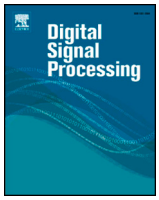
Digit. Signal Process. 147 (2024) 104426

Contents lists available at ScienceDirect

Digital Signal Processing

journal homepage: www.elsevier.com/locate/dsp



Exploring varying color spaces through representative forgery learning to improve deepfake detection

Muhammad Ahmad Amin [a,*], Yongjian Hu [a], Yu Guan [b], Muhammad Zain Amin [c]

[a] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, Guangdong, China
[b] Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK
[c] Duke University, Durham, 27708, NC, USA

ARTICLE INFO

Keywords:
Deepfake detection
Color spaces
Forgery mining
Handcrafted cues image
Multimedia forensics
Color spaces-based forgery detection network

ABSTRACT

In the digital age, the rise of deepfake technology has brought unprecedented challenges to multimedia content authentication. The existing deepfake detection methods generally perform well in known settings. However, generalization and robustness are still challenging tasks. Observing that most conventional methods adopt the RGB color space, we introduce a novel deepfake detection approach by utilizing multiple color spaces to enhance the identification of deepfakes. Overall, our proposed detection framework comprises two primary stages, i.e., representative forgery learning through multi-color space reasoning and the color spaces-based forgery detection network (FDN). The representative forgery learning task is realized in succession through the manipulation cue boosting network (MCBN), color space transformations, and the forgery highlighting network (FHN). MCBN improves the feature representation, alternate color spaces provide distinctive advantages over traditional RGB color space, while FHN plays an auxiliary role, where it not only mines the texture inconsistency but also points out high-level semantic forgery clues, aiding in the robustness ability of FDN to discern subtle alterations in digital imagery accurately. Through rigorous evaluation on the benchmark datasets, including the FaceForensics++, DFDC, and CelebDF datasets, our proposed approach exhibits promising results in identifying forged multimedia content across varying color representations, outperforming the state-of-the-art methods.

1. Introduction

Deepfake forgery generation and detection, a critical component of digital forensics, encompasses a wide range of techniques aimed at generating and identifying alterations or manipulations in multimedia content. While representative facial forgeries methods like Deepfakes [1] and Face2Face [2] blend faces of unique identities, which usually leaves behind blending boundary artifacts, the generative adversarial networks (GANs) [3] based manipulation approaches [4–8] leave inherent fingerprints over the global facial area. Either global fingerprints or local artifacts, these clues are key manipulation traces to expose facial forgeries.

Traditional methods of deepfake forgery detection heavily rely on low-level features such as texture, noise distribution, physiological signal, and other minute discrepancies introduced during the generation of deepfakes [9–12]. Recently, deep learning-based detection approaches have gradually become the mainstay, which can be categorized into video-level and frame-level (image-level) detectors. Frame-level detectors rely on spatial artifacts such as color distortion and texture inconsistency [13], and artifacts in the frequency domain [14]. Furthermore, many detectors analyze artifacts in the latent space to identify fake images [15]. In more intuitive attempt, video-level detectors also considered temporal inconsistencies by exploring forged video frame-by-frame [16]. Although most deepfake forgery detectors [17–21] can achieve nearly perfect detection accuracy in known forgery scenarios, simple post-processing operations like image blurring or compression can compromise the low-level features of images or videos, which makes these forgery traces harder to detect and limits generalized performance.

Observing these aspects, our research is motivated by two points. First, the traditional deepfake forgery generation and detection methods operate in the RGB color space, and the artifacts in this space are better concealed due to better generation training and post-processing.

* Corresponding author.
E-mail addresses: eeahmadamin@mail.scut.edu.cn (M.A. Amin), eeyjhu@scut.edu.cn (Y. Hu), yu.guan@warwick.ac.uk (Y. Guan), muhammad-zain_amin@etu.u-bourgogne.fr (M.Z. Amin).

https://doi.org/10.1016/j.dsp.2024.104426

Available online 12 February 2024

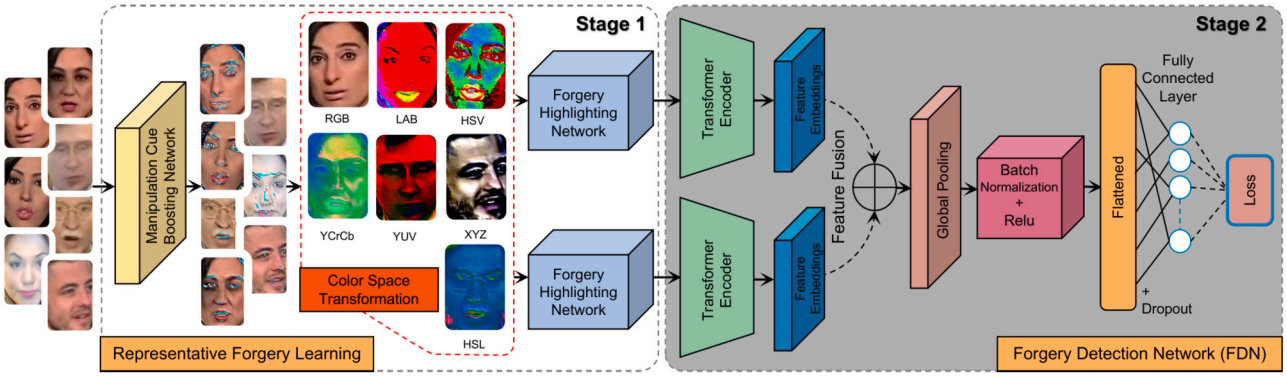


Fig. 1. An illustration of our proposed Color Spaces-based Deepfake Detection Framework.

However, alternative color spaces [22], the fundamental aspect of any visual content, offer a plethora of information that has, until now, been underutilized in the domain of forgery detection. Second, the forged contents will likely escape from most current single- or multi-domain-based deepfake detection approaches, which only focus on limited, low-level forgery cue mining. It, therefore, necessitates a robust detection technique with greater generalization ability, which pay more attention to the corresponding intrinsic manipulation clues rather than over-fitting specific forgeries.

Given this, we introduce a novel deepfake forgery detection framework that draws manipulation traces from varying color space transformations rather than the RGB space. Specifically, we explore seven color spaces and their best possible combinations. The color spaces are combined based on the properties of their distinct components in a manner that complements other space information to improve detection performance rather than playing a redundant role or not contributing at all. The proposed framework comprises several key network modules: the manipulation cue boosting network (MCBN), alternate color space transformations, the forgery highlighting network (FHN), and the color spaces-based forgery detection network (FDN). All these networks collectively facilitate representative forgery learning in two stages. From the image-processing point of view, the MCBN accentuates the blending cues and boundary disparity information within an facial image, resulting in an improved representation of forgery traces. In order to make the network attend to different potential forged regions, we design the FHN in two steps, including 1) generating the image-level handcrafted cues map of the spatially manipulated region and 2) utilizing forgery cues region masking (FCRM) to intentionally occlude the sensitive low-level and high-level structured cues within the facial image, pushing the detector to explore representative forgery from the previously ignored facial region. Finally, the FDN is a transformer-based model specifically tailored to capture high-level semantic textural discrepancies within forged elements across multiple-color spaces. By capitalizing on these discrepancies, we aim to bolster the resilience of our deepfake forgery detection approach against the evolving sophistication of deepfake generation techniques. Fig. 1 showcases our proposed deepfake forgery detection framework.

Our main contributions are as follows,

- We introduce a novel deepfake forgery detection framework that exploits the complementary information across various color spaces, RGB, YCrCb, LAB, XYZ, YUV, HSL, and HSV, to improve the detection performance.
- We also introduce the Manipulation Cue Boosting Network (MCBN), which significantly enhances the distinguishability of forgery cues representations.
- We present a Forgery Highlighting Network (FHN) that plays an auxiliary role in pinpointing and masking the forged regions within multimedia content, facilitating the high-level semantic feature learning process.
- We propose the Color Spaces-Based Forgery Detection Network (FDN), a transformer-based model designed explicitly for the robust forgery detection.
- Through rigorous evaluation across diverse multimedia deepfake datasets, this paper provides empirical evidence which demonstrates the effectiveness and practical utility of our proposed approach.

This paper is structured as follows: Section 2 summarizes related work. Section 3 describes our proposed framework. Section 4 details the experimental settings. Section 5 presents the evaluation analysis and ablation study. Lastly, in Section 6, we present our conclusions and future work.

2. Related work

This section reviews several color spaces and deepfake forensics-related studies.

2.1. Color spaces

A color space serves as a unique geometric representation of colors, offering a metric to measure the distance between its color components [23], [24]. Color spaces can be broadly categorized into four families: primary spaces, luminance-chrominance spaces, perceptual spaces, and independent axis spaces. Fig. 1 illustrates the diverse color spaces we employed in our study.

Primary spaces, such as red-green-blue (RGB) space and XYZ space, are foundational. RGB can be transformed into XYZ through linear transformations, with the X, Y, and Z primaries being imaginary. XYZ serves as a basis for numerous color space conversions, providing a standardized reference for color measurement. In the luminance-chrominance category, YUV and YCbCr spaces, derived from RGB, separate luminance (Y) from chrominance (color information). YCrCb and YUV efficiently represent human perception, emphasizing luminance sensitivity and color separation, making them advantageous for compression tasks. The perceptually uniform LAB space, derived from XYZ, comprises three components: L (lightness), A (green to magenta), and B (blue to yellow). LAB is beneficial for various image processing tasks, including color correction, analysis, and transformations, aligning with human perceptions of colors based on luminosity, hue, and saturation. Perceptual spaces like hue-saturation-lightness (HSL) and hue-saturation-value (HSV) cater to human visual perception. HSL represents colors in terms of hue (color tone), saturation (color purity), and lightness (brightness). HSV, similar to HSL but using value (brightness) instead of lightness, effectively separates color information from brightness, enhancing intuitive image analysis and manipulation.

In terms of adequate comparison, each family has its own set of advantages and use cases, as stated above. In our proposed detection

framework case, the influence of color space transformations on the detection performance of the forgery detection network (FDN) is explicitly demonstrated in Fig. 8 and 9. The results clearly showcase how different color spaces, obtained through these transformations, affect the efficiency of distinctive deepfake forgery detection. The unique color components of each color space are discussed in Fig. 4.

2.2. Deepfake forensics

With the rise of sophisticated deepfake manipulation methods [1,2,4–8], distinguishing doctored videos or images from authentic ones has become increasingly challenging due to their lifelike quality. Researchers are actively enhancing deepfake detectors from diverse perspectives, including CNN-based backbones [12] inherited from image classification models, attention mechanisms for high-level texture feature extraction [13], and the integration of multiple domains like spatial, frequency, and temporal domains [14,18,19,25]. The exploration extends to examining consistency between consecutive frames [15–17], and incorporating multimodal networks, such as content-aware, label-unaware, and residual Federated learning [20,21,26,27], including the near-infrared light modality.

Early works, exemplified by [12], leverage popular image classification backbones like Xception as a baseline. However, these CNN backbones, designed for image classification, emphasize category-level differences, leading to suboptimal performance when confronted with unseen datasets. Wang et al. [13] addressed this limitation by proposing an attention-guided data augmentation mechanism (RFM) to randomly erase forgery cues in the spatial domain, aiming to reduce overfitting. Despite these efforts, generalized cue learning remains a challenge.

To enhance fidelity in the frequency domain, Chen et al. [14] introduced the local relation learning network (LRLNet), where spatial and frequency features are fused in different layers of the backbone network. The authenticity of a face is determined by calculating the multi-scale similarity between different patches. While providing robust preprocessing, this method does not perform well against unseen manipulations. Similarly, Liu et al. [25] investigated up-sampling operations in the frequency domain, common in facial forgery techniques, and introduced spatial phase shallow learning (SPSL). SPSL utilizes both spatial images and phase spectrums to capture up-sampling cues. However, due to content-irrelevant learning of fine-grained spatial frequency features, these methods lack information interaction between two domains and exhibit limited generalized performance.

To capture inconsistencies in forgery videos, Sun et al. [15] proposed an intra-frame inconsistency learning framework (DCL), generating annotated forged location labels by subtracting the forgery image from its corresponding real image. However, this approach focuses on category-level discrepancies rather than intrinsic disparities between genuine and falsified images. In an effort to capture more generic clues, Zheng et al. [16] introduced a 3D spatio-temporal neural network (FTCN) for detecting temporal incoherence artifacts. While promising, this method lacks the interpretability of both temporal and spatial information. For improved robustness, Cao et al. [17] presented reconstruction-classification learning (RECCE) based on reconstruction differences to identify forgery traces but faced challenges in overfitting real faces for specific demographic groups. In another approach, Yu et al. [18] introduced a common forgery learning framework (CFFE) focusing on blending boundary cues left by different manipulation methods. However, due to its exclusive emphasis on boundary cues, it exhibits limited performance in cross-manipulation scenarios.

Most recent works address these limitations through multi-domain feature learning methods, including content-aware spatial-frequencies-temporal learning in visible and near-infrared light [19,20,26], residual federated learning [27], and label-unaware learning [21]. Wang et al. [19] introduced a spatial-frequency dynamic graph-based network (SFDG) using relation-aware spatial-frequency features for generalized forgery detection. Wang et al. [20] proposed a spectrum spatial-temporal frequency clue network (FCAN-DCT) exploring both spatial and temporal frequency features with an attention mechanism, but it fails to learn more intrinsic features. Dong et al. [21] highlighted the sensitivity of deepfake detection models to label training data and introduced an ID-unaware deepfake detection model (CADDM) to improve generalization across various manipulations. Nonetheless, this method performs well on face-swapping manipulation but has limited performance in other cross-manipulation tests.

3. Proposed methodology

In this section, we discuss our multi-stage forgery detection framework in detail, i.e., representative forgery learning and forgery detection framework (FDN). The unified framework pipeline is illustrated in Fig. 1. In the representative forgery learning framework, first, we introduce the manipulation cue boosting network (MCBN). Second, we explore the color spaces and the impact of MCBN on the discernibility between the color components of real and deepfake images in varying color spaces. Third, we introduce an auxiliary supervision network, the forgery highlighting network (FHN), to guide the FDN to discover undetectable deepfake clues through forgery artifact mining and less network-sensitive texture changes masking. Lastly, in the second stage, we presented the color spaces-based FDN, which utilizes the advantages of key schemes, i.e., representative forgery learning through multi-color space reasoning and convolutional transformer encoder-guided attention.

3.1. Manipulation cue boosting network

In the case of face forgeries, manipulation traces are often present as small, isolated dots or linear textures that are difficult to realize with the naked eye. From the image-processing field perspective, these manipulation cues can be highlighted by traditional gradient operators. Hence, keeping this simple idea in mind, we introduce the MCBN at the input layer of our proposed representative forgery learning framework. The MCBN module is applied directly to the three input channels of a facial image, enhancing the blending cues and boundary disparity information. Specifically, MCBN takes an input image I_i and then applies a simple diagonal gradient operator G to perform convolutional operations on the images of various color channels. Once processed by G, the resultant output is added element-wise to its original input I_i, shown in Fig. 2. As given in (1), this process results in a refined and boosted image I_B, which retains both the enhanced and the original characteristics.

$$I_B = I_i + \left(\sum_{i=1}^{C} I_i \circledast G \right) \tag{1}$$

where G denotes the possible gradient operator and C is the number of channels. Commonly used gradient operators are summarized in Fig. 3, and we also performed an ablation study to select and analyze the impact of MCBN framework with the mentioned gradient operators on the detection performance (supplementary material, section 1.1). Although MCBN utilizes convolutional operations with fixed kernel coefficients, it provides a straightforward yet potent preprocessing effect.

3.2. Color spaces and their components analysis

The mentioned color spaces, such as RGB, YCrCb, LAB, YUV, HSL, HSV, and XYZ, represent images from different perspectives, and variations in these color components can provide important clues for identifying deepfake images or videos. In the context of deepfake generation, the Wasserstein distance (WD) [28] serves as a divergence metric, highlighting differences in pixel value distributions between real and generated images, with the generative network working to minimize the WD gap among them to improve their convergence.

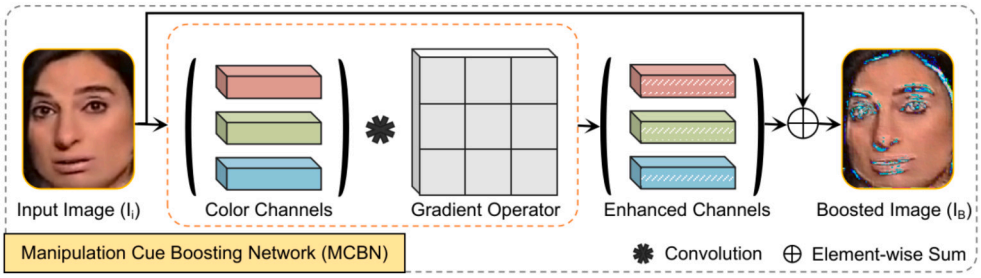


Fig. 2. Manipulation Cue Boosting Network improves the representation of manipulation clues left by facial forgery methods. Yet, it still retains the original image by combining it with the enhanced one.

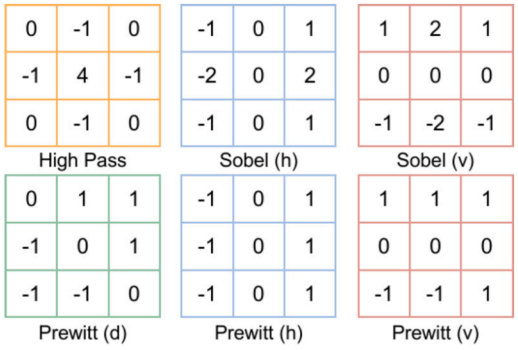


Fig. 3. The classical gradient operators are denoted by d, h, and v for diagonal, horizontal, and vertical directions, respectively. In this context, we utilize the Prewitt (d) operator to perform convolution operations.

Similar to [29], we opted to utilize WD for the analysis of real and synthesized deepfakes across various color spaces. To conduct this analysis, we specifically chose pristine and deepfake samples of the same subject from the FF++ [30] dataset. Also, we test the WD metric with and without applying the MCBN to demonstrate its effectiveness, as illustrated in Fig. 4. Precisely, we compare the WD difference between the real (I_{real}) and deepfake (I_{fake}) images color component $I^C \in \{R, G, B, Y, Cr, Cb, L, A, B, Y, U, V, H, S, L, H, S, V\}$ distributions. The WD for a specific channel in a unique color space is calculated as follows:

$$WD\left(I_{real}, I_{fake}\right) = \sum_{i=1}^{n} |I_{real}^{C(i)} - I_{fake}^{C(i)}| \cdot d\,(i) \tag{2}$$

Where $I_{real}^{C(i)}$ and $I_{fake}^{C(i)}$ are the color component values for the real and deepfake images at position i, n denotes the total number of pixels being considered in the color component distributions, and $d\,(i)$ represents the distance between positions of real and fake samples color components in the different color spaces.

As shown in Fig. 4 and Table 1, it is evident that when we examine the color components C of the I_{real} and I_{fake} samples - specifically in the RGB color space - the WD between the red channel component increases from 8.52 to 10.83. The same trend is observed for the green channel component, which increases from 8.45 to 12.28, and the blue channel component, which increases from 6.93 to 13.29. Other color spaces like YCrCb, YUV, and HSV exhibit a similar pattern. Hence, this phenomenon demonstrates that the WD between the color components C of the real and deepfake images or videos in various color spaces becomes more pronounced after adopting the MCBN, which indicates that the spatial dependencies enhanced by the MCBN amplify the subtle discrepancies inherent between synthetic and real content.

3.3. Forgery highlighting network

To point out the potentially forged areas within the facial region and avoid over-fitting to specific forgery patterns, FHN mines the high-level structured semantic forgery clues in the facial image in two steps, as shown in Fig. 5. The first step generates the sensitive position maps for each mini-batch of input facial images through the manipulated region identifier network ($MRIN$). In the second step, FCRM within FHN highlights the noteworthy local cues as NTC regions (sketched with red circles) and the minute texture changes or inconsistencies as MTC regions (marked with yellow circles). FHN generates corresponding masks by selectively erasing patches in each batch of facial images as NTC and MTC regions. Finally, the handcrafted mini-batch is fed into the FDN.

3.3.1. Manipulated region identifier network

In order to guide the FDN model towards semantic forgery cues, the $MRIN$ needs to precisely point out the potential forgery cue regions left by manipulation methods that the detection network is sensitive to. Concretely, the most sensitive region is defined as the region where alteration has left artifacts, which can critically impact detection performance. To realize the sensitive regions, in the forward propagation, $MRIN$ receives facial image I_B as input and outputs logits $L = \left[L_{real}, L_{fake}\right]$, where the backpropagation process updates image gradient location values adaptively under the guidance of binary labels (3),

$$L = \left[L_{real}, L_{fake}\right] = MRIN\left(I_B\right) \tag{3}$$

As any alteration Δ_{I_B} in a real or fake image would affect both logits $\left(L_{real}, L_{fake}\right)$, the resultant sensitive region gradient values I_g of input I_B should be determined by the relative magnitude of the two logit values. Hence, by utilizing the $\Delta_{I_B}.L_{real}$ and $\Delta_{I_B}.L_{fake}$ to separately represent how alteration Δ_{I_B} in I_B impacts the logits, we can determine the I_g as,

$$I_g = \Delta_{I_B} \cdot L_{fake} - \Delta_{I_B} \cdot L_{real} \tag{4}$$

Hence, the maximum absolute difference values of I_g are regarded as a manipulated region map $\left(MRM_{sensitive}\right)$, thus revealing the regions where the potential artifacts are located. In other words, each value in $MRM_{sensitive}$ precisely indicates the sensitivity of the detector to the corresponding pixel in an image or video. Fig. 6 depicts various examples of handcrafted cues images. Formally, $MRM_{sensitive}$ can be formulated as,

$$MRM_{sensitive} = max\left(abs\left(I_g\right)\right) \tag{5}$$

Where the function $max\,(.)$ calculates the maximum value along channel axis and $abs\,(.)$ obtains the absolute value of each pixel.

3.3.2. Forgery cues region masking

The $MRM_{sensitive}$ generated by $MRIN$ guides subsequent operations in FCRM, where the $MRM_{sensitive}$ contains two kinds of forged areas. First, the NTC regions imply that there are many significant artifacts in the facial area and the network will focus on. Second, the minute texture inconsistency regions (MTC) contain fewer tampering artifacts but more information about texture changes, which is usually ignored by the network.

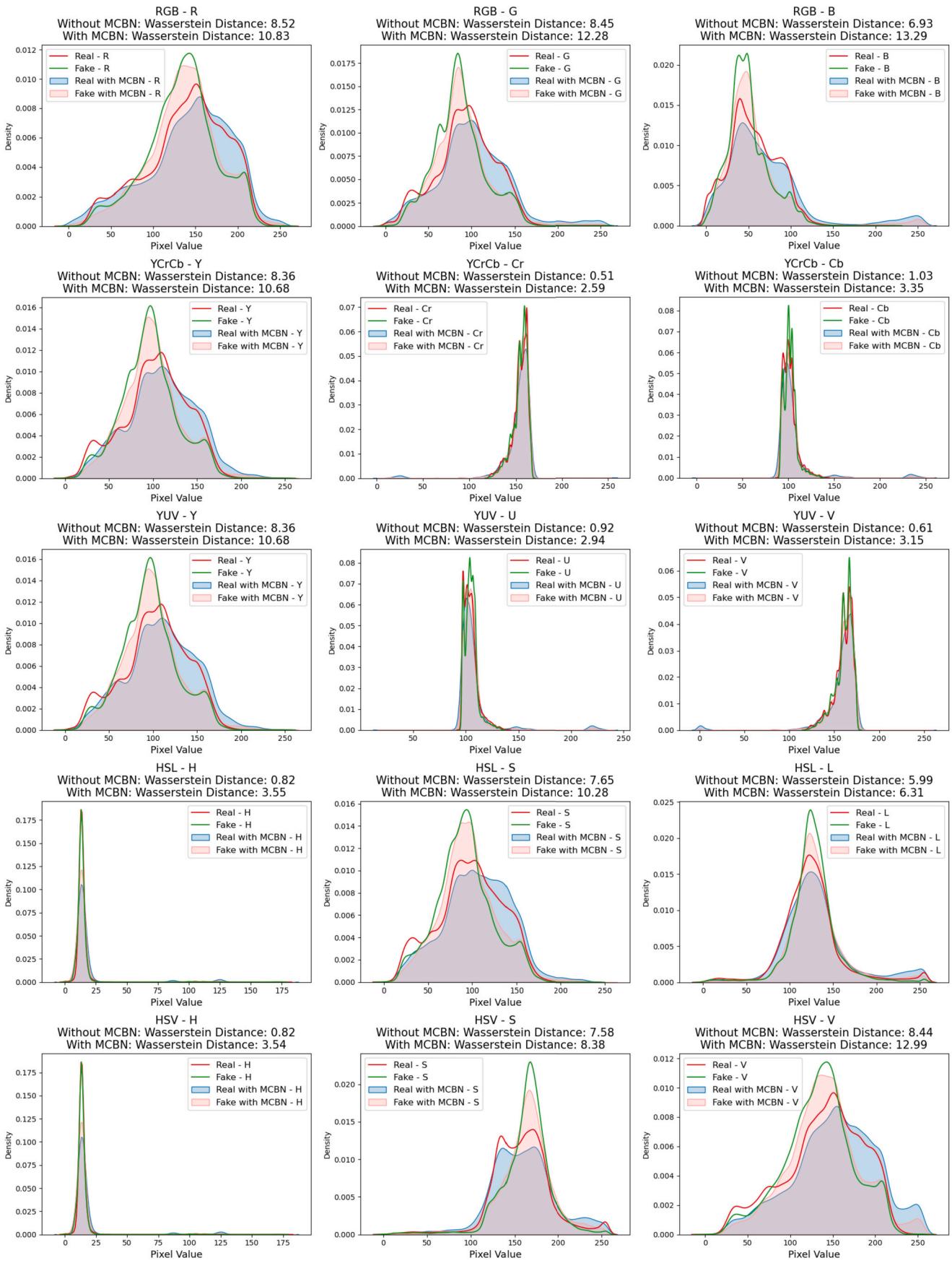


Fig. 4. We assess the effectiveness of the color spaces and MCBN by comparing the WD between the pixel value distributions of authentic and deepfake images of the same subject, both with and without applying the MCBN module.

Table 1
The effectiveness of the color spaces and MCBN is given by comparing the WD between the pixel value distributions of authentic and deepfake images in distinct color space components.

No.	Color Spaces	Color Components	WD Without MCBN	WD With MCBN
1	RGB	R	8.52	10.83
		G	8.45	12.28
		B	6.93	13.29
1	YCrCb	Y	8.36	10.68
		Cr	0.51	2.59
		Cb	1.03	3.35
2	YUV	Y	8.36	10.68
		U	0.92	2.94
		V	3.15	0.61
3	HSL	H	0.82	3.55
		S	7.65	10.25
		L	5.99	6.31
4	HSV	H	0.82	3.54
		S	7.58	8.38
		V	8.44	12.99

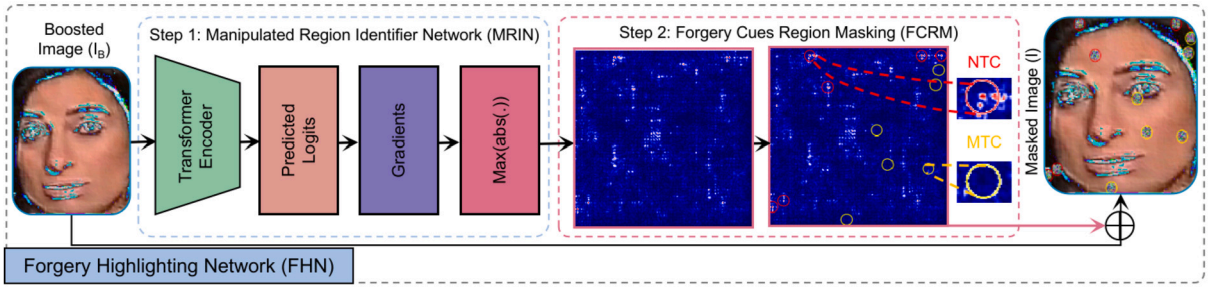


Fig. 5. A overview of FHN framework, which can be divided into two parts. Firstly, we generate manipulated region map $(MRM_{sensitive})$ for each original image of a single mini-batch. Then, we utilize FCRM to partially mask the original images under the guidance of generated $MRM_{sensitive}$.

To fully mine the high-level semantic clues in the global facial image, FCRM simultaneously erases some patches in both the NTC and MTC regions. Specifically, FCRM only erases the most sensitive Top-5 patches in the NTC regions to occlude the significant artifacts partially. Similarly, the MTC regions are erased randomly. Thus, there are a total of $2N$ erased patches from each image. To keep the structural information between artifacts, the patch size $H_p \times W_p$ is set to a relatively small value to prevent the forgery clues from being completely occluded. In our FCRM, the patching hyperparameters N, H_p, and W_p are set to 5, 10, and 10, respectively, and these hyperparameters are selected through experimentation, where we analyze the impact of FHN patching hyperparameters on FDN detection performance and select the optimal hyperparameters (supplementary material, section 1.2). The main steps for FCRM are summarized in Algorithm 1, and portrayed in Fig. 6.

Algorithm 1: Forgery Cues Region Masking (FCRM).

Input: Input image: I_B
Input image size: $W \times H$
Erased patch size: $W_p \times H_p$
Number of patches: N
Maps generated by MRIN: $MRM_{sensitive}$
Output: Masked image: I
$count \leftarrow 0$;
while $count < N$ **do**
  $I_B[x_{NTC}, y_{NTC}] \leftarrow$ coordinates of the maximum pixel value in $MRM_{sensitive}$;
  $NTC_{top} = max(x_s - H_p/2)$;
  $NTC_{bottom} = min(x_s - H_p/2)$;
  $NTC_{left} = max(x_s - W_p/2)$;
  $NTC_{right} = min(x_s - W_p/2)$;
  Fill input $I_B[NTC_{top} : NTC_{bottom}, NTC_{left} : NTC_{right}]$ with random integers as the mask;
  if *input* $I_B[x_{MTC}, y_{MTC}] \leftarrow$ *random coordinates in* $MRM_{sensitive}$ **then**
    $MTC_{top} = x_w - H_p/2$;
    $MTC_{bottom} = x_w - H_p/2$;
    $MTC_{left} = x_w - W_p/2$;
    $MTC_{right} = x_w - W_p/2$;
  Fill input $I_B[MTC_{top} : MTC_{bottom}, MTC_{left} : MTC_{right}]$ with random integers as the mask;
  $count \leftarrow count + 1$;
return *sensitive region masked image:* I;

3.4. Forgery detection network

In order to detect the deepfake facial clues highlighted by FHN in different color spaces, we designed the color spaces-based FDN, as depicted in Fig. 1. The FDN model employs two convolutional vision transformer encoders [31] as part of the backbone network base. The encoders structure is illustrated in Fig. 7(a). In the encoder model of FDN, a CNN-inspired architecture with attention is used [32] to set up a hierarchy for semantic forgery cues extraction. The encoder model within FDN comprises three stages, as shown in Fig. 7(a). Each stage starts with a convolutional token embedding, using convolution with a specific stride on a reshaped 2D token map. Then, layer normalization is applied, following [33]. This process cycle helps in capturing local information while reducing the sequence size and increasing token features, resulting in spatial down-sampling and more feature maps. Specifically, the output token map from a previous stage I_{i-1} can be ex-

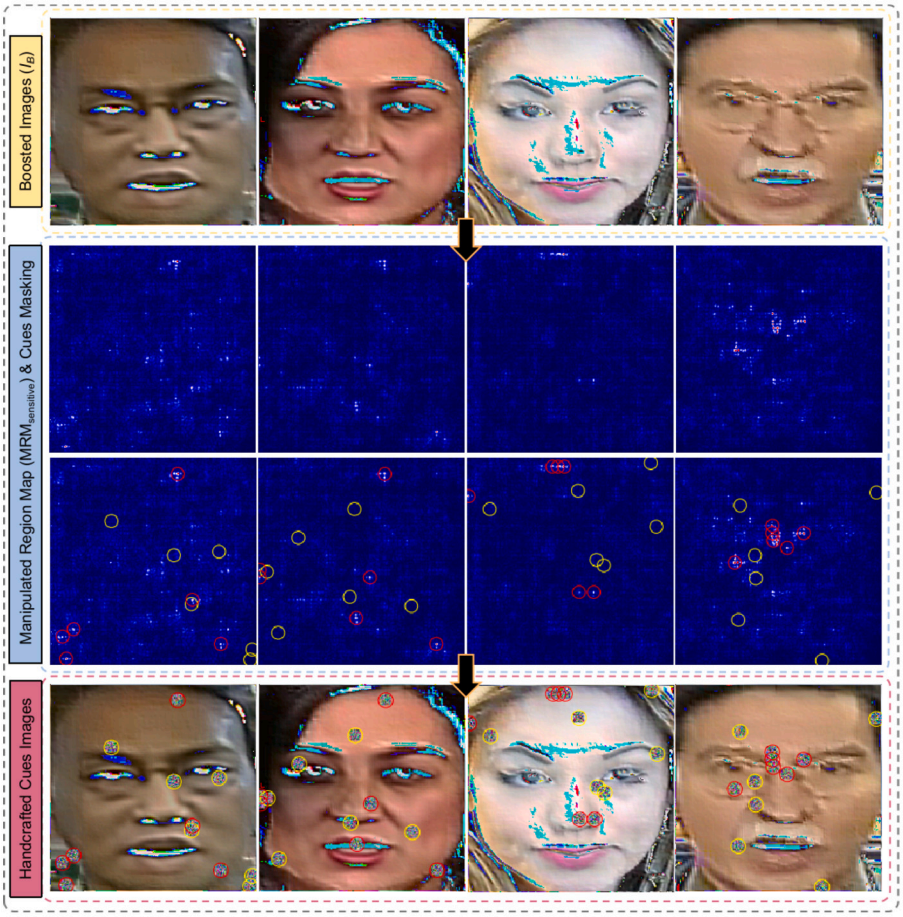


Fig. 6. The FHN framework complete working principle is shown. First row shows the boosted images (I_B). The second and third row show the manipulated region maps $(MRM_{sensitive})$ and partially masked original input boosted images (I_B) maps. The last row shows the mined and masked handcrafted high-level clues images.

pressed as a 2D image or video frame I_i. This approach enables the modeling of local spatial contexts in a hierarchical manner.

$$I_{i-1} \in \mathbb{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}} \tag{6}$$

The FDN goal in stage i is to create a functional map of I_{i-1} to new tokens called $f\left(I_{i-1}\right)$, with a channel size of C_i. To accomplish this, we use a function called $f(.)$, which involves a two-dimensional convolution operation with a kernel size of $s \times s$, a stride of $s - o$, and padding p to handle boundary issues. The resulting map of tokens is given as in (7),

$$f\left(I_{i-1}\right) \in \mathbb{R}^{H_i \times W_i \times C_i} \tag{7}$$

Here, H_i and W_i denote the height and width, respectively. Each stage involves applying a convolutional projection to self-attention blocks within the FDN encoder module. A depth-wise separable convolution [12] is utilized on a 2D token map to enhance local context comprehension and mitigate semantic ambiguity. To streamline the process, the convolution stride is used to sub-sample the key K and value V matrices, resulting in improved efficiency with minimal performance impact, as demonstrated in Fig. 7(c). In order to apply the convolutional projection, a convolution layer that is depth-wise separable with a kernel dimension of is utilized, and the tokens that have been projected are flattened into a 2D shape for further processing.

$$I_i^{q/k/v} = Flatten\left(Conv2D\left(Reshape2D\left(I_i\right), s\right)\right) \tag{8}$$

At layer i in matrices $Q/K/V$, the input token is represented by $I_i^{q/k/v}$, while I_i represents the original undisturbed token. To carry out the Conv2D process, a depth-wise separable convolution [12] is used, which is achieved by combining Depth-wise Conv2D, BatchNorm2D, and Point-wise Conv2D. The convolution kernel size is denoted by s. By incorporating a convolutional projection layer, the transformer block within FDN can easily achieve the original position-wise linear projection layer using an 1×1 convolution layer.

Building upon this strong foundation after stage 3, the FDN model introduces a unique fusion strategy in the following layers by employing element-wise addition that seamlessly merges feature representations extracted by two encoders from two color space representative embeddings containing potential distinct complementary cues information. This approach is motivated to avoid redundancy and ensures that each contributes meaningfully to the overall feature representation rather than concatenation, where simultaneously combining more than two color spaces does not provide additional distinctive information. The resulting fused representation F can be expressed mathematically as in (9).

$$F = F_{I_1} + F_{I_2} \tag{9}$$

After the fusion step at stage 4 and 5, the fused feature representation F is passed through convolutional layers to capture local patterns and relationships. The first convolutional layer $(Conv_1)$ applies a 3×3 convolution to the fused features, followed by batch normalization and ReLU activation. The second convolutional layer $(Conv_2)$ further refines the features by applying another 3×3 convolution, followed by batch normalization and ReLU activation. The output of convolutional layers, denoted as x, represents locally enhanced features. The convolution operation (for each layer) can be stated as:

$$x = \text{Relu}\Big(\text{batchNorm}\Big(\text{Conv2d}\big(F, \text{ConvWeights}\big) + \text{ConvBias}\Big)\Big) \tag{10}$$

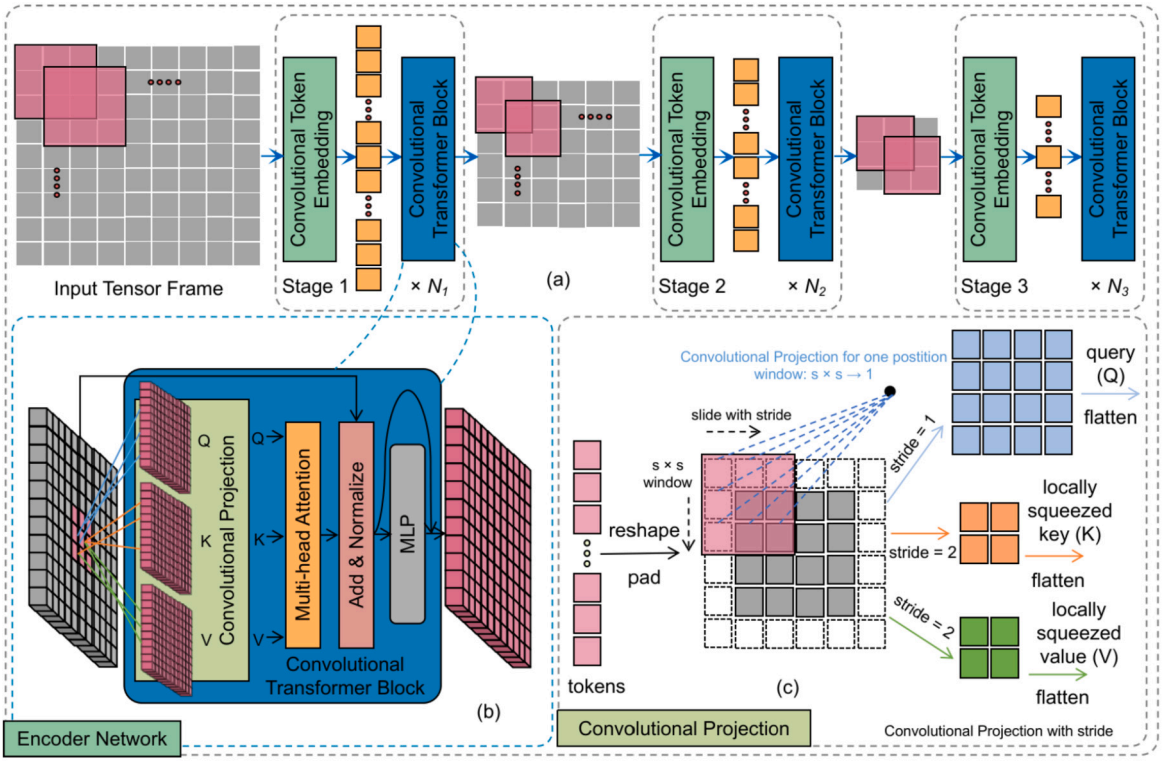


Fig. 7. (a) The proposed encoder architecture comprises a hierarchical multi-stage structure, as depicted in the overall architecture in Fig. 1, enabled by the Convolutional Token Embedding layer. (b) Illustration of the convolution projection as the initial layer in the convolutional transformer block, and (c) shows the squeezed convolutional projection we employed.

In subsequent operations of FDN, the globally enhanced feature maps x obtained from the convolutional layers are subjected to global average pooling (GAP), which computes the average value of each channel across the spatial dimensions of the feature maps. This operation reduces the spatial dimensions to a single value per channel, effectively summarizing the presence of different features across the entire image.

$$x_p = GAP(x) \tag{11}$$

Further, the pooled feature maps x_p resulting from global average pooling are flattened into a 1D vector. This vector retains the information captured by the convolutional layers but reshapes it into a format suitable for feeding into subsequent fully connected layers. Flattening transforms the 3D tensor into a 1D tensor, as in Equation (12).

$$x_f = flatten\left(x_p\right) \tag{12}$$

Now, the dropout is applied to the flattened feature vector x_f to randomly drop a fraction of the elements during training, as in (13), which helps prevent overfitting and improve generalized performance.

$$x_d = dropout\left(x_f\right) \tag{13}$$

Finally, after the dropout, feature vector x_d is passed through a fully connected layer, which performs a weighted sum of input features with learnable weights (parameters). The output of this layer represents the final prediction.

$$y = Linear\left(x_d, W, b\right) = x_d \cdot W^T + b \tag{14}$$

Here, x_d represents the input features after dropout, y denotes the output class scores, b represents the bias vector, and W is the weight matrix. The successive layers after the fusion process, including convolutional layers, global average pooling, and flattening, refine the fused features F, enabling the FDN to grasp intricate local patterns and global characteristics to culminate in accurate predictions y regarding the authenticity of a given facial image or video frame. The FDN parameters are given in Table 2.

Table 2
Details regarding the model parameters of The FDN.

Stages	Layer Name	FDN	Output Size
Stage 1	Conv. Embedding	7 × 7, 64, **stride** 4	56 × 56
	Conv. Projection	[3 × 3, 192] × 2	
	MHSA	[H1 = 1 , D1 =64] × 2	56 × 56
	MLP	[R1 = 4] × 2	
Stage 2	Conv. Embedding	3 × 3, 192, **stride** 2	28 × 28
	Conv. Projection	[3 × 3, 768] × 2	
	MHSA	[H2 = 3 , D2 =192] × 2	28 × 28
	MLP	[R2 = 4] × 2	
Stage 3	Conv. Embedding	3 × 3, 384, **stride** 2	14 × 14
	Conv. Projection	[3 × 3, 1024] × 2	
	MHSA	[H3 = 6 , D3 =384] × 2	14 × 14
	MLP	[R3 = 4] × 2	
Stage 4	$Conv_1$	576	64 × H × W
	$batchNorm_1$	128	
Stage 5	$Conv_2$	73856	128 × H × W
	$batchNorm_2$	256	
Stage 6	Global Avg. Pool	0	128 × 1 × 1
Stage 7	Flatten	0	128
Stage 8	Dropout	0	128
Stage 9	Linear	128	2 (*classes*)

4. Experimental settings

This section outlines the datasets, the evaluation metric, hyperparameters, and the loss function used for experiments.

Table 3
A synopsis of the datasets employed in our experimental evaluation.

Dataset	Manipulation Methods	Real/Fake Videos
FF++ [30]	DF, F2F, FS, NT	1,000/4,000
DFDC [34]	Audio Swaps, MM/NN-FS, NTH, DF-128, DF-256, Refinement, DF-AE, StyleGAN, FSGAN	23,654/104,500
CDF [35]	Enhanced Deepfake Facial Swaps	590/5,639

4.1. Datasets and evaluation metric

We conducted experiments on three widely recognized benchmark datasets to assess the effectiveness of our proposed framework. These datasets include FaceForensics++ (FF++) [30], which consists of subsets such as DeepFakes (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT) with three compression ratios (raw (c0), high quality (c23), and low quality (c40)), as well as Deepfake Detection Challenge (DFDC) [34] and Celeb-DF-v2 (CDF) [35]. Table 3 provides a comprehensive overview of all the mentioned datasets, including detailed information of their generation methodologies. We selected 1000 videos from the each subset of the FF++ dataset and divided them into three distinct subsets, namely training, validation, and testing, with a distribution as in [30]. We utilize the Area Under the Receiver Operating Characteristic Curve [36] (AUC) as an evaluation metric following the previous works [13,14,18,15,16,21], which is standard practice in most studies. The AUC values are calculated based on frame-level evaluation.

4.2. Model initialization

The two base encoders within our FDN model are initialized from the pre-trained convolutional vision transformer (CvT) [31] backbone weights trained on the ImageNet1k dataset. This choice was motivated by the success of CvT in capturing hierarchical features in image data, as demonstrated through its strong performance in image understanding tasks. The rest of the FDN model weights are randomly initialized. Subsequently, all the FDN parameters are fine-tuned on deepfake datasets.

4.3. Hyperparameters and preprocessing settings

In this subsection, we outline the key hyperparameters and preprocessing steps employed in the training and testing of the FDN model. We carefully choose the learning rate (lr) of $5e^{-5}$, and a batch size of 128 for training and testing processes through a hit-and-trial approach for optimal hyperparameters, balancing convergence speed and stability. The FDN models initialized from pre-trained CvT encoder weights undergo training for 5 epochs using the AdamW optimizer, and to dynamically adjust the learning rate during training, a Cosine Annealing Learning Rate Scheduler is applied with a maximum number of epochs $\left(T_{max}\right)$ set to 3, $eta_{min}=0$, and $last_{epoch}=$-1. Further, the dropout rate was meticulously selected through an experimental process, ranging from low to high values, aimed at striking a balance between model capacity and generalization to prevent overfitting. Concerning data preprocessing, a normalization transformation is applied using dataset-specific mean and standard deviation values on the input images with the 224 × 224 size. These values, approximately [0.6051, 0.4498, 0.3932] for means and [0.2209, 0.1863, 0.1834] for standard deviations, ensure that the dataset is appropriately standardized.

4.4. Loss function

The cross-entropy loss function [37], denoted as L_{CE}, is employed in the training of the FDN model, considering that deepfake detection is a binary classification problem. Given the predicted class scores $\hat{y}_i$ and the ground truth labels y for a classification task, the L_{CE} is computed as:

$$L_{CE}\left(y_i, \hat{y}_i\right) = -\sum_i y_i \cdot log\left(\hat{y}_i\right) \tag{15}$$

Here, y_i represents the binary ground truth label (1 for real images, 0 for fake images) for class i, and $\hat{y}_i$ denotes the predicted raw score of class i produced by the FDN. An epoch-wise analysis of the FDN loss function is given in the ablation study (supplementary material, section 1.3).

5. Results and evaluations

In this section, we evaluated our proposed forgery detection framework by conducting experiments in intra-, cross-manipulation, and cross-dataset scenarios using seven different color spaces, including RGB, YCbCr, HSV, LAB, HSL, XYZ, and YUV, and their combinations. By doing so, we were able to simulate previously unseen forgeries in different cross-domain settings and compression ratios. We utilized the FF++ [30] subsets at the c0 compression level for training and c0, c23, and c40 as testing sets (see Fig. 8) to perform within- and cross-manipulation evaluations. For cross-dataset performance evaluations, we trained our proposed FDN on complete FF++ (c0) [30] and tested it on the DFDC [34] and CDF [35] datasets (refer to Fig. 9). Second, we compare the best results from the first step study in intra-, cross-manipulation, and cross-dataset settings against state-of-the-art methods. In the third and last part, we performed the ablation study to analyze the contribution of individual building blocks of our proposed framework. We implemented the comparison methods with the star ($*$) symbol in the same environment as ours by using publicly available codes. Further, we have provided an extended robustness evaluation study against unseen post-processing operations in the ablation study (supplementary material, section 1.4).

5.1. Analysis with respect to multiple color spaces and their impact on the performance of FDN

This subsection of our study comprises two important analyses in various color spaces.

5.1.1. Intra- and cross-manipulation evaluation under three compression ratios in multiple color spaces

The extensive evaluation, as depicted in Fig. 8, provides a thorough analysis of intra- and cross-manipulation scenarios within the FF++ dataset [30] (encompassing DF, FS, F2F, and NT) across three compression levels (c0, c23, and c40) to elucidate the influence of diverse color spaces on detection performance.

The analysis, detailed in Fig. 8 (a1, a2, a3, and a4), where DF serves as the source domain dataset and target datasets include DF, FS, F2F, and NT, reveals intriguing insights. In this case, the HSV color space emerges as a top performer overall, achieving 94.54% detection accuracy, especially at compression ratio c0 in cross settings, FS (86.99%), F2F (97.61%), and NT (93.90%). However, its performance wanes at higher compression levels of c23 and c40. HSV represents colors in terms of hue, saturation, and value, which helps perform well in cross-manipulation settings on higher-quality images. The HSL secures the second position, excelling overall at the c0 compression level (93.24%), as shown in Fig. 8(a4), but performs relatively poorly at c23 and c40 among all color spaces in cross-manipulation evaluation. HSL focuses on hue, saturation, and lightness. High sensitivity to color information might factor into its poor performance at c23 and c40. Notably, under c23 and c40 compression settings, YCrCb and YUV in terms of individual color spaces consistently performed well. YCrCb demonstrates consistent detection accuracy across intra-settings, notably excelling in DF at c23 (92.90%) along with YUV (97.67%) and XYZ (96.96%). In cross-settings, YCrCb also delivers strong results in F2F (c0) (95.46%) and NT (c23) (78.90%). Similarly, YUV shows competitive accuracy in intra-settings, with notable the best performance in DF (c23) (97.67%). In another interesting individual color space case, XYZ is extremely resistive to compression in terms of intra-evaluation settings (DF at

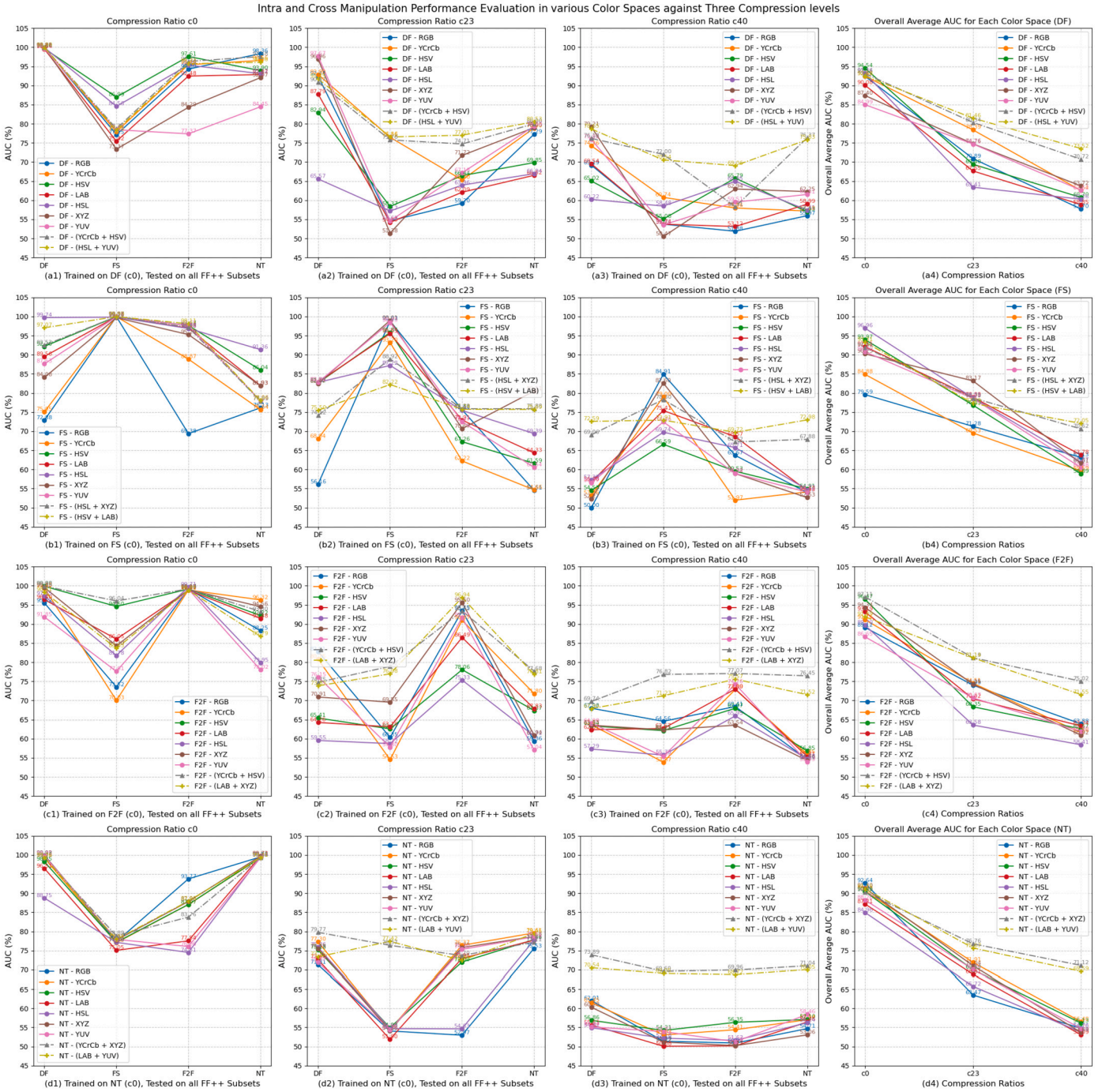


Fig. 8. Our Proposed Method is Evaluated for Intra and Cross-manipulation on the FF++ [30] Subsets. Results are compared by employing different Color Spaces in terms of AUC (%).

c0, c23, and c40: 99.86%, 96.96%, and 62.25%). In terms of overall performance, the combination of YCrCb and YUV with HSV and HSL, respectively, yields the best performance under c23 and c40 compression levels (at c23, 80.34%, 81.46%, and at c40, 73.52%, 70.72%), which highlights the significance of merging multiple color spaces. In other words, HSL and YUV demonstrate robustness to compression, while others, like individual color space HSL, show sensitivity to color variations. The combinations of color spaces improve performance across different compression levels, making them better for feature learning and detection in our proposed detection framework.

Fig. 8 (b1, b2, b3, and b4) showcases the performance when the FS dataset is used for training; the rest are used as test sets (DF, FS, F2F, and NT). All color spaces demonstrated exceptionally high performance for the intra-settings on FS with a c0 compression, nearly reaching 100%. In cross-manipulation settings, HSL performed the best at the c0 compression level. As we increased the compression to c23 and c40, there were noticeable drops in performance in all color spaces, but in terms of individual color space, XYZ performed the best with an AUC score of 83.17% at c23, where the HSL color space showed a drastic decline, as shown in Fig. 8 (b2 and b3), respectively. Observation suggests that when we tested on DF, F2F, and NT under c0 compression, the combination of HSV and LAB spaces exhibited superior performance at compression level c40 (72.05%). Specifically, in cross-setting on the F2F at the c0 level, the said combination achieves an impressive

AUC score of 98.11%. However, under the c40 compression setting, the HSV and LAB combination remained resilient, reaching an overall AUC score of 70.62%. In the overall intra- and cross-manipulation average AUC scenario, the HSL yielded the best performance, with an average AUC score of 96.96% at the c0 compression level as an individual color space. While some color spaces, like XYZ and LAB, demonstrate more robustness to compression, others, like RGB, are less effective. Combinations of HSL with XYZ and HSV with LAB color spaces offer balanced performance across various compression settings, especially at c40.

Based on the plots in Fig. 8 (c1, c2, c3, and c4), we glean the following insights: When FDN is trained on the F2F dataset, The HSV and XYZ color spaces show the second and third best overall results as single color spaces at compression level c0 (96.47% and 94.34%). Based on observation of Fig. 8(c4), it is evident that when tested on DF, FS, F2F, and NT under all compression settings, the combination of YCrCb and HSV spaces exhibited superior performance overall. Remarkably, at all compression levels, the said combination achieved impressive AUC scores of c0 (97.11%), c23 (81.19%), and c40 (75.02%). This is due to the fact that YCrCb separates luminance and chrominance, which helps preserve color information under compression; on the other hand, saturation and value components of HSV aid in preserving color details, which especially help in cross-manipulation cases. The XYZ color space also performs steadily across compression ratios from c0 to c40, with overall AUC values ranging from c40 (60.91%) to c0 (94.35%). In short, when we combine the best-performing individual color spaces, the FDN offers robustness to compression.

Further, the in-depth comparison given in Fig. 8 (d1, d2, d3, and d4) scrutinizes the performance of various color spaces while the FDN is trained on NT forgeries at compression levels (c0). In NT, the RGB as an individual color space yields the best intra- and cross-setting overall results at compression ratio c0 (92.64%). However, it exhibits notably less robustness to compression at levels c23 and c40. While LAB grabbed second place and displayed impressive intra-settings performance for NT (c0) at 99.73% as an individual color space, its performance in cross-settings varies significantly, particularly for DF at compression level c0 (96.51%), which shows great robustness, but for FS and F2F, performance decreases significantly. Under increased compression, the YCrCb combination with XYZ enhances detection performance significantly and performs the best at compression levels c23 and c40, achieving an average AUC score of c0 (91.15%), c23 (76.76%), and c40 (71.12%) in intra- and cross-settings. Similar robustness is shown by the combination of LAB and YUV color spaces, as shown in Fig. 8(d4).

In summary, our proposed deepfake detection framework demonstrates robust performance across compression levels c0 and c23 in intra- and cross-manipulation evaluation scenarios, as depicted in Fig. 8 concerning individual or combined color spaces. However, its efficacy reduces at the c40 compression level. The primary reason for this reduction is the lossy compression of test data. Image compression discards some high-frequency image details and textures that contain forgery artifacts, which makes our FDN models unable to extract the features as sensitive as those from the data at compression level c0. Despite this limitation, our incorporation of multiple color space combinations, including HSL, HSV, YCrCb, and YUV, has resulted in noteworthy improvements at the c40 compression level, as evidenced by a margin of 10–15% in Fig. 8 (a3, a4, b3, b4, c3, c4, d3, and d4).

5.1.2. Cross-dataset evaluation in different color spaces

In cross-dataset analysis, as illustrated in Fig. 9, we assess the performance of the proposed forgery detection network (FDN) when it is trained on FF++ [30] dataset and tested on the DFDC [34] and CDF [35] datasets. We focus on identifying the top-performing color spaces and their combinations, which are essential for ensuring generalized performance across diverse environments.

The analysis presented in Fig. 9 reveals that the combination of XYZ and HSV color spaces consistently outperforms other configurations, delivering the highest accuracy across both the DFDC (81.93%) and CDF (86.56%) datasets. This combination stands out as the top choice for cross-dataset deepfake detection. The XYZ and HSL combination secures the second position, demonstrating exceptional performance with high accuracy in both the DFDC (81.15%) and CDF (84.47%) datasets, establishing itself as a robust alternative. The YCrCb and HSV combination takes the third spot, showcasing remarkable performance in both the DFDC (80.65%) and CDF (85.09%) datasets, positioning it among the top choices.

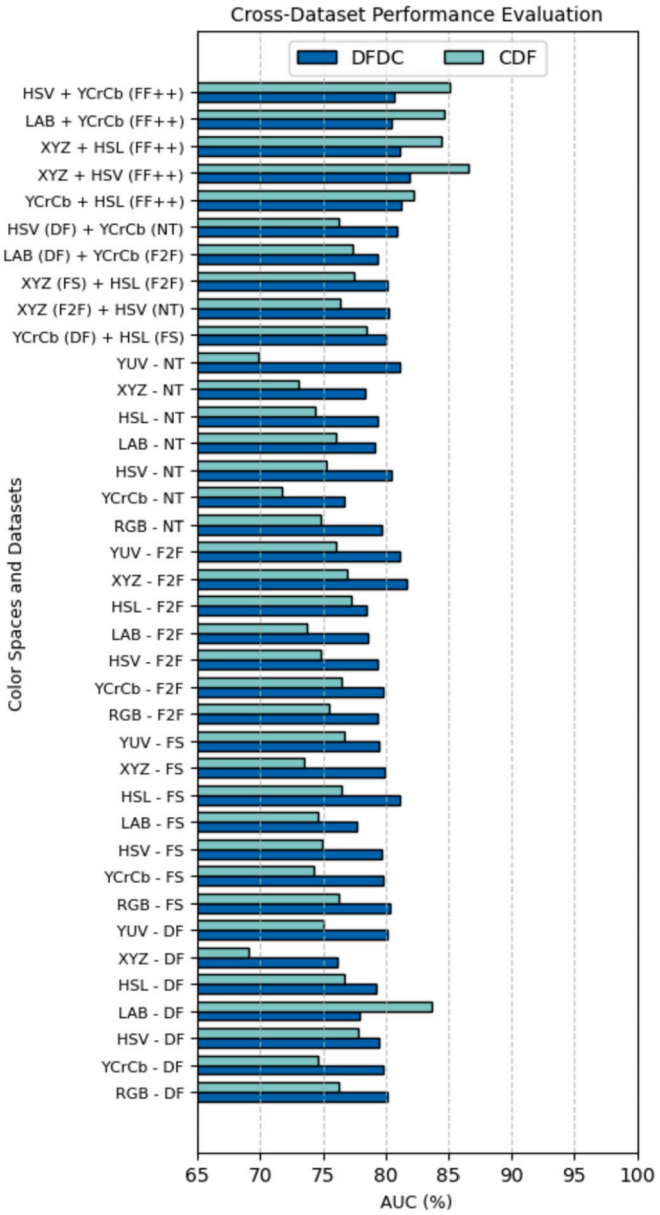


Fig. 9. Our Proposed Method is Evaluated in Cross-Dataset Settings on the DFDC [34] and CDF [35] Datasets by employing the combinations of various Color Spaces. Results are compared in terms of AUC (%).

In a unique scenario, we also explore the different combinations of FF++ [30] subsets for cross-evaluation; specific pairs demonstrate noteworthy performance in terms of color spaces. In one of the many cases, we pair the DF dataset with the HSV color space and the NT dataset with the YCrCb color space. This combination yields competitive results, especially in DFDC (80.94%). Similarly, we combined the FS with XYZ and F2F, along with HSL color space. This quad pair achieved consistent accuracy in both the DFDC (80.08%) and CDF (77.44%) datasets, highlighting robust performance in cross-dataset settings.

Table 4
Our Proposed Method is Evaluated for Intra and Cross-manipulation Settings on the FF++ [30] Subsets. Results are compared with the State-of-the-art Methods in terms of AUC (%).

Train	Methods	Test				Avg.
		DF	FS	F2F	NT	
DF	Xception* [12]	99.32	49.05	73.60	73.61	73.89
	RFM* [13]	98.80	72.69	65.18	63.44	75.02
	LRLNet* [14]	99.45	62.53	75.64	80.73	79.58
	DCL* [15]	99.95	54.19	81.25	90.72	81.53
	FTCN* [16]	99.30	53.50	76.00	87.40	79.05
	RECCE* [17]	99.19	57.42	74.39	85.04	79.01
	CFFE* [18]	99.36	50.00	52.87	61.52	65.93
	SFDG [19]	99.73	75.34	86.45	86.64	87.03
	FCAN-DCT [20]	99.90	65.50	82.50	84.60	83.13
	CADDM [21]	99.38	58.33	83.94	68.98	77.66
	FDN (RGB)	**99.99**	77.15	94.30	**98.36**	92.45
	FDN (HSV)	99.66	**86.99**	**97.61**	93.90	**94.54**
	FDN (YCrCb + HSV)	**99.99**	78.71	96.31	97.58	93.15
FS	Xception* [12]	66.45	99.40	88.83	71.32	81.50
	RFM* [13]	81.34	98.26	61.53	55.02	74.03
	LRLNet* [14]	73.69	99.61	87.24	85.10	86.41
	DCL* [15]	62.00	99.87	84.53	53.80	75.05
	FTCN* [16]	88.10	98.90	69.70	76.70	83.35
	RECCE* [17]	66.66	99.76	73.66	57.46	74.39
	CFFE* [18]	51.99	98.53	52.82	53.09	64.10
	SFDG [19]	81.71	99.53	77.30	60.89	79.85
	FCAN-DCT [20]	69.40	99.90	98.90	98.10	91.57
	CADDM [21]	93.42	99.92	74.00	49.86	79.30
	FDN (RGB)	72.88	99.96	69.38	76.13	79.58
	FDN (HSL)	**99.74**	99.85	96.88	**91.36**	**96.96**
	FDN (HSV + LAB)	97.07	**99.99**	**98.11**	78.06	93.31
F2F	Xception* [12]	80.33	76.25	99.47	69.66	81.42
	RFM* [13]	67.80	64.67	96.44	64.55	73.36
	LRLNet* [14]	81.93	84.60	99.33	86.57	88.10
	DCL* [15]	81.09	54.24	99.02	64.97	74.83
	FTCN* [16]	81.70	76.40	98.90	85.90	85.72
	RECCE* [17]	88.04	67.35	98.93	74.16	82.12
	CFFE* [18]	73.56	50.26	98.35	56.19	69.59
	SFDG [19]	97.38	73.54	99.36	72.61	85.72
	FCAN-DCT [20]	95.70	95.20	99.90	84.40	93.80
	CADDM [21]	99.88	79.40	99.97	82.38	90.41
	FDN (RGB)	95.52	73.52	99.18	88.25	89.12
	FDN (HSV)	99.88	94.60	99.16	92.22	96.47
	FDN (YCrCb + HSV)	**99.90**	**96.04**	**99.18**	**93.30**	**97.11**
NT	Xception* [12]	79.98	73.17	81.36	99.15	83.41
	RFM* [13]	75.39	62.83	72.24	85.51	73.99
	LRLNet* [14]	85.71	89.39	90.43	99.40	91.23
	DCL* [15]	98.27	60.44	71.44	99.10	82.31
	FTCN* [16]	90.00	64.40	88.10	98.00	85.12
	RECCE* [17]	90.20	58.06	76.65	97.17	80.52
	CFFE* [18]	88.69	50.64	65.21	97.26	75.45
	SFDG [19]	91.73	83.58	70.85	99.74	86.47
	FCAN-DCT [20]	96.70	93.30	85.10	98.90	93.50
	CADDM [21]	99.99	97.93	86.76	99.46	96.03
	FDN (RGB)	**99.93**	77.34	**93.77**	99.51	**92.64**
	FDN (XYZ)	99.92	77.73	87.86	**99.71**	91.31
	FDN (LAB + YUV)	99.36	**77.93**	87.98	99.31	91.15

5.2. Comparison with the state-of-the-art methods

This evaluation section compares our proposed deepfake forgery detection framework with ten standard methods. The baseline benchmark method is Xception [12], published in 2017. The other state-of-the-art methods include RFM [13], LRLNet [14], DCL [15], FTCN [16] (published in 2021), RECCE [17], CFFE [18] (published in 2022), SFDG [19], FCAN-DCT [20] and CADDM [21] (published in 2023). The overall comparison is made in two evaluation categories at compression level c0, including intra-, cross-manipulation, and cross-dataset.

5.2.1. Intra- and cross-manipulation evaluation comparison with the state-of-the-art methods

In our thorough intra- and cross-manipulation evaluation, illustrated in Table 4, we analyzed our proposed deepfake forgery detection framework across different training and testing scenarios with ten contemporary methods. When trained on the DF dataset and subsequently tested against the DF, FS, F2F, and NT datasets, our proposed models (denoted with 'FDN') significantly outperformed most state-of-the-art techniques. Notably, HSV exhibited the highest average AUC value of 94.54%, emphasizing the potency of leveraging the HSV color space. From the comparison methods, the SFDG [19] showed a commendable overall performance of 87.03%, attributable to its spatial and frequency feature fusion in a content-aware manner. Lastly, FCAN-DCT [20], which focuses on spatial and temporal frequency features with an attention mechanism, registered an average AUC of 83.13%. Considering the nuances of the DF dataset, our models are adept at discerning manipulations by exploiting color spaces, combined with other framework innovations. The inability of models like RFM [13] (75.02%) to surpass our model underscores the significance of addressing category-level differences rather than just deepfake nuances.

With training on the FS dataset, our HSL-based approach took the lead, reaching an average AUC of 96.96%, showcasing the robustness of harnessing HSL color space in understanding facial manipulations inherent in the FS dataset, as HSL preserves image quality across different manipulations. Again, the FCAN-DCT [20] followed closely with an AUC of 91.57%, highlighting its ability to harness spatial and temporal features in visual and near-infrared lights. Similarly, LRLNet [14], which focuses on spatial and frequency clues, earned a later place with an average AUC of 86.41%. Our proposed FDN model dominance in this training scenario emphasizes their robustness, even when confronted with the FS dataset that primarily targets facial swaps. For all models trained on F2F, the combination of YCrCb with HSV stood out with an astounding average AUC of 97.11%, further solidifying the notion that combining different color spaces can be particularly advantageous. While FCAN-DCT [20] and CADDM [21] achieved notable AUC scores of 93.80% and 90.41%, respectively, showcasing the potential of spatial-temporal combined feature learning and the ID-unaware deepfake detection framework, given the intricacies of the Face2Face manipulation technique, models benefiting from multi-modal feature understanding or those capable of ID-unaware detection exhibited superior performance. In the last case, when models were trained on the NT dataset, the CADDM [21] achieved an outstanding average AUC of 96.03%, underscoring the importance of ID-unaware deepfake detection models when handling neural texture manipulations, followed by FCAN-DCT [20] (93.50%). In third place is our RGB space-based FDN model, whose configuration outperformed our other color spaces based FDN model, which showcased an impressive AUC of 92.64%, indicating the utility of the standard RGB color space for neural textures. The NT dataset, rich in intricate manipulations, posed challenges that were effectively navigated by models tailored to understand neural textures specifically or those with broader, versatile feature-capturing abilities. In short, our rigorous testing against varied datasets highlights the ability of our models to generalize well, primarily due to the exploitation of different color spaces and representative forgery learning.

5.2.2. Cross-dataset evaluation comparison with the state-of-the-art methods

In the cross-dataset evaluation scenario, we train our proposed FDN model on the comprehensive FF++ [30] dataset at compression level c0 and rigorously test its performance against the challenging DFDC [34] and CDF [35] datasets. In Table 5, we compare the performance of our models with other state-of-the-art methods.

Our model showcased remarkable adaptability and robustness when evaluated on the DFDC [34] and CDF [35] datasets. The FDN model trained on the pair of XYZ and HSV color spaces demonstrated the highest AUC of 81.25% when tested on the DFDC [34] dataset, proving the effectiveness of a combination of color spaces in understanding diverse manipulations. The second place was another combination of YCrCb and HSL, performing exceptionally well on the DFDC [34] dataset with an average AUC of 81.25%. In comparison with the recent methods, the FTCN [16] is leading the pack with an AUC of 79.97%, showcasing its commendable generalization ability. While CADDM [21] follows closely with an AUC of 79.57%, indicating its adaptive strength. Also, LRLNet [14], with an AUC of 76.53%, demonstrates that spatial-frequency features based on multi-scale similarity learning are beneficial when dealing with a different source domain dataset. Regarding the DFDC [34] dataset, the superior performance of our proposed FDN models can be attributed to the amalgamation of multiple color spaces. The XYZ and HSV combinations effectively capture subtle inconsistencies and nuances prevalent in deepfakes through the interplay of luminance and chrominance components.

On the CDF [35] dataset, the CADDM [21] is exceptionally proficient, with the best AUC score of 93.88%, outperforming our color spaces-based FDN models. This suggests that the ID-unaware features it harnesses can generalize excellently across face-swapped datasets. Our FDN model achieved second and third places when the XYZ and HSV combination and the HSV pairing with YCrCb were employed, yielding AUC scores of 86.56% and 85.09%, respectively. When we analyzed the other state-of-the-art methods, the FCAN-DCT [20] presented an AUC of 83.46%, indicating its combined spatial and temporal frequency feature fusion strength. Similarly, DCL [15] secured an AUC of 82.30%, further highlighting its intra-frame learning capabilities.

Table 5
Our Proposed Method is Compared against the State-of-the-art Methods in terms of AUC (%) on Cross-Dataset Settings (DFDC [34] and CDF [35]).

Train	Methods	Test		Avg.
		DFDC	CDF	
FF++	Xception* [12]	67.90	59.46	63.68
	RFM* [13]	66.01	65.63	65.82
	LRLNet* [14]	76.53	78.26	77.40
	DCL* [15]	75.24	80.12	77.68
	FTCN* [16]	79.97	79.85	79.91
	RECCE* [17]	68.34	68.94	68.64
	CFFE* [18]	72.09	74.20	73.15
	SFDG [19]	73.64	75.83	74.74
	FCAN-DCT [20]	-	83.46	-
	CADDM [21]	79.57	93.88	86.73
	FDN (XYZ + HSL)	81.15	84.47	82.81
	FDN (LAB + YCrCb)	80.45	84.68	82.56
	FDN (HSV + YCrCb)	80.65	85.09	82.87
	FDN (YCrCb + HSL)	81.25	82.26	81.76
	FDN (XYZ + HSV)	**81.93**	**86.56**	**84.25**

The dominance of CADDM [21] in the CDF [35] dataset reinforces the notion that learning generalized ID-unaware features can be pivotal for unseen deepfake detection, specifically when faced with face swap data variations. In other words, the strong performance on the CDF [35] dataset and limited performance on the DFDC [34] indicate the weakness of the CADDM [21] model. The underlying reason is that the CDF [35] dataset is enhanced deepfakes with improved facial swapping operation, consisting of fewer videos compared to the DFDC [34], which is the largest dataset with a greater number of deepfake generation manipulations. Nevertheless, our FDN model performed consistently on the DFDC [34] and CDF [35] datasets. It is worth noting that our proposed models exhibited resilience and adaptability, effectively dealing with unseen datasets and ensuring that detection capabilities are not confined to a singular training dataset.

5.3. Ablation study

In this study, we present an ablation study to investigate the impact of various components within our proposed forgery detection framework. We consider seven different experimental configurations (No. 1 to 7), each representing the presence or absence of specific components: FDN, MCBN, Color Spaces, and FHN. These configurations are evaluated using DF, FS, F2F, and NT as testing datasets, each corresponding to specific forgeries present in the FF++ [30] dataset. Table 6 summarizes the experimental configurations and their related AUC scores.

The presence of FDN in all configurations (No. 1 to 7) demonstrates its central role in detecting forgeries across various datasets. It serves as the backbone for our deepfake forgery detection framework. Secondly, the inclusion of MCBN (Configurations 2 and 3) leads to performance improvements over Configuration 1, indicating that MCBN enhances the ability of the network to detect manipulations by improving forgery cue representation. Third, experimentation with different color spaces (Configurations 1, 3, 4, 6, and 7) highlights that considering multiple color spaces is beneficial. In this case, specifically, HSV. This diversification allows the network to capture a broader range of manipulation cues, contributing to enhanced performance. Finally, including FHN in Configuration 4 leads to the highest performance across all datasets (average AUC of 94.54%). FHN plays a crucial role in highlighting potential

Table 6
Ablation study is performed by using different combinations of our Proposed Framework components. Our Proposed Method is Trained on FF++ (DF) [30] and Tested against all four subsets of FF++ in terms of AUC (%). For comparison, we are using HSV as a Color Space.

No.	Proposed Framework Modules				Trained and Tested on FF++				
	FDN	MCBN	Color Space	FHN	DF	FS	F2F	NT	Avg.
1	✓				99.99	61.41	90.56	96.51	87.11
2	✓	✓			99.99	70.87	91.09	96.50	89.61
3	✓	✓	✓		99.99	76.86	91.70	94.57	90.78
4	✓	✓	✓	✓	99.66	86.99	97.61	93.90	**94.54**
5	✓	✓		✓	99.99	77.15	94.30	98.36	92.45
6	✓		✓	✓	99.99	77.02	90.95	98.26	91.55
7	✓			✓	99.99	78.52	88.69	96.46	90.91

areas of forgery, aiding the network in accurate detection. Hence, configuration 4, which includes FDN, MCBN, unique color space, and FHN, emerges as the best-performing setup. This configuration combines the core forgery detection capabilities of the FDN with the boosting effect of MCBN. Additionally, the use of multiple color spaces allows for a comprehensive analysis of manipulation cues, and the FHN provides essential guidance by highlighting potential forgery regions. The success of Configuration 4 underscores the importance of combining robust forgery detection architecture, cue enhancement, and multi-modal analysis. It reflects the capacity of our proposed detection framework to excel in identifying various types of forgeries, making it a promising choice for real-world applications where the detection of manipulated media is paramount.

6. Conclusions and future work

In conclusion, this paper presents a comprehensive forgery detection framework that employs multiple color spaces. In our framework, the color spaces are combined based on the properties of their distinct components in a manner that complements other space information to improve detection performance rather than playing a redundant role. The overall operations in the detection framework are comprised of forgery cue representation enhancement, color space transformations, auxiliary supervision to point out the sensitive forgery regions, and employing a dedicated FDN that extracts forgery clues from multiple color space-based handcrafted feature frames. The empirical validation of our approach was conducted on a range of deepfake datasets, such as intra-, cross-manipulation, and robustness evaluations on FF++ subsets. The FDN with the HSV and HSL color spaces achieved outstanding performances compared to state-of-the-art methods. At the same time, YCrCb and YUV demonstrated strong resistance to compression under increased compression ratios. When we combined the four mentioned color spaces, the overall generalized performance and robustness of the proposed FDN increased significantly. On the cross-dataset evaluation, our method also performed consistently, and it can be improved further by expanding it beyond the spatial domain. Over and above, our method exhibits versatility and adaptability in detecting forged elements. Further research into hybrid approaches, multiple feature domains, and adaptive color space selection may lead to more robust forgery detection methods.

CRediT authorship contribution statement

Muhammad Ahmad Amin: Conceptualization of this study, Methodology, Software, Data Curation, Experimentation & Original draft preparation.

Yongjian Hu: Supervision for Overseeing Experiments, Paper drafting, Review & Editing.

Yu Guan: Collaborative Supervision for Experiments, Paper review & Editing.

Muhammad Zain Amin: Resources for Experiments, Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dsp.2024.104426.

References

[1] DeepFakes, https://github.com/deepfakes/, 2019.
[2] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nießner, Face2Face: real-time face capture and reenactment of RGB videos, Commun. ACM 62 (1) (2018) 96–104, https://doi.org/10.1145/3292039, publisher: Association for Computing Machinery Place: New York, NY, USA.
[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks, vol. 63, Association for Computing Machinery, New York, NY, USA, 2020, pp. 139–144.
[4] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 43 (12) (2021) 4217–4228, https://doi.org/10.1109/TPAMI.2020.2970919.
[5] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: 6th International Conference on Learning Representations, ICLRs, 2018.
[6] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, Towards open-set identity preserving face synthesis, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6713–6722.
[7] R. Natsume, T. Yatagawa, S. Morishima, Rsgan: face swapping and editing using face and hair representation in latent spaces, in: ACM SIGGRAPH 2018 Posters, SIGGRAPH '18, Association for Computing Machinery, New York, NY, USA, 2018.
[8] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Advancing high fidelity identity swapping for forgery detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5073–5082.
[9] Y. Li, M.C. Chang, S. Lyu, In Ictu Oculi: exposing AI created fake videos by detecting eye blinking, in: 10th IEEE International Workshop on Information Forensics and Security, WIFS 2018, Institute of Electrical and Electronics Engineers Inc., ISBN 9781538665367, Jan. 2019.
[10] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW 2019, Institute of Electrical and Electronics Engineers Inc., ISBN 9781728113920, 2019, pp. 83–92.
[11] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, DeepRhythm: exposing DeepFakes with attentional visual heartbeat rhythms, in: MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia, ACM Place, New York, NY, USA, ISBN 9781450379885, 2020.
[12] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ISSN 1063-6919, IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 1800–1807.
[13] C. Wang, W. Deng, Representative forgery mining for fake face detection, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14918–14927.

[14] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, R. Ji, Local relation learning for face forgery detection, Proc. AAAI Conf. Artif. Intell. 35 (2) (2021) 1081–1088, https://doi.org/10.1609/aaai.v35i2.16193.
[15] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, R. Ji, Dual contrastive learning for general face forgery detection, Proc. AAAI Conf. Artif. Intell. 36 (2) (2022) 2316–2324, https://doi.org/10.1609/aaai.v36i2.20130.
[16] Y. Zheng, J. Bao, D. Chen, M. Zeng, F. Wen, Exploring temporal coherence for more general video face forgery detection, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 15024–15034.
[17] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, X. Yang, End-to-end reconstruction-classification learning for face forgery detection, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4103–4112.
[18] P. Yu, J. Fei, Z. Xia, Z. Zhou, J. Weng, Improving generalization by commonality learning in face forgery detection, IEEE Trans. Inf. Forensics Secur. 17 (2022) 547–558, https://doi.org/10.1109/TIFS.2022.3146781.
[19] Y. Wang, K. Yu, C. Chen, X. Hu, S. Peng, Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 7278–7287.
[20] Y. Wang, C. Peng, D. Liu, N. Wang, X. Gao, Spatial-temporal frequency forgery clue for video forgery detection in VIS and NIR scenario, IEEE Trans. Circuits Syst. Video Technol. 33 (12) (2023) 7943–7956, https://doi.org/10.1109/TCSVT.2023.3281475.
[21] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, Z. Ge, Implicit Identity Leakage: the stumbling block to improving deepfake detection generalization, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 3994–4004.
[22] H. Li, B. Li, S. Tan, J. Huang, Identification of deep network generated images using disparities in color components, Signal Process. 174 (2020) 107616, https://doi.org/10.1016/j.sigpro.2020.107616.
[23] L. Busin, N. Vandenbroucke, L. Macaire, Color spaces and image segmentation, in: Advances in Imaging and Electron Physics, Elsevier, 2009, pp. 65–168.
[24] S.N. Gowda, C. Yuan, Colornet: investigating the importance of color spaces for image classification, in: Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Springer-Verlag, Berlin, Heidelberg, 2018, pp. 581–596, Revised Selected Papers, Part IV.
[25] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu, Spatial-phase shallow learning: rethinking face forgery detection in frequency domain, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 772–781.
[26] Y. Wang, C. Peng, D. Liu, N. Wang, X. Gao, Forgerynir: deep face forgery and detection in near-infrared scenario, IEEE Trans. Inf. Forensics Secur. 17 (2022) 500–515, https://doi.org/10.1109/TIFS.2022.3146766.
[27] D. Liu, Z. Dang, C. Peng, Y. Zheng, S. Li, N. Wang, X. Gao, Fedforgery: generalized face forgery detection with residual federated learning, IEEE Trans. Inf. Forensics Secur. 18 (2023) 4272–4284, https://doi.org/10.1109/TIFS.2023.3293951.
[28] C. Villani, The Wasserstein Distances, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 93–111.
[29] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 214–223, https://proceedings.mlr.press/v70/arjovsky17a.html.
[30] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Faceforensics++: learning to detect manipulated facial images, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11.
[31] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: introducing convolutions to vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 22–31.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[33] J. Lei Ba, J.R. Kiros, G.E. Hinton, Layer normalization, arXiv e-prints, arXiv:1607.06450, 2016, https://doi.org/10.48550/arXiv.1607.06450.
[34] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The DeepFake detection challenge (DFDC) dataset, arXiv:2006.07397v4, Jun. 2020.
[35] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: a large-scale challenging dataset for deepfake forensics, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3204–3213.
[36] A.P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159, https://doi.org/10.1016/S0031-3203(96)00142-2.
[37] Z. Zhang, M.R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, pp. 8792–8802, https://par.nsf.gov/biblio/10083366.

Muhammad Ahmad Amin received an M.E. degree in Information and Communication Engineering from the South China University of Technology (SCUT), China, in 2018. He is currently pursuing his Ph.D. degree in information and communication engineering at SCUT. His research interests include image forensics, information security, and artificial intelligence.

Yongjian Hu is a professor at the School of Electronic and Information Engineering, South China University of Technology, China. He is a senior member of the IEEE and has published over 130 peer-reviewed papers. His research interests include image forensics, information security, and deep learning.

Yu Guan is an associate professor in the Dept. of Computer Science, University of Warwick, U.K. He has published 60+ peer-reviewed papers, including at top venues like IEEE T-PAMI, T-IP, CVPR, ECCV, ACM IMWUT, ACM Multimedia, etc. His research agenda is centered on activity recognition, AI healthcare, wearable computing, computer vision, and applied machine learning.

Muhammad Zain Amin is currently enrolled in the Erasmus Mundus Joint Master Degree in Medical Imaging and Application at the University of Burgundy, France; the University of Cassino, Italy; the University of Girona, Spain; and Duke University, USA. He received his bachelor degree in computer science from the University of Engineering and Technology, Lahore, Pakistan. His primary research interests include image forensics, information security, medical imaging, and deep learning.