Fusion-SSAT: Unleashing the Potential of Self-supervised Auxiliary Task by Feature Fusion for Generalized Deepfake Detection

Shukesh Reddy, Srijan Das, Abhijit Das, *Senior Member IEEE*

Abstract—In this work, we attempted to unleash the potential of self-supervised learning as an auxiliary task that can optimise the primary task of generalised deepfake detection. To explore this, we examined different combinations of the training schemes for these tasks that can be most effective. Our findings reveal that fusing the feature representation from self-supervised auxiliary tasks is a powerful feature representation for the problem at hand. Such a representation can leverage the ultimate potential and bring in a unique representation of both the self-supervised and primary tasks, achieving better performance for the primary task. We experimented on a large set of datasets, which includes DF40, FaceForensics++, Celeb-DF, DFD, FaceShifter, UADFV, and our results showed better generalizability on cross-dataset evaluation when compared with current state-of-the-art detectors.

Index Terms—Generalized Deepfake Detection, Self-supervised Learning, Auxiliary Task, Local Directional pattern.

✦

1 INTRODUCTION

RAPID increase in the use of Large Language Models (LLMs)[47], [19], deep denerative models[81] and diffusion models[56] has revolutionized content creation across text, image, video, and audio domains. These advancements enabled the ease in creating of extraordinarily realistic synthetic media, usually indistinguishable from real content. Accelerated growth in deepfake generation techniques[30], [70], [55], [9], [49], [13], [48], [4], [62], [54], [2], [26], [27] with respective to Face swapping[62], [73], Facial expressions[2], [9], [13], face reenactment[4], [28], [31], [39], talking face generation[27], [80], [59], [67], facial attribute editing[15], [42], [77], [71] from a source image or video to a target video. This forgery poses a risk to subject identifications that have been widely employed in digital payments, video surveillance, and social media. To mitigate these concerns, there is a rapidly expanding research on deepfake detectors for the identification of facial forgeries. As a result, many research efforts have focused on developing methods to detect manipulated media using CNN and RNN-based models: XceptionNet[14], EfficientNetB7[63]; Transformer-Based Models: vision transformers (ViT)[20], swin transformer[41].

Although current state-of-the-art deepfake detectors, UCF[75], RECCE[6], CORE[46], FFD[16], Face X-Raycite[34], DSP-FWA[35], Capsule[45], and XceptionNet[14] exhibit strong performance in within-domain evaluations, they demonstrate a significant decrease in performance when employed in cross-domain or cross-dataset analyses.

State-of-the-art models[10], [43], [6], [45], [51], [35], [16], [34], [65], [7], [68], [9]often struggle to generalize across datasets, as they tend to learn superficial, dataset-specific patterns rather than learning the underlying generative mechanisms of manipulated content. The over-reliance on dataset-specific artefacts limits their ability to perform reliably in cross-dataset evaluations. The resulting drop in performance reveals a critical shortcoming in the model's ability to learn robust and transferable features that hold up across varied domains.

While these models may achieve high accuracy in controlled or domain-specific settings [75], [46], they frequently overfit to idiosyncratic cues present in the training data, such as compression artefacts or content regularities thereby diminishing their practical utility. In real-world scenarios, where deepfakes vary widely in terms of generation methods, compression levels, and content diversity, such limitations become especially pronounced, underscoring the need for more resilient and generalizable detection approaches.

To mitigate the above-mentioned limitation, we proposed a novel approach, Fusion-SSAT, which consists of an auxiliary task reconstruction and deepfake detection as the main task. The reconstruction task aims to learn the local texture features from a given masked RGB face image. The primary task is a global feature encoder for binary classification of deepfake detection. We fuse the feature from the encoder of the auxiliary task along with the feature from the encoder of the primary detection task to blend both the global and local features. This fusion of local and global features significantly enhances the performance of deepfake detection across datasets, as the reconstruction of local texture patterns from the corresponding RGB pushes the encoder to identify intricate face artefacts and distortions created during manipulation, hence empowering the detection feature representation. To be specific, this significantly more arduous process of featuring is prompted to acquire manipulation resilient representations. Hence, it served as a means of implicit regularisation, pushing the model to more effectively detect local cues along with the global cues that often differ between real and fake content.

To summarize, our contributions are as follows:

● *S. Reddy, A. Das are with the Machine Intelligence Group, Department of CS&IS, Birla Institute of Technology and Sciences, Pilani, India.*
S. Das, University of North Carolina at Charlotte, United States of America.
E-mail: abhijit.das@hyderabad.bits-pilani.ac.in

- A novel approach, Fusion-SSAT, to improve deepfake detection and to generalization in cross-domain deepfake detection scenarios.
- Demonstration of Fusion-SSAT's ability to learn fine-grained texture and artefact patterns via local texture pattern from its corresponding face images.
- Extensive evaluation on DF40 and deepfake benchmark datasets validating the effectiveness of the proposed approach.

2 RELATED WORKS

In deepfake detection, CNN's are frequently preferred over alternative architectural models. Preliminary research in deepfake detection has primarily utilised known CNN architectures, such exception-net[14] and efficientnet B4[63], as binary classifiers. Nonetheless, these techniques encounter difficulties in generalizing to unseen manipulation methods. To tackle this difficulty, various solutions have been explored, including disentanglement learning, multi-task learning[18], [82], [3], and pseudo fake synthesis[44], [61]. Advancements in generative modeling have led to deepfakes becoming increasingly realistic, resulting in imperceptible localized errors. The previously described methods are susceptible to obsolescence. They continue to depend on conventional CNNs[14], [63], [58], hence experiencing a loss of local information with successive convolutional layers. To enhance the capture of low-level features, techniques employing implicit attention strategies[65] have been suggested nevertheless, they exhibit inadequate generalization.

Recent advancements in vision transformers (ViTs)[20], [17], [32], [3] have resulted in the creation of multiple types for diverse applications. Optimal performance of these models necessitates comprehensive datasets and pre-training. The DeiT model[65] was improved to reduce these requirements by advanced regularization, data augmentation, and token extraction from convolutional layers. The tokenisation method utilised by T2T[79] was implemented to discern and document local structural information. Convolutional filters were employed in alternative models to incorporate inductive bias. Hierarchical transformers[8] have incorporated inductive bias by reducing the number of tokens through patch merging. Nonetheless, they still require datasets of considerable size for pre-training.

SSL has shown effective usage in image-based techniques such as SimCLR[10], MoCo[24], MAE[23], and DINO[7], as well as in video-based methodologies like CoCLR[22], VideoMAE[64], and SVT[52]. SSL emphasis the acquisition of representations by utilizing the intrinsic structure of data, hence obviating the dependence on labeled data. This approach has demonstrated considerable enhancements in generalizability, with current research exceeding the performance of supervised models in zero-shot testing. Leveraging these developments, we expect that ssl will facilitate the model's acquisition of superior representations to improve generalizability in deep fake detection. The SSL-based method enables the model to accurately differentiate between authentic and counterfeit instances, regardless of the manipulation strategy utilized.

MTL has been extensively studied across multiple domains of machine learning and deep learning applications. It has been utilized in natural language processing applications, including unified representations and representation learning. Furthermore, MTL has been utilized in voice recognition, drug discovery, and computer vision applications such as facial analysis, pedestrian identification, facial alignment, and attribute prediction, among others. Moreover, MTL has recently acquired importance in face attribute learning, sometimes referred to as semantic features, as they offer a more authentic depiction of objects and actions. This method facilitates a thorough comprehension of the visual domain by concurrently modeling several aspects in face-related activities. Although MTL is frequently examined through Supervised Learning, its investigation within the framework of SSL is still mostly uncharted, rendering it particularly pertinent to the current issue. Furthermore, it is important to highlight that deep fake detection has not been thoroughly examined within the MTL framework.

3 PROPOSED METHODOLOGY
3.1 Preliminaries

Self-Supervised Auxiliary Task (SSAT)[18]: Jointly optimizes ViTs for the Primary task (classification) and a self-supervised auxiliary task (reconstruction) when the amount of training is limited. The primary task utilizes the latent representation of full image for the classification, while masked image is used for the reconstruction task. The framework is jointly trained on losses of primary task denoted by L_{cls} and auxiliary task denoted by L_{SSAT}, where λ is set to 0.1 thereby assigning a higher relative weight of 0.9 to L_{SSAT} during training shown in equation 1. Figure 1 provides an overview of the framework.

$$L = \lambda \cdot L_{\text{cls}} + (1 - \lambda) \cdot L_{\text{SSAT}} \tag{1}$$

SSAT showed a better generalizability in high quality deepfakes but fails to learn the local and global features from highly compressed deepfake's.

Local Texture Features: encode fine-grained spatial and directional variations in facial regions, offering robustness against illumination changes, noise, and image compression. By analysing edge responses across extended neighbourhoods, these descriptors preserve intricate local structures around critical regions such as the eyes, nose, and mouth. Advanced sampling and directional weighting enhance their ability to represent subtle structural differences that may be overlooked by global descriptors. This enables local texture features to serve as a powerful tool in applications that demand precise discrimination of subtle and fine-grained visual patterns.

3.2 Fusion-SSAT

In contrast to state-of-the-art deepfake detectors, training solely on classification tasks will limit the ability to understand various patterns of deepfake artifacts. In Fusion-SSAT, we introduced a multitask learning framework where in RGB video serves as input for the primary task of deep fake detection i.e real/ fake classification, conducted in a supervised manner, while LDP video/ local texture pattern is reconstructed as auxiliary task executed in a self-supervised manner, with shared features between the tasks. To enhance the generalisation of deepfakes, we train the classifier using a combination of features from the encoder of the auxiliary task and global features from RGB, which facilitates improved pattern recognition compared to training only on global features.

Let $V = \{v_1, v_2, \ldots, v_n\}$ represent a dataset of video sequences, where each v_i is a video comprising a sequence of frames organized as a tensor $v_i \in \mathbb{R}^{B \times T \times C \times H \times W}$. Here, B denotes the batch size, T is the number of frames, C is the
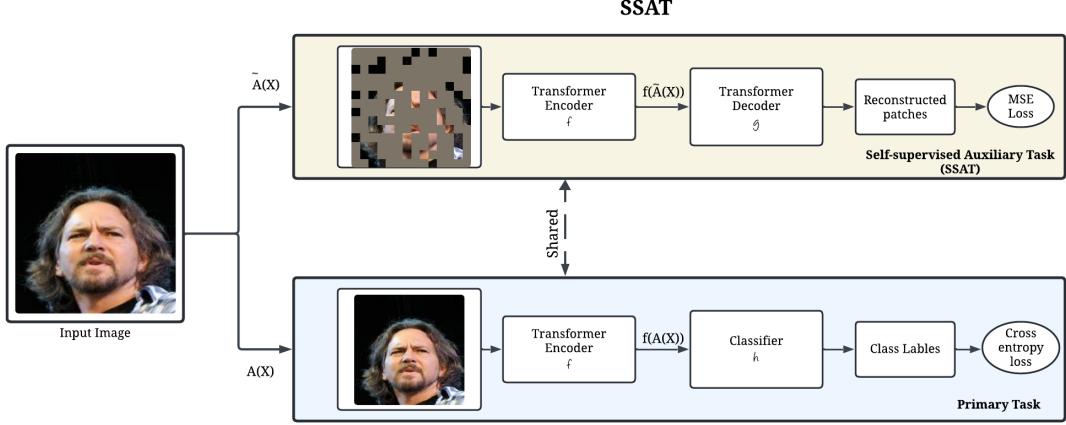
Fig. 1: Overview of the self-supervised auxiliary task (SSAT) framework.

number of channels, H is the frame height, and W is the frame width. Each video is provided in two forms: a masked RGB video $L(v_i)$ and an full patch RGB video $R(v_i)$, both with dimensions $\mathbb{R}^{B \times T \times C \times H \times W}$. The RGB video undergoes random masking with a masking ratio of 0.75, denoted as $\tilde{L}(v_i) = M(L(v_i))$, where M randomly masks 75% of the spatio-temporal patches.

Both $\tilde{L}(v_i)$ and $R(v_i)$ are processed by a shared ViT encoder $f : \mathbb{R}^{B \times T \times C \times H \times W} \to \mathbb{R}^{B \times S \times D}$, where S is the number of encoded tokens, and D is the dimension of the latent representation. The encoder generates latent representations:

$$f(R(v_i)) \in \mathbb{R}^{B \times S \times D} \quad \text{for the RGB video,}$$

$$f'(\tilde{L}(v_i)) \in \mathbb{R}^{B \times S' \times D} \quad \text{for the masked RGB video,}$$

where $S' \leq S$ due to the masking operation, and f' denotes the encoder operating on masked inputs.

Primary Task of Deepfake Classification The latent representation $f(R(v_i))$ of the RGB video is passed to a classifier $h : \mathbb{R}^{B \times S \times D} \to \mathbb{R}^{B \times 2}$ for binary deepfake classification. Each video v_i is associated with a ground-truth label $y_i \in \{0, 1\}$, where $y_i = 1$ indicates a deepfake, and $y_i = 0$ indicates an authentic video. The classification loss is computed using cross-entropy:

$$L_{\text{cls}} = -\frac{1}{B} \sum_{i=1}^{B} [y_i \log(h(f(R(v_i)))_1) + (1 - y_i) \log(h(f(R(v_i)))_0)]$$

where $h(f(R(v_i)))_k$ denotes the predicted probability for class $k \in \{0, 1\}$.

Auxiliary Task: Masked LDP2Video Prediction task The latent representation $f'(\tilde{L}(v_i))$ of the masked RGB video is passed to a shallow decoder $g : \mathbb{R}^{B \times S' \times D} \to \mathbb{R}^{B \times T \times C \times H \times W}$, inspired by the VideoMAE approach[64]. The decoder reconstructs the local texture patches from original RGB $L(v_i)$ masked video representation. Following the MAE framework, g takes $f'(\tilde{L}(v_i))$ and learnable masked tokens as input. Each token at the decoder's output is linearly projected to a vector of pixel values representing a patch. The reconstructed video is:

$$\hat{L}(v_i) = g(f'(\tilde{L}(v_i))).$$

The reconstruction loss is the mean squared error (MSE) computed only over the masked patches:

$$L_{\text{rec}} = \frac{1}{|\mathcal{M}_i|} \sum_{(t,h,w) \in \mathcal{M}_i} \left\| L(v_i)[:, t, :, h, w] - \hat{L}(v_i)[:, t, :, h, w] \right\|_2^2$$

where $\mathcal{M}_i$ is the set of masked patch indices in $\tilde{L}(v_i)$, and $|\mathcal{M}_i|$ is the number of masked patches.

The framework jointly optimizes the primary classification task and the auxiliary reconstruction task using a convex combination of losses:

$$L = \lambda * L_{\text{cls}} + (1 - \lambda) * L_{\text{rec}}$$

where $\lambda = 0.1$, assigning a weight of 0.1 to the classification loss and 0.9 to the reconstruction loss, emphasizing the auxiliary task.
Fusion strategy: A joint feature embedding strategy is introduced, where features from both the LDP[29] and RGB denoted as $f'(\tilde{L}(v_i))$ and $f(R(v_j))$, respectively are combined via element-wise multiplication: $z = f'(\tilde{L}(v_i)) \odot f(R(v_j))$. The fused feature vector z is then fed into the classifier h. This joint embedding enables the classifier to concurrently leverage fine-grained texture information from LDP features[29] and high-level semantic cues from RGB inputs, thereby improving classification performance compared to the earlier independent processing approach as shown in Figure 2 This formulation ensures that the shared ViT encoder[20] learns robust representations for deepfake detection from RGB videos while leveraging self-supervised learning to reconstruct masked patches in LDP videos, enhancing generalization and capturing texture-based features critical for deepfake analysis.

4 EXPERIMENTAL RESULTS

In this section, we described the experiments' details and results on the proposed model Fusion-SSAT in comparison with state-of-the-art deepfake detectors. Section 4.1 describes the details of datasets on which proposed models has been evaluated, While Section 4.2 describes the implementation details, hardware and evaluation metrics used in reporting the performance of proposed models along with state-of-the-art models. Section 4.3 describes results comparison between proposed model with existing state-of-the-art models.

4.1 Datasets

We evaluted our proposed model Fusion-SSAT on FaceForensics++[57], Celeb-DF v1[36], Celeb-DF v2[36], FaceShifter[33], deep fake detection[57], UADFV[78], DF40[74] datasets. All proposed models are trained on FF++ c23, and tested
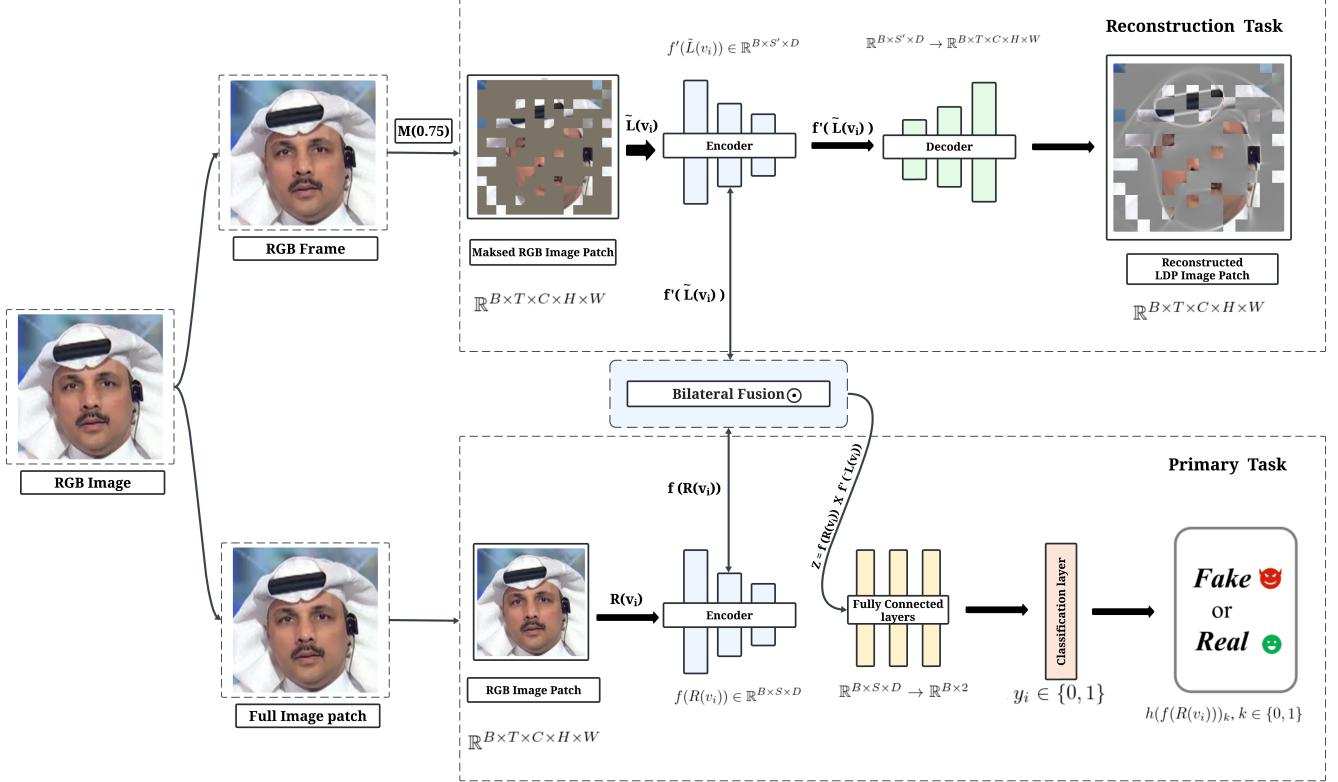
Reconstruction Task

$f'(\tilde{L}(v_i)) \in \mathbb{R}^{B \times S' \times D}$ $\mathbb{R}^{B \times S' \times D} \to \mathbb{R}^{B \times T \times C \times H \times W}$

M(0.75) $\tilde{L}(v_i)$ Encoder $f'(\tilde{L}(v_i))$ Decoder

RGB Frame   Maksed RGB Image Patch

$\mathbb{R}^{B \times T \times C \times H \times W}$ $f'(\tilde{L}(v_i))$

Reconstructed LDP Image Patch

$\mathbb{R}^{B \times T \times C \times H \times W}$

RGB Image

Bilateral Fusion $\odot$

$f(R(v_i))$

Primary Task

$R(v_i)$ Encoder $z_i = f(R(v_i)) \times f'(\tilde{L}(v_i))$ Fully Connected layers   Classification layer

Fake 😈 or *Real* 😊

Full Image patch   RGB Image Patch

$f(R(v_i)) \in \mathbb{R}^{B \times S \times D}$ $\mathbb{R}^{B \times S \times D} \to \mathbb{R}^{B \times 2}$ $y_i \in \{0, 1\}$ $h(f(R(v_i)))_k, k \in \{0, 1\}$

$\mathbb{R}^{B \times T \times C \times H \times W}$

Fig. 2: Overview of the proposed Fusion-SSAT approach

on cross-domain datasets mentioned. Detailed description of each dataset is given below:

FaceForensics++[57]: is a huge benchmark dataset that consists of 1000 real videos and 4000 fake videos, while 720 is used as the train split, 140 each for test and validation splits. Dataset contains manipulations created with state-of-the-art methods, namely, Face2Face, FaceSwap, DeepFakes, and Neuraltextures. There exist three variants of FF++ relating to video compression levels i.e raw, lightly compressed (c23), and highly compressed (c40).

Celeb-DF v1, v2[36]: includes high quality face-swapping videos, while Celeb-DF v1 has 408 original videos and 795 fake videos while v2 has 590 original videos and 5639 fake videos.

Faceshifter[33]: is a method to created high fidelity and occlusion aware face swapping videos. It is a subset of the FaceForensics++ dataset, consists of 1000 fake videos.

Deep Fake Detection[57]: is a dataset which is developed by google, includes 363 real videos and 3000 fake videos. Now it's available as part of FaceForensics++ dataset. **UADFV[78]:** dataset consists of 98 videos, in which 49 videos are real and 49 videos are fake.
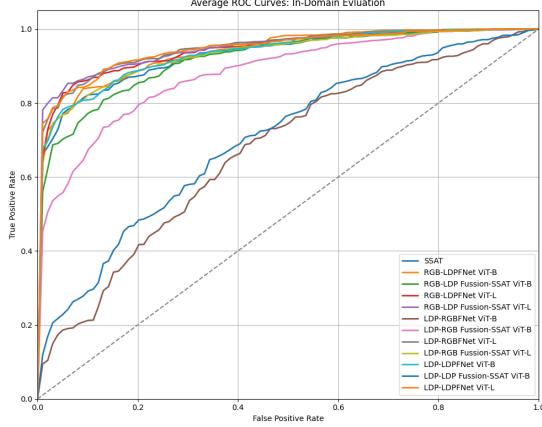
DF40[74]: is a huge dataset which compress different state-of-the-art deepfake generation techniques Contains 0.1M+ fake videos. Used real videos from FaceForensics++ and Celeb-DF dataset's for creating fake videos. Dataset is broadly categorized into 4 parts namely, Face-Swapping (FS), Face-Reenactment (FR), Entire Face Synthesis (EFS), Face Editing (FE). DF40 is created by 40 deepfake techniques, including 10 FS methods, 13 FR methods, 12 EFS methods, and 5 FE methods.
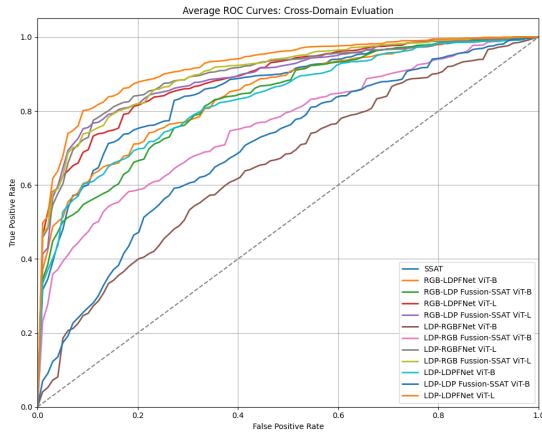
4.2 Implementation details and Evaluation Metrics

All proposed models are trained on images/videos of dimension 224 x 224 pixels using patches of 16 x 16 pixels as input (LDP or RGB) to the respective task (classification or reconstruction) of the model. The training is conducted in batch size of 8 on an NVIDIA A100 GPU utilizing the stochastic gradient descent optimizer for weight adjustments. The learning rate is determined by a cosine schedule that starts at 0.00005 and targets a minimum learning rate of 1e-6. The learning rate is consistently diminished by incorporating a decay rate of 0.05 across each of the 75 training epochs the model experiences. The stochastic depth is implemented with an initial drop path rate of 0.01, and the different loss terms are balanced using a Lambda (λ) value of 0.1. A masking ratio of 0.75 has been established. The official partitions are utilized for all datasets.

4.3 Results and discussion

We have evaluated our proposed model Fusion-SSAT in two different scenario's. In section 4.3.1 all models are trained on FaceForensics++ c23[57], and tested on in-domain and cross-domain datasets which includes Celeb-DF v1[36], Celeb-DF v2[36], DFD[57], FaceShifter[33], UADFV[78]. In section 4.3.2 all model's are trained on DF40 train set of FF++ domain with FS (FF), FR(FF), EFS(FF) forgery data and evaluated on within-forgery, cross-forgery (FF)[57] and cross-domain(CDF)[36] of all forgery methods. In our proposed model, Fusion-SSAT naming convention, the modality preceding the hyphen signifies the masked video input supplied to the reconstruction task, whereas the modality following the hyphen indicates the domain in which the prediction occurs. For instance, In RGB-LDP Fusion-SSAT, the masked video input originates from the RGB modality, while

(a) In-domain evaluation on FF++ dataset

(b) Cross-domain evaluation on CDF, DFD, UADFV datasets

Fig. 3: ROCs for in-domain and cross-domain evaluations.

(a) FS (FF)

(b) FR (FF)

(c) EFS (FF)

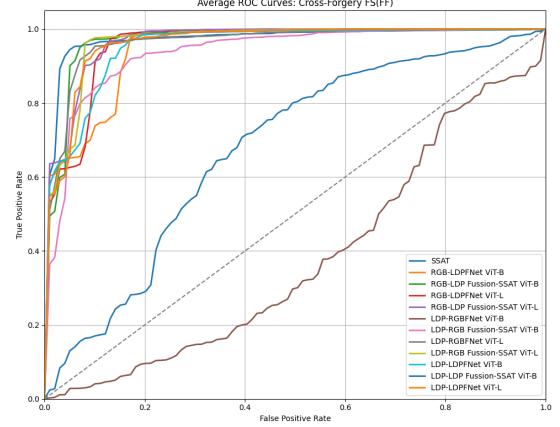Fig. 4: ROCs of cross-forgery evaluations on DF40 dataset.

the model predicts in the LDP. We employed the AUC as the evaluation metric. The best AUC scores achieved by our models are underlined, while the best SOTA results are *italicized*, and the highest-performing method in each column is indicated in **bold**.

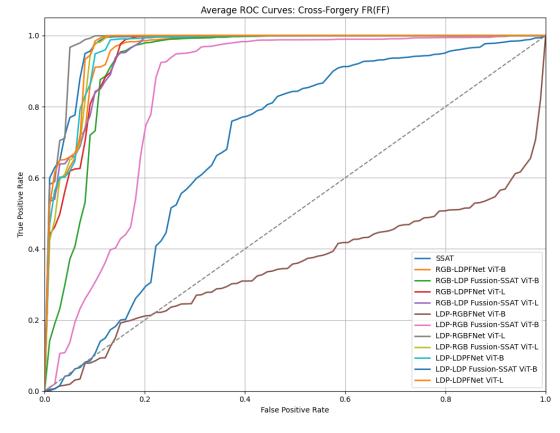### 4.3.1 Experimental results on Within-Domain and Cross-Domain

Table 1 exhibits the experimental results of our proposed model trained on FF++ c23[57] (Deepfakes(FF-DF), Face2Face(FF-F2F), FaceSwap(FF-FS), Neuraltextures(FF-NT)) and evaluated on within domain evaluation and cross domain datasets. We benchmark our results against deepfakebench's[76] spatial detector's which exhibited better performance than existing state-of-the-art deepfake detectors.

Column-1 defines the specifications of the detector, whereas Column-2 defines the backbone network employed in the respective detector. Column-3, 4 defines the experimental results (AUC score) of with domain evaluation and cross-domain evaluation, that includes test scores on the complete datasets FF-c23[57], FF-c40[57], CDFv1[36], CDFv2[36], DFD[57], Fsh[33], UADFV[78] and specific forging methods. The 'Avg' column outlines the performance of each detector over all datasets, both within-domain and cross-domain.

Our proposed model Fusion-SSAT, demonstrates superior generalizability compared to the detectors listed in Table 1. Among all

the above mentioned state-of-the-art detectors, UCF exhibits the highest average at 0.9527%, however our technique, Fusion-SSAT, achieves an AUC score of 0.9613%, demonstrating a 2% above the existing state-of-the-art in within domain evaluation. Our model Fusion-SSAT showed superior performance in individual evaluations of each forgery method, specifically in FF-c23, FF-DF, FF-F2F, FF-FS, and FF-NT. In identifying compressed deepfake FF-c40, our model Fusion-SSAT exhibits a performance decrease of 0.005%. In Cross-Domain Evaluation, each dataset demonstrates a 5% improvement, and on average, we achieved an 8% higher AUC

TABLE 1: *In-domain and cross-domain evaluations using the AUC metric. All detectors are trained on FF-c23 and evaluated on other data. "Avg." donates the average AUC for within-domain and cross-domain evaluation, and the overall results.*

Detector	Backbone	In-Domain Evaluation							Cross Domain Evaluation					
		FF-c23	FF-c40	FF-DF	FF-F2F	FF-FS	FF-NT	Avg.	CDFv1	CDFv2	DFD	Fsh	UADFV	Avg.
Meso4 [1]'WIFS-18	MesoNet[1]	0.6077	0.5920	0.6771	0.6170	0.5946	0.5701	0.6097	0.7358	0.6091	0.5481	0.5660	0.7150	0.6348
MesoIncep [1]'WIFS-18	MesoNet[1]	0.7583	0.7278	0.8542	0.8087	0.7421	0.6517	0.7571	0.7366	0.6966	0.6069	0.6438	0.9049	0.7177
CNN-Aug [69]'CVPR-20	ResNet[25]	0.8493	0.7846	0.9048	0.8788	0.9026	0.7313	0.8419	0.7420	0.7027	0.6464	0.5985	0.8739	0.7127
Xception [57]'ICML-19	Xception[14]	0.9637	0.8261	0.9799	0.9785	0.9833	0.9385	0.9450	0.7794	0.7365	0.8163	0.6249	0.9379	0.7790
EfficientB4 [63]'ICCV-19	Efficient[63]	0.9567	0.8150	0.9757	0.9758	0.9797	0.9308	0.9389	0.7909	0.7487	0.8148	0.6162	0.9472	0.7835
Capsule[45]'CASSP-19	Capsule[45]	0.8421	0.7040	0.8669	0.8634	0.8734	0.7804	0.8217	0.7909	0.7472	0.6841	0.6465	0.9078	0.7553
FWA[35]'CVPRW-19	Xception[14]	0.8765	0.7357	0.9210	0.9000	0.8843	0.8120	0.8549	0.7897	0.6680	0.7403	0.5551	0.8539	0.7214
X-ray[34]'CVPR-20	HRNet[68]	0.9592	0.7925	0.9794	0.9872	0.9871	0.9290	0.9391	0.7093	0.6786	0.7655	0.6553	0.8989	0.7415
FFD[16]'CVPR-20	Xception[14]	0.9624	0.8237	0.9803	0.9784	0.9853	0.9306	0.9434	0.7840	0.7435	0.8024	0.6056	0.9450	0.7761
CORE[46]'CVPRW-20	Xception[14]	0.9638	0.8194	0.9787	0.9803	0.9823	0.9339	0.9431	0.7798	0.7428	0.8018	0.6032	0.9412	0.7738
Recce[6]'CVPR-20	Designed	0.9621	0.8190	0.9797	0.9779	0.9785	0.9357	0.9422	0.7677	0.7319	0.8119	0.6095	0.9446	0.7731
F3Net [50]'ECCV-20	Xception[14]	0.9635	0.8271	0.9793	0.9796	0.9844	0.9354	0.9449	0.7769	0.7352	0.7975	0.5914	0.9347	0.7614
SPSL [40]'CVPR-21	Xception[14]	0.9610	0.8174	0.9781	0.9754	0.9829	0.9299	0.9408	0.8150	0.7650	0.8122	0.6437	0.9424	*0.7956*
SRM [43]'CVPR-21	Xception[14]	0.9576	0.8114	0.9733	0.9696	0.9740	0.9295	0.9359	0.7926	0.7552	0.8120	0.6014	0.9427	0.6222
UCF[75]'ICCV-23	Xception[14]	0.9705	0.8399	0.9883	0.9840	0.9896	0.9441	*0.9527*	0.7793	0.7527	0.8074	0.6462	0.9528	*0.7877*
MARLIN [5]'CVPR-23	ViT-L	0.9370	0.8059	0.9399	0.9250	0.9580	0.9708	0.9228	0.7140	0.7030	0.6850	0.6671	0.8150	0.7168
AVAD [21]'CVPR-23	Designed	0.9410	0.7980	0.9872	0.9534	0.9267	0.9093	0.9193	0.7382	0.7089	0.6892	0.7120	0.8910	0.7479
DF-Adapter [60]'IJCV-24	ViT/ViT-Adapter[20]	0.9862	0.9683	0.9984	0.9976	0.9933	0.9597	*0.9839*	0.7174	0.7071	0.7194	0.7292	0.8948	0.7536
StyleGRU [12]'CVPR-24	StyleGRU[12]	0.8297	0.7091	0.9400	0.6850	0.8880	0.8060	0.8297	0.6250	0.6051	0.7220	0.9070	0.8145	0.7347
Fairness [38]'CVPR-24	Xception[14]	0.9828	0.8091	0.9905	0.9865	0.9923	0.9635	0.9541	0.7442	0.7002	0.8482	0.8342	0.8523	0.7958
OPR [11]'NeurIPS-24	EfficientNetB4[63]	0.9591	0.6078	0.9861	0.9272	0.9272	0.9071	0.8858	0.9094	0.8448	0.8091	0.7190	0.8581	0.8281
PFF [37]'ICLR-24	Xception[14]	0.8518	0.6982	0.8591	0.9281	0.8271	0.8561	0.8367	0.7721	0.7041	0.7826	0.6982	0.8125	0.7539
SSAT [18]'WACV-24	VideoMAE/ViT-B	0.9682	0.7438	0.9991	0.9603	0.9922	0.9215	0.9308	0.8073	0.7757	0.8109	0.9902	0.8920	0.8552
LDP-RGBFNet[53]'ICPRW-24	VideoMAE/ViT-B	0.6766	0.6566	0.8622	0.6047	0.6112	0.6281	0.6732	0.4607	0.6828	0.5654	0.8052	0.7344	0.6497
LDP-RGBFNet[53]'ICPRW-24	VideoMAE/ViT-L	0.9868	0.7687	0.9998	0.9871	0.9972	0.9631	0.9505	0.8256	0.8165	0.8973	0.9933	0.9545	0.8974
LDP-LDPFNet[53]'ICPRW-24	VideoMAE/ViT-B	0.9687	0.7727	0.9995	0.9667	0.9893	0.9193	0.9360	0.7676	0.7499	0.7537	0.9915	0.8971	0.8319
LDP-LDPFNet[53]'ICPRW-24	VideoMAE/ViT-L	0.9857	0.7673	0.9996	0.9844	0.9969	0.9617	0.9493	0.9193	0.8676	0.8766	0.9946	0.9517	0.9219
RGB-LDPFNet[53]'ICPRW-24	VideoMAE/ViT-B	0.9740	0.7724	0.9996	0.9688	0.9914	0.9364	0.9404	0.8055	0.7550	0.7634	0.9904	0.9077	0.8444
RGB-LDPFNet[53]'ICPRW-24	VideoMAE/ViT-L	0.9815	0.8283	0.9996	0.9828	0.9940	0.9496	0.9559	0.8727	0.8477	0.7955	0.9902	0.9409	0.8894
RGB-LDP Fusion-SSAT	VideoMAE/ViT-B	0.9491	0.7796	0.9941	0.9451	0.9866	0.8710	0.9209	0.7170	0.7659	0.7551	0.9807	0.9311	0.8299
RGB-LDP Fusion-SSAT	VideoMAE/ViT-L	0.9865	**0.8349**	0.9996	**0.9921**	0.9978	0.9571	<u>0.9613</u>	0.8740	0.8256	0.8122	**0.9962**	**0.9570**	<u>0.8930</u>

score compared to existing state-of-the-art approaches, indicating that our model, Fusion-SSAT, exhibits superior generalizability to on unseen deepfake manipulation techniques. The ROCs are in Fig. 3.

4.3.2 Experimental results on Cross-Forgery and Domain

Table 2 summaries the experimental results of our proposed model trained on the DF40 ff domain, which includes FS(FF), FR(FF), and EFS(FF), and evaluated on within-forgery, cross-forgery, and cross-domain data. Column-1 specifies the forgery technique utilised for model training, column-2 defines the model employed for evaluation, and column-3 presents the experimental results of the DF40 test set (FF and CDF domains).

Cross-Forgery Evaluation: Our proposed model, Fusion-SSAT, outperformed all existing models with an average AUC of 0.970% on FS (FF), 0.958% on FR (FF), and 0.999% on EFS (FF). In direct comparisons, our model shows a significant upward trend in AUC: an improvement of 10% when trained on FF (FF), 12% when trained on FR (FF), and up to 30% when trained on EFS (FF).

While all state-of-the-art models exhibit poor performance on the EFS test data, our models demonstrate better generalizability on both FS and FR test sets. When trained on the EFS (FF) training set, SOTA models show a performance drop of up to 50% across all test scenarios. In contrast, our proposed method consistently outperforms them, achieving an AUC of 0.99% on FS (FF) and FR (FF) test sets. The ROCs for these experiments can be found in Fig. 4.

Cross-Domain evaluation: Most of the current SOTA deepfake outperforms when trained and tested on FF domain witnessed in Table 2, but exhibit poor performance when it is tested on cross-domain data (CDF) as shown in Table 2.

Our proposed model, Fusion-SSAT outperformed all existing SOTA models listed by average AUC of 0.986% on FS(CDF), 0.981% on FR(CDF), 0.991% on EFS(CDF) CDF Test set. Even in one to one cross-domain test, all SOTA models' performance is only up to 70% AUC, while our proposed models' AUC is not less than 89%, which shows a significant and better generalizability capability when compare to the existing deepfake detectors. In the FR (CDF) test, the majority of current methodologies exhibit

TABLE 2: ***Cross-forgery and Cross-domain evaluation:*** *Models trained on DF40 FF domain with different forgery methods (FS, FR, EFS). Evaluation is performed on both FF and CDF domains.*

Training Set	Model	Testing Set (FF)				Testing Set (CDF)			
		FS	FR	EFS	Avg. (FF)	FS	FR	EFS	Avg. (CDF)
FS (FF)	Xception[14]'CVPR-17	0.991	0.892	0.810	0.898	0.922	0.657	0.642	*0.740*
	CLIP[51]'ICML-21	0.996	0.908	0.837	*0.914*	0.967	0.744	0.730	0.814
	SRM[43]'CVPR-21	0.988	0.867	0.703	0.853	0.919	0.621	0.603	0.714
	SPSL[40]'CVPR-21	0.987	0.849	0.735	0.857	0.938	0.656	0.648	*0.747*
	RECCE[6]'CVPR-20	0.991	0.855	0.758	0.868	0.926	0.632	0.610	0.723
	RFM[66]'CVPR-21	0.992	0.884	0.821	*0.899*	0.939	0.637	0.628	0.735
	SSAT [18]'WACV-24	0.988	0.963	0.972	0.974	0.985	0.860	0.990	0.945
	LDP-RGBFNet ViT-B[53]'ICPRW-24	0.458	0.354	0.322	0.378	0.577	0.432	0.711	0.573
	LDP-RGBFNet ViT-L[53]'ICPRW-24	0.991	0.950	0.954	0.965	0.994	0.882	0.992	0.956
	LDP-LDPFNet ViT-B[53]'ICPRW-24	0.991	0.978	0.899	0.956	0.990	0.895	0.968	0.951
	LDP-LDPFNet ViT-L[53]'ICPRW-24	0.993	0.970	0.932	0.965	0.995	0.916	0.988	0.966
	RGB-LDPFNet ViT-B[53]'ICPRW-24	0.994	0.980	0.870	0.948	0.992	0.913	0.976	0.960
	RGB-LDPFNet ViT-L[53]'ICPRW-24	0.994	0.981	0.905	0.960	0.998	0.951	0.979	<u>0.976</u>
	RGB-LDP Fusion-SSAT ViT-B	0.978	0.982	0.950	<u>0.970</u>	0.995	0.949	0.989	0.977
	RGB-LDP Fusion-SSAT ViT-L	**0.996**	**0.993**	0.920	<u>0.970</u>	**0.998**	**0.976**	0.984	<u>0.986</u>
FR (FF)	Xception[14]'CVPR-17	0.838	0.996	0.670	0.835	0.481	0.857	0.369	0.569
	CLIP[51]'ICML-21	0.932	0.999	0.798	*0.910*	0.638	0.933	0.209	*0.593*
	SRM[43]'CVPR-21	0.893	0.998	0.698	0.863	0.454	0.869	0.326	0.550
	SPSL[40]'CVPR-21	0.901	0.998	0.695	0.865	0.479	0.852	0.256	0.529
	RECCE[6]'CVPR-20	0.865	0.997	0.716	0.859	0.452	0.881	0.332	0.555
	RFM[66]'CVPR-21	0.892	0.999	0.776	*0.889*	0.492	0.882	0.359	*0.578*
	SSAT [18]'WACV-24	0.991	0.997	0.937	0.975	0.968	0.982	0.955	0.968
	LDP-RGBFNet ViT-B[53]'ICPRW-24	0.379	0.661	0.017	0.352	0.348	0.550	0.096	0.331
	LDP-RGBFNet ViT-L[53]'ICPRW-24	**0.996**	**0.996**	0.955	<u>0.982</u>	**0.994**	0.994	0.980	<u>0.989</u>
	LDP-LDPFNet ViT-B[53]'ICPRW-24	0.980	0.993	0.919	0.964	0.970	0.982	0.932	0.961
	LDP-LDPFNet ViT-L[53]'ICPRW-24	0.988	0.993	0.924	<u>0.968</u>	0.991	0.992	0.966	<u>0.983</u>
	RGB-LDPFNet ViT-B[53]'ICPRW-24	0.972	0.987	0.914	0.958	0.972	0.979	0.962	0.971
	RGB-LDPFNet ViT-L[53]'ICPRW-24	0.975	0.985	0.894	0.951	0.981	0.980	0.949	0.970
	RGB-LDP Fusion-SSAT ViT-B	0.936	0.954	0.894	0.928	0.964	0.973	0.947	0.961
	RGB-LDP Fusion-SSAT ViT-L	0.984	0.994	0.890	0.956	0.990	**0.995**	0.959	0.981
EFS (FF)	Xception[14]'CVPR-17	0.665	0.807	0.999	0.824	0.586	0.594	0.983	0.721
	CLIP[51]'ICML-21	0.688	0.889	0.999	*0.859*	0.617	0.735	0.988	*0.780*
	SRM[43]'CVPR-21	0.596	0.776	0.999	0.790	0.589	0.620	0.964	0.724
	SPSL[40]'CVPR-21	0.659	0.811	0.999	0.823	0.635	0.651	0.975	0.754
	RECCE[6]'CVPR-20	0.691	0.801	0.999	*0.830*	0.623	0.603	0.984	0.737
	RFM[66]'CVPR-21	0.653	0.795	0.999	0.816	0.644	0.666	0.981	*0.764*
	SSAT [18]'WACV-24	0.997	0.997	1.0	0.998	0.995	0.959	1.0	0.985
	LDP-RGBFNet ViT-B[53]'ICPRW-24	0.924	0.837	1.0	0.920	0.823	0.719	1.0	0.847
	LDP-RGBFNet ViT-L[53]'ICPRW-24	0.999	0.998	1.0	0.999	0.997	0.981	1.0	0.993
	LDP-LDPFNet ViT-B[53]'ICPRW-24	0.999	0.997	1.0	0.998	0.997	0.979	1.0	<u>0.992</u>
	LDP-LDPFNet ViT-L[53]'ICPRW-24	0.999	0.998	1.0	0.999	0.997	0.981	**1.0**	0.993
	RGB-LDPFNet ViT-B[53]'ICPRW-24	0.999	0.996	0.999	<u>0.998</u>	0.996	0.965	0.990	0.983
	RGB-LDPFNet ViT-L[53]'ICPRW-24	0.999	0.998	0.999	<u>0.998</u>	0.995	0.980	0.990	0.988
	RGB-LDP Fusion-SSAT ViT-B	0.996	0.964	0.999	0.986	0.971	0.840	0.990	0.933
	RGB-LDP Fusion-SSAT ViT-L	**0.999**	**0.999**	0.999	<u>0.999</u>	**0.997**	**0.986**	0.990	<u>0.991</u>
All	MARALIN [5]'CVPR-23	-	-	-	0.981	-	-	-	0.796
	GrDT [72]'WACVW-25	-	-	-	0.986	-	-	-	0.824
	Ours	-	-	-	0.999	-	-	-	0.991

diminished performance, with average AUCs falling below 0.60%. On the other hand, our models attain significantly higher average AUCs of 0.97% and 0.98%. In all CDF domain tests where current methods demonstrate poor performance, particularly in the FR (FF) scenario where average AUCs frequently fall below 0.60, our models demonstrate exceptional reliability, with AUCs consistently above 0.940 across all test sets. The ROCs for these experiments can be found in Fig. 5.

Fusion-SSAT consistently outperforms all other models across all testing scenarios, achieving the highest average AUC scores. This demonstrates its remarkable generalisation capability across diverse types of forgeries and areas. These findings confirm the effectiveness of combining texture-based features with global features to improve cross-domain deepfake detection.

4.4 Ablation Study

As part of the ablation study, we evaluate the impact of different feature fusion strategies, including LBP-LBP, LDP-LDP, and LDP-RGB. The results, summarised in Table 3, 4 reveal that relying solely on local texture features (LBP-LBP, LDP-LDP) yields suboptimal performance. Notably, combining local descriptors with global representations (LDP-RGB) leads to more discriminative features and improved overall performance.

5 CONCLUSION

In this work we introduced a novel approach Fusion-SSAT which fuses the feature representation from self-supervised auxiliary task with the primary task. We utilized masked LDP local descriptor as input to the SSAT task and try to reconstruct the RGB patches which makes SSL task harder to learn the local patterns rather than

TABLE 3: *In-domain and cross-domain evaluation with fusion of LDP-LDP and LDP-RGB features trained on FF++[57] dataset*

Detector	Backbone	In-Domain Evaluation							Cross Domain Evaluation					
		FF-c23	FF-c40	FF-DF	FF-F2F	FF-FS	FF-NT	Avg.	CDFv1	CDFv2	DFD	Fsh	UADFV	Avg.
LDP-RGB Fusion-SSAT	VideoMAE/ViT-B	0.8833	0.7694	0.9899	0.8163	0.9467	0.7814	0.8645	0.5963	0.7071	0.6500	0.9274	0.9115	0.7585
LDP-RGB Fusion-SSAT	VideoMAE/ViT-L	**0.9866**	0.7731	0.9997	0.9887	0.9973	**0.9605**	<u>0.9510</u>	**0.8748**	**0.8254**	**0.8557**	**0.9947**	**0.9540**	<u>0.9009</u>
LDP-LDP Fusion-SSAT	VideoMAE/ViT-B	0.7109	0.6696	0.9242	0.5941	0.7051	0.6209	0.7041	0.5309	0.754	0.6157	0.7482	0.8426	0.6983
LDP-LDP Fusion-SSAT	VideoMAE/ViT-L	0.9846	**0.7786**	**0.9999**	**0.9849**	**0.9966**	0.9581	0.9505	0.8382	0.8158	0.8058	0.9937	0.9281	0.8763
LBP-RGB Fusion-SSAT	VideoMAE/ViT-B	0.8243	**0.7023**	0.9689	0.8262	**0.9052**	0.7325	0.8266	0.5262	0.7001	0.6234	**0.9532**	0.9013	0.7408
LBP-RGB Fusion-SSAT	VideoMAE/ViT-L	**0.9123**	0.6512	**0.9781**	**0.9345**	0.8469	0.7934	0.8527	**0.7876**	**0.8011**	**0.7350**	0.8692	**0.9021**	<u>0.8190</u>
LBP-LBP Fusion-SSAT	VideoMAE/ViT-B	0.5311	0.6087	0.7420	0.5803	0.6234	0.5698	0.6092	0.4882	0.5405	0.6120	0.5931	0.6598	0.5787
LBP-LBP Fusion-SSAT	VideoMAE/ViT-L	0.8243	0.6625	0.9711	0.8442	0.9107	**0.7981**	<u>0.8351</u>	0.7821	0.7940	0.7011	0.8250	0.8792	0.7963

TABLE 4: *Cross-forgery and Cross-domain evaluation with fusion of LDP-LDP and LDP-RGB features trained on DF40 dataset*

Training Set	Model	Testing Set (FF)				Testing Set (CDF)			
		FS	FR	EFS	Avg. (FF)	FS	FR	EFS	Avg. (CDF)
FS (FF)	LDP-RGB Fusion-SSAT ViT-B	0.948	0.906	**0.955**	0.936	0.940	0.826	**0.995**	0.920
	LDP-RGB Fusion-SSAT ViT-L	0.989	0.980	0.932	0.967	0.995	0.962	0.988	0.982
	LDP-LDP Fusion-SSAT ViT-B	0.687	0.620	0.721	0.676	0.774	0.654	0.918	0.782
	LDP-LDP Fusion-SSAT ViT-L	**0.991**	**0.985**	0.963	<u>0.979</u>	**0.997**	**0.966**	0.991	<u>0.985</u>
FR (FF)	LDP-RGB Fusion-SSAT ViT-B	0.834	0.895	0.799	0.842	0.865	0.896	0.929	0.897
	LDP-RGB Fusion-SSAT ViT-L	0.983	0.990	0.917	0.963	**0.991**	**0.994**	0.960	0.982
	LDP-LDP Fusion-SSAT ViT-B	0.609	0.760	0.70	0.689	0.654	0.726	0.827	0.736
	LDP-LDP Fusion-SSAT ViT-L	**0.992**	**0.995**	**0.985**	<u>0.990</u>	0.990	0.992	**0.984**	<u>0.989</u>
EFS (FF)	LDP-RGB Fusion-SSAT ViT-B	0.940	0.836	1.0	0.925	0.767	0.600	1.0	0.789
	LDP-RGB Fusion-SSAT ViT-L	**0.999**	**0.998**	**1.0**	0.999	**0.997**	**0.981**	**1.0**	<u>0.993</u>
	LDP-LDP Fusion-SSAT ViT-B	0.929	0.786	1.0	0.905	0.69	0.496	0.999	0.728
	LDP-LDP Fusion-SSAT ViT-L	0.999	0.998	**1.0**	<u>0.999</u>	0.997	0.976	0.999	0.991

relaying on superficial features. This increased the performance of the primary task to better generalize on unseen manipulations or cross-dataset evaluation. Our experimental results showed a better generalization on large scale DF40 dataset which includes latest deepfake generation techniques than state-of-the-art deepfake detectors.

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security*, pages 1–7. IEEE, 2018.

[2] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu. Partial fc: Training 10 million identities on a single machine. In *ICCVW*, 2021.

[3] P. Balaji, A. Das, S. Das, and A. Dantcheva. Attending generalizability in course of deep fake detection by exploring multi-task learning. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 475–484, 2023.

[4] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos. Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[5] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. Marlin: Masked autoencoder for facial video representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1504, 2023.

[6] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4122, June 2022.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[8] I. Chalkidis, X. Dai, M. Fergadiotis, P. Malakasiotis, and D. Elliott. An exploration of hierarchical attention transformers for efficient long document classification, 2022.

[9] R. Chen, X. Chen, B. Ni, and Y. Ge. Simswap: An efficient framework for high fidelity face swapping. In *MM '20: The 28th ACM International Conference on Multimedia*, 2020.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[11] J. Cheng, Z. Yan, Y. Zhang, Y. Luo, Z. Wang, and C. Li. Can we leave deepfake data behind in training deepfake detector? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[12] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi. Exploiting style latent flows for generalizing deepfake video detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1133–1143, 2024.

[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[14] F. Chollet. Xception: Deep learning with depthwise separable convolutions, 2017.

[15] Y. Dalva, S. F. Altındiş, and A. Dundar. Vecgan: Image-to-image translation with interpretable latent directions. In *European Conference on Computer Vision*, pages 153–169. Springer, 2022.

[16] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5780–5789, 06 2020.

[17] A. Das, S. Das, and A. Dantcheva. Demystifying attention mechanisms for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7, 2021.

[18] S. Das, T. Jain, D. Reilly, P. Balaji, S. Karmakar, S. Marjit, X. Li, A. Das, and M. Ryoo. Limited data, unlimited potential: A study on vits augmented by masked autoencoders. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[19] DeepSeek-AI. Deepseek-v3 technical report, 2025.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[21] C. Feng, Z. Chen, and A. Owens. Self-supervised video forensics by audio-visual anomaly detection. In *2023 IEEE/CVF Conference on*

(a) FS (CDF)

(b) FR (CDF)

(c) EFS (CDF)

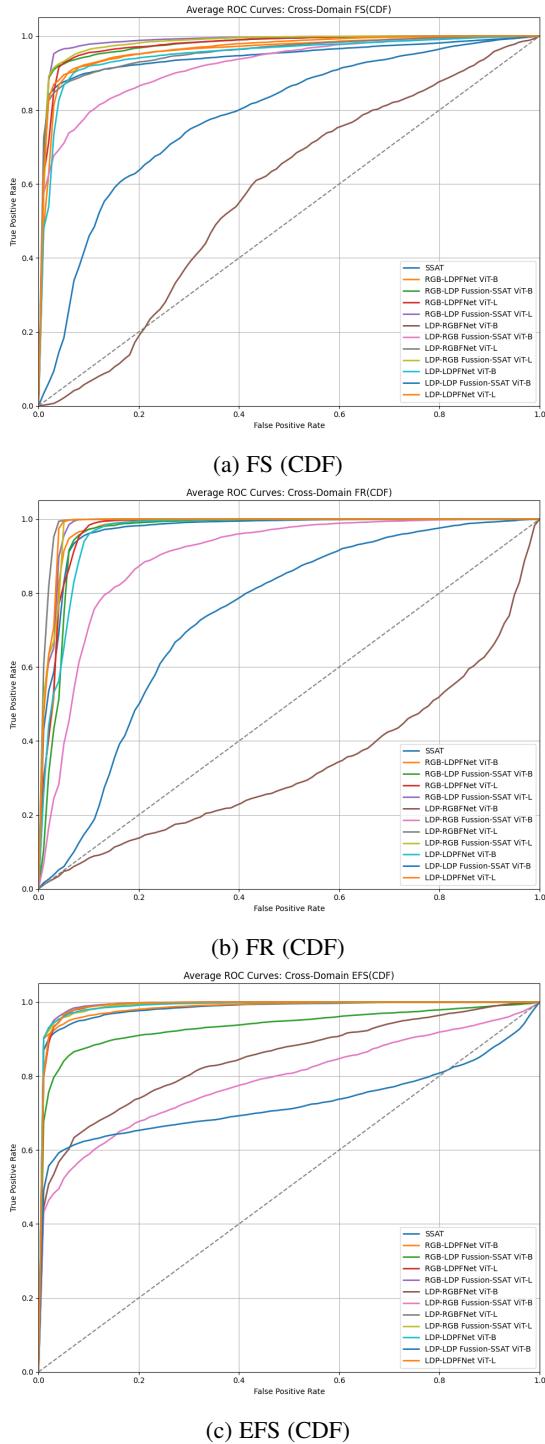Fig. 5: ROCs of cross-domain evaluations on the DF40 dataset.

Computer Vision and Pattern Recognition (CVPR), pages 10491–10503, 2023.

[22] T. Han, W. Xie, and A. Zisserman. Self-supervised co-training for video representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[26] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[27] F.-T. Hong, L. Zhang, L. Shen, and D. Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022.

[28] G.-S. Hsu, C.-H. Tsai, and H.-Y. Wu. Dual-generator face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 642–650, June 2022.

[29] T. Jabid, M. H. Kabir, and O. Chae. Local directional pattern (ldp) - a robust image descriptor for object recognition. In *Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, AVSS '10, page 482–487, USA, 2010. IEEE Computer Society.

[30] A. C. K, A. A V, S. Das, and A. Das. Latent flow diffusion for deepfake video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3781–3790, June 2024.

[31] M. Koujan, M. Doukas, A. Roussos, and S. Zafeiriou. Head2head: Video-based neural head synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 319–326, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

[32] K. Kuckreja, X. Hoque, N. Poddar, S. Reddy, A. Dhall, and A. Das. Indiface: Illuminating india's deepfake landscape with a comprehensive synthetic dataset. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9, 2024.

[33] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020.

[34] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020.

[35] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[36] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020.

[37] J. Liang, S. Liang, A. Liu, X. Jia, J. Kuang, and X. Cao. Poisoned forgery face: Towards backdoor attacks on face forgery detection. In *The Twelfth International Conference on Learning Representations*, 2024.

[38] L. Lin, X. He, Y. Ju, X. Wang, F. Ding, and S. Hu. Preserving fairness generalization in deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16815–16825, 2024.

[39] D. Liu, L. Y. Wu, B. Li, Y. Zhao, Z. Ge, and J. Zhang. T-person-gan: Text-to-person image generation with identity-consistency and manifold mix-up. *Expert Systems with Applications*, 245:123456, 2024. Available online.

[40] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and 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)*, pages 772–781, 2021.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[42] Y. Lu and T. Ebrahimi. Towards the detection of ai-synthesized human face images. *arXiv preprint arXiv:2402.08750*, 2024.

[43] Y. Luo, Y. Zhang, J. Yan, and W. Liu. Generalizing face forgery detection with high-frequency features. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16321, 2021.

[44] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection, 2024.

[45] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos, 2018.

[46] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 12–21, June 2022.

[47] OpenAI. Gpt-4 technical report, 2024.

[48] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021.

[49] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery.

[50] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference Computer Vision*, pages 86–103. Springer, 2020.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[52] K. Ranasinghe, M. Naseer, S. Khan, F. Khan, and M. Ryoo. Self-supervised video transformer. In *CVPR*, pages 2864–2874, 06 2022.

[53] S. Reddy, N. Poddar, S. Das, and A. Das. Self-supervised auxiliary learning for texture and model-based hybrid robust and fair featuring in face analysis. In S. Palaiahnakote, S. Schuckers, J.-M. Ogier, P. Bhattacharya, U. Pal, and S. Bhattacharya, editors, *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 386–401, Cham, 2025. Springer Nature Switzerland.

[54] X. Ren, A. Lattas, B. Gecer, J. Deng, C. Ma, and X. Yang. Facial geometric detail recovery via implicit representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023.

[55] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering, 2021.

[56] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[57] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[58] R. Roy, I. Joshi, A. Das, and A. Dantcheva. *3D CNN Architectures and Attention Mechanisms for Deepfake Detection*, pages 213–234. Springer International Publishing, Cham, 2022.

[59] Z. Sha, Y. Tan, M. Li, M. Backes, and Y. Zhang. Zerofake: Zero-shot detection of fake images generated and edited by text-to-image generation models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*, Salt Lake City, UT, USA, October 2024. ACM.

[60] R. Shao, T. Wu, L. Nie, and Z. Liu. Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision (IJCV)*, 2024.

[61] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[62] K. Shiohara, X. Yang, and T. Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[63] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[64] Z. Tong, Y. Song, J. Wang, and L. Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

[65] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

[66] C. Wang and W. Deng. Representative forgery mining for fake face detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14918–14927, 2021.

[67] H. Wang. Comparative analysis of gans and diffusion models in image generation. *AMCCE 2024, Transactions on Computer Science and Intelligent Systems Research*, 120:59–68, 2024.

[68] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition, 2020.

[69] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.

[70] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022.

[71] Z. Wang, Z. Chi, Y. Zhang, et al. Fregan: Exploiting frequency components for training gans under limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 33387–33399, 2022. Official results cited in Yang et al. (2024).

[72] H. Xie, H. He, B. Fu, and V. Sanchez. Grdt: Towards robust deepfake detection using geometric representation distribution and texture. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 686–696, 2025.

[73] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, and Y. Liu. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7632–7641, June 2022.

[74] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, L. Yuan, C. Wang, S. Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.

[75] Z. Yan, Y. Zhang, Y. Fan, and B. Wu. Ucf: Uncovering common features for generalizable deepfake detection, 2023.

[76] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 4534–4565. Curran Associates, Inc., 2023.

[77] Q. Yang, Z. Zhao, Y. Pu, S. Pan, J. Gu, and D. Xu. Fncontra: Frequency-domain negative sample mining in contrastive learning for limited-data image generation. *Expert Systems with Applications*, 245:123456, 2024. Available online.

[78] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses, 2018.

[79] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[80] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8652–8661, 2023.

[81] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You. Open-sora: Democratizing efficient video production for all, 2024.

[82] M. Zou, B. Yu, Y. Zhan, S. Lyu, and K. Ma. Semantics-oriented multitask learning for deepfake detection: A joint embedding approach, 2024.

Shukesh Reddy is currently a PhD scholar in the Department of Computer Science and Information Systems at BITS Pilani, Hyderabad Campus. He earned his Bachelor's degree in Computer Science in 2023. Before pursuing his PhD, he worked as a Software Engineer, leading the development of an IoT platform and IoT development kit. His research interests focus on learning representations of human faces, with his current work centred on advancing face forgery detection techniques.

Srijan Das I am an Assistant Professor in the Department of Computer Science at the University of North Carolina at Charlotte. At UNC Charlotte, I am working on Video Representation Learning, and Robotic Vision. I am a member of the AI4Health Center and one of the founding members of the Charlotte Machine Learning Lab (CharMLab) at UNC Charlotte.

Abhijit Das is an assistant professor at BITS Pilani Hyderabad. Previously, he worked as a Post-Doc Researcher at Inria Sophia Antipolis–Méditerranée, France. He has completed his PhD from the School of Information and Communication Technology, Griffith University, Australia. He is an accomplished machine learning and computer vision researcher with more than 15 years of research and teaching experience. He is presently pursuing an investigation on learning representations and human analysis employing facial and corporeal-based visual features.