPs in Each Layer
INPUT	224×224×3	-	-	-	-
$CONV_1$	224×224×3	3×3	32	1	896
$RELU_1$	224×224×32	-	-	-	0
BN_1	224×224×32	-	-	-	128
$POOL_1$	74×74×32	2×2	-	2	0
		DROPOUT (25%)			0
$CONV_2$	74×74×32	3×3	32	1	9248
$RELU_2$	74×74×32	-	-	-	0
BN_2	74×74×32	-	-	-	128
$POOL_2$	24×24×32	2×2	-	2	0
		DROPOUT (25%)			0
$CONV_3$	24×24×64	3×3	64	1	18496
$RELU_3$	24×24×64	-	-	-	0
BN_3	24×24×64	-	-	-	128
$CONV_4$	24×24×64	3×3	64	1	36928
$RELU_4$	24×24×64	-	-	-	0
BN_4	24×24×64	-	-	-	256
$POOL_3$	12×12×64	2×2	-	2	0
		DROPOUT (25%)			
$CONV_5$	12×12×128	3×3	128	1	73856
$RELU_5$	12×12×128	-	-	-	
BN_5	12×12×128	-	-	-	512
$CONV_6$	12×12×128	3×3	128	1	147584
$RELU_6$	12×12×128	-	-	-	0
BN_5	12×12×128	-	-	-	512
$POOL_4$	6×6×128	2×2	-	2	0
		DROPOUT (25%)			
$CONV_7$	6×6×128	3×3	128	1	147584
$RELU_7$	6×6×128	-	-	-	0
BN_6	6×6×128				512
$CONV_8$	6×6×128	3×3	128	1	147584
$RELU_8$	6×6×128	-	-	-	0
BN_7	6×6×128	-	-	-	512
$POOL_5$	3×3×128	2×2	-	2	0
		DROPOUT (50%)			
FC_1	200×1	-	-	-	230600
$RELU_9$	200×1	-	-	-	0
BN_8	200×1	-	-	-	800
		DROPOUT (50%)			0
FC_2	2×1	-	-	-	402
$RELU_{10}$	2×1	-	-	-	0
		SIGMOID ACTIVATION WITH TWO CLASSES (2×1)			

$Conv_n$- n^{th} Convolution Layer; $RELU_n$: n^{th} Activation Layer; $Pool_n$: n^{th} Pooling Layers, FC_n- n^{th} Fully Connected Layer; BN_n- n^{th} Batch Normalization, LPs: Learnable Parameters

4.5 Architecture Description of our Proposed FvFc-Net

The Convolutional Neural Network (CNN) stands out as a remarkably efficient method for identification, incorporating an Artificial Neural Network (ANN) structured with multiple sequential layers: convolution layer (conv), pooling layer (pool), and fully connected layers (FC). The CNN takes in the entire video frame as input, and its output layer comprises numerous neurons, each representing a distinct class. In this paper, we have employed the ELA [17] technique to enhance classification accuracy. The block diagram illustrating the proposed DCNN architecture, dubbed "FvFc-NET (Forged Video Frame Classification Network)", for discerning between forged and authentic frames within videos, is depicted in Fig. 3. In designing the FvFc-NET, we adopted two-part approaches within the CNN network. Initially, we feed the input data in its raw form I_r into a CNN to extract features. Subsequently, we perform ELA on the input data I_e and pass it through another CNN to extract additional features. Finally, both sets of features $F_M I_r$, $F_M I_e$ obtained from the Feature Extraction Layer are passed through the Feature Fusion Module (FFM). The FFM will concatenate both the feature

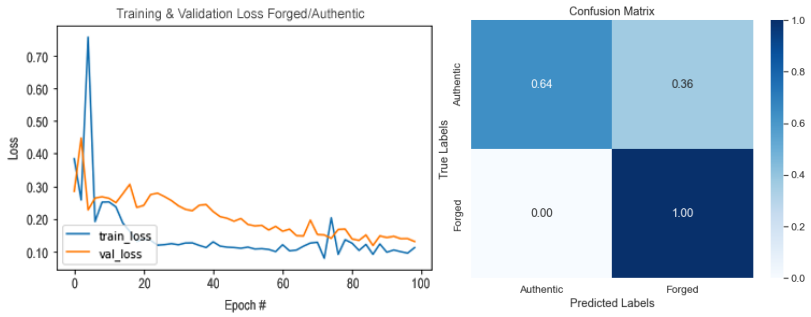
Fig. 4. Some of the output of our proposed FvFc-Net (a) and (b) are the frame sequences of our own designed dataset; (c) and (d) are the Frame sequences of the REWIND Dataset

maps $F_M I_r$, $F_M I_e$ along the channel dimension i.e., $\prod FF_M$, where $\prod FF_M = F_M I_r \oplus F_M I_e$ and forwarded to the classification layer. Here, $\oplus$ represents element wise addition. While testing, $\prod FF_M$ will be shared for better classification. The classification layer evaluates all outputs from the FFM layer, providing a binary classification (i.e., Authentic or Forged). If a significant variance is detected, the frame is classified as forged O_f ; otherwise, it is labeled as authentic O_a .

The description of each layers and number of learnable parameters associated with our proposed FvFc-NET architecture for classification of video frames on our created dataset are summarized in Table 2.

4.6 Training of proposed FvFc-Net Architecture

We have implemented our proposed architecture using the Python platform, specifically leveraging the TensorFlow and Keras libraries. This implementation was conducted on a system equipped with an Nvidia TITAN XP GPU and 64 GB of installed RAM. To train the FvFc-Net for forged frame classification, 110000 Frame sequences are arbitrarily selected from own created forged video dataset. In order to produce the forged videos, we have captured videos using a variety of camera models, such as the Nikon D5100 and the FLIR T650sc. These videos contains indoor and outdoor scenes. To bolster our dataset, we've captured videos under both daylight and nighttime conditions, ensuring a more comprehensive and robust collection. The spliced video clips are generated through a multi-step process. Initially, the background is eliminated from the authentic videos. The salient objects are then extracted using the background subtraction method facilitated by the Rotobrush tool [19] within Adobe After Effects 2022 [20]. Finally, these salient objects are seamlessly in targeted or “spliced” into other authentic videos, resulting in the creation of the manipulated video clips. To enhance learning, training frames are randomly shuffled before inputting them into the network. The frame sequences are split into training and validation sets in a 4:1 ratio. The learning rate is set at 0.001 and weight decay at 0.0002 based on experimentation. Additionally, random weight initialization is performed. We conducted empirical comparisons of the training and validation accuracy of our proposed architecture, utilizing categorical cross-entropy loss [21], with three commonly employed optimizers: Adam [22], RMSprop (root mean square propagation) [23], and SGD (stochastic gradient descent) [21]. Observations indicate that the Adam optimizer yields the highest learning accuracy. In our proposed approach, we employ a total of 100 epochs with a batch size 32 to minimize the softmax loss. Fig. 5(a) displays the convergence plot of the training process depicting the training and validation loss over epochs of our proposed network.



(a) (b)

Fig. 5. (a) Confusion Matrix of FvFc-Net; (b) Convergence Graph of the Proposed FvFc-Net

Table 3. Accuracy Comparison with other state-of-the-art techniques

State-of-the-Art Technique	Accuracy (in %)
Our Proposed Framework	**96.73**
TF-GLCM [8]	95.21
HAC [10]	93.51
DBN-DNN + PCA [11]	94.52
GKSWF + SVM [12]	95.87

5 Experimental Results and Discussion

In this section, we demonstrate the performance evaluation of FvFc-NET, for classifying video frame sequences into authentic and forged frames. For quantitative evaluation of our proposed model, our model has been tested with another set of frame sequences from our own designed dataset those are not used for training our model. Consequently, we also tested our proposed model (FvFc-NET) on publically available REWIND dataset [24]. Fig. 5 (b) displays the confusion matrix of the proposed model in terms of classification accuracy for forgery classification task on the used datasets. Therefore, this plot offers valuable insights into the model's ability to correctly classify instances across different classes, aiding in the evaluation of its overall performance. To assess the efficacy of our proposed model, we conducted a comparative analysis with state-of-the-art pre-trained CNN models. Numerous pre-trained CNN models have been released to the public, accompanied by their learned kernels and weights, specifically designed for the ImageNet challenge dataset [25]. In our study, we utilize ten renowned and extensively employed pre-trained CNN models sourced from literature: VGG-16 [26], VGG-19 [26], AlexNet [27], Inception-V3 [28], GoogleNet [29], Resnet-101 [30], Resnet-50 [30], Resnet-18 [30], Densenet-201 [31] and Inception-Resnet-V2 [32]. To evaluate the effectiveness of our proposed model (FvFc-NET) with state of the art models, we employed the pre-trained CNNs mentioned earlier, utilizing them as both a fixed feature extraction module (CNN_{FFE}) and a transfer learning module (CNN_{TLM}). In our work, we employ a fixed feature extraction approach by removing the fully connected layers from the CNNs while preserving the rest of the network architecture. Following the feature extraction process using these pre-trained networks, we integrate support vector machines (SVM) with various kernels and employ k-fold cross-validation on the fixed feature extractor. This results in the classification of authentic or forged frames within our own designed dataset. For evaluating the performance of (CNN_{FFE}), similar 10000 frames of our dataset used for testing our proposed CNN models are used. The performance of the state of the art CNNs against the best performing kernel (i.e. SVM-Linear) is provided in Fig. 6 (a). Alternatively, employing CNN alongside a transfer learning module proves to be a viable approach in contemporary times for training the

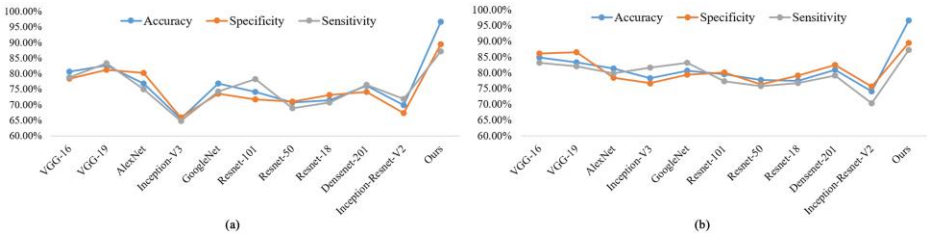


Fig. 6. Performance of our Proposed (FvFc-Net) in Comparison to the State-of-the-Art Pre-Trained CNNs for Classification of forged video frames on our own created Dataset. (a) Pre-Trained CNN as a Fixed Feature Extraction (CNN_{FFE}); (b) Pre-Trained CNN with Transfer Learning Module (CNN_{TLM}).

network with specific datasets. In our comparative study, we also have used transfer learning module on the pre-trained CNN networks (as shown in Fig. 6 (b)). Here, we initially train the base networks using a base dataset, such as the ImageNet challenge dataset. Subsequently, the generic features acquired from this extensive dataset, represented by pre-defined weights, are transferred to the target network while preserving the entirety of the network structure. This enables training these models on our own dataset. For transfer learning with pre-trained CNN models, we leveraged a subset of 100,000 frames from our dataset, which corresponds to the data used in training our proposed model. Meanwhile, for evaluating the performance of (CNN_{TLM}), we have used similar test set of 10000 frames from our dataset used for testing our proposed model.

The effectiveness of both the pre-trained models and our novel model in classifying forged frames is assessed by examining accuracy, sensitivity, and specificity, as depicted in Fig. 6. In case of CNN as a fixed feature extraction module (CNN_{FFE}), it can be observed that VGG-19 [26] outperforms remaining nine pre-trained CNN models with an average accuracy, specificity and sensitivity of 82.70%, 81.34% and 83.43% respectively. Another interesting observation can be clearly observed from Fig. 5 that use of pre-trained CNN models as a transfer learning module (CNN_{TLM}) enhances the Forged Frame classification performance in all of the specific cases as compared to the pre-trained CNN models as a fixed feature extraction module (CNN_{FFE}). In this case, VGG-16 [26] outperforms with an average accuracy, specificity and sensitivity of 84.85%, 86.16% and 83.22% respectively. In comparison, we can observe from Fig. 5 that our proposed CNN model enhances the classification performance of forged frame classification with an average testing accuracy, specificity and sensitivity of 96.73%, 89.58% and 87.31% respectively. Moreover, in order to assess the efficacy of the proposed network's performance, we conducted a comparative analysis of its classification capabilities against those of state-of-the-art methods employed for similar tasks on our dataset. The compared methods are TF-GLCM [8], HAC [10], DBN-DNN + PCA [11], and GKSWF + SVM [12]. The prediction performance of the proposed network and the existing methods on our own dataset is provided in Table 3 in terms of average accuracy. It can be observed from Table 3 that as compared to the state-of-the-art techniques, the method GKSWF + SVM [12] has been observed to perform well on our created dataset with an average accuracy of 95.87%. Moreover, it can also be concluded that FvFc-Net on our proposed framework has superior performance for classification of forged and authentic frames as compared to the state-of-the-art methods with an average accuracy of 96.73%. In addition, our proposed CNN model is also tested on REWIND Dataset [24]. The proposed framework able to predicts the class of video frames (i.e., authentic or forged) with an average accuracy of 95.31% on REWIND dataset.

6 Conclusion

In this paper, we delve into the occurrence of ghost artifacts in manipulated videos and assess their impact on the reliability of conventional forensic techniques for verifying video authenticity. Additionally, we introduce a novel Forged Video Frame Classification Network (FvFc-NET) for classifying forged and authentic frames within videos. This framework integrates a backbone DCNN network with both a Feature Fusion Module (FFM) and an ELA Analysis module. Through rigorous experimentation on our own designed dataset and publically available REWIND dataset, we demonstrate the efficacy of our proposed FvFc-NET method with respect to the existing state-of-the-art methodologies. The findings presented in this paper pave the way for enhanced authenticity verification and integrity assurance in digital imagery, particularly in applications where image tampering can have severe consequences. In the future, we will refine or fine-tune the proposed model to perform well on forged images or videos generated using generative models. Moreover, a key limitation of our approach is the forgery classification task in multiple compression settings which occurs in the videos compressed under various social media platforms and is an example of social media laundering in the domain of multimedia forensics. Therefore, generalization of the proposed model with respect to different compression settings will play a significant role to meet up the forgery classification task and make it applicable for various real-world forensics applications and will be extended in our future work.

Acknowledgement

The work is supported by DST SERB International Research Experience (SIRE) Fellowship awarded to Mrinal Kanti Bhowmik for the year 2022-2023 under the Grant No. SIR/2022/000387, Dated: 12/05/2022 from Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India as a visiting Research scholar (Post Ph.D. Experience) at the NYU Center for Cybersecurity (CCS), Department of Computer Science and Engineering, Tandon School of Engineering, New York University, New York City, USA. Mrinal Kanti Bhowmik would also like to thank his mentor/supervisor, Prof. Nasir Memon, Emeritus Professor of New York University, USA for providing him infrastructural and research supports during the Post-Ph.D. SERB International Research Experience.

References

1. Sharma, P., Kumar, M., & Sharma, H.: Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation. Multimedia Tools and Applications 82(12), 18117–18150 (2023).
2. Alghamdi, M. I.: "Digital forensics in cyber security—recent trends, threats, and opportunities." In: Cybersecurity Threats with New Perspectives, pp. (pages), (2021).
3. Rawat, W., & Wang, Z.: Deep convolutional neural networks for image classification: A comprehensive review. Neural computation 29(9), 2352–2449 (2017).
4. Hakimi, F., Hariri, M., & GharehBaghi, F.: Image splicing forgery detection using local binary pattern and discrete wavelet transform. In: 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), pp. 1074–1077. IEEE (November 2015).
5. Kaur, G., & Khehra, B. S.: An efficient approach for digital image splicing detection using adaptive SVM. International Journal of Computer Science and Information Security 14(6), 168 (2016).
6. Liu, Q., & Sung, A. H.: A new approach for JPEG resize and image splicing detection. In: Proceedings of the First ACM workshop on Multimedia in forensics, pp. 43–48. ACM, New York (2009).
7. He, Z., Lu, W., Sun, W., & Huang, J.: Digital image splicing detection based on Markov features in DCT and DWT domain. Pattern Recognition 45(12), 4292–4299 (2012).
8. Shen, X., Shi, Z., & Chen, H.: Splicing image forgery detection using textural features based on the grey level co-occurrence matrices. IET Image Processing 11(1), 44–53 (2017).

9. Jaiswal, A.K., & Srivastava, R.: Image splicing detection using deep residual network. In: Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), (March 2019).
10. Tiwari, S. K., Mazumdar, A., & Bora, P. K.: Detection of splicing forgery using CNN-extracted camera-specific features. In: Pattern Recognition and Machine Intelligence: 8th International Conference, PReMI 2019, Tezpur, India, December 17-20, 2019, Proceedings, Part I, pp. 473–481. Springer International Publishing (2019).
11. Hussien, N. Y., Mahmoud, R. O., & Zayed, H. H.: Deep learning on digital image splicing detection using CFA artifacts. International Journal of Sociotechnology and Knowledge Development (IJSKD) 12(2), 31–44 (2020).
12. Baleanu, D., Al-Shamayleh, A. S., & Ibrahim, R. W.: Image Splicing Detection Using Generalized Whittaker Function Descriptor. Computers, Materials & Continua 75(2), 1–2 (2023).
13. Chen, W., Verbist, F., Deligiannis, N., Schelkens, P., & Munteanu, A.: Efficient intra-frame video coding for low resolution wireless visual sensors. In: 2013 18th International Conference on Digital Signal Processing (DSP), pp. 1–6. IEEE (2013).
14. Lin, Z., He, J., Tang, X., & Tang, C. K.: Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. Pattern Recognition 42(11), 2492–2501 (2009).
15. Adobe Photoshop. [Online]. Available: https://www.adobe.com/in/products/photoshop/free-trial-download.html, Last accessed 2023/06/06.
16. Filmora 9. [Online]. Available: https://filmora.wondershare.net/, Last accessed 2022/11/23.
17. Abd Warif, N. B., Idris, M. Y. I., Wahab, A. W. A., & Salleh, R.: An evaluation of Error Level Analysis in image forensics. In: 2015 5th IEEE International Conference on System Engineering and Technology (ICSET), pp. 23–28. IEEE (2015).
18. Sudiatmika, I. B. K., Rahman, F., Trisno, T., & Suyoto, S.: Image forgery detection using error level analysis and deep learning. TELKOMNIKA (Telecommunication Computing Electronics and Control), 17(2), pp. 653-659 (2019).
19. Rotobrush. [Online]. Available: https://cmsc426.github.io/rotobrush/#ref, Last accessed 2013/01/21.
20. Adobe: Stand out with after effects. [Online]. Available: https://www.adobe.com/in/products/aftereffects.html, Last accessed 2022/09/06.
21. Sutskever, I., Martens, J., Dahl, G., & Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proceedings of the International Conference on Machine Learning, pp. 1139–1147 (2013).
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
23. Author, F.: Overview of different Optimizers for neural networks. [Online]. Available: https://medium.com/datadriveninvestor/overview-of-different-optimizers-for-neural-networks-e0ed119440c3, Last accessed 2023/03/12.
24. REWIND – Video (2017). Copy-move forgeries dataset. Available: https://sites.google.com/site/rewindpolimi/downloads/datasets/videocopy-move-forgeries-dataset, Last accessed 2024/01/12.
25. Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L.: "Imagenet: A large-scale hierarchical image database." In: Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 248-255 (2009).
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., & Berg, A.C.: "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115(3), 211–252 (2015).
27. Krizhevsky, A., Sutskever, I., & Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, ACM, pp. 1097-1105 (2012).
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 2818–2826 (2016).
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A.: "Going deeper with convolutions." In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 1-9 (2015).
30. He, K., Zhang, X., Ren, S., & Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 770-778 (2016).
31. G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger: "Densely connected convolutional networks." In: Proc. IEEE conference on computer vision and pattern recognition, IEEE, pp. 4700–4708 (2017).
32. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A.A.: "Inception-v4, inception-resnet and the impact of residual connections on learning." In: Thirty-First AAAI Conference on Artificial Intelligence Proceedings, ACM, Vol. 31, No. 1, (2017).