






DeepSpace: Navigating the Frontier of Deepfake Identification Using Attention-Driven Xception and a Task-Specific Subspace

Ayush Roy¹^a, Sk Mohiuddin²^b, Maxim Minenko³^c, Dmitrii Kaplun^{4,3,*}^d and Ram Sarkar⁵^e

¹Jadavpur University, Department of Electrical Engineering, Kolkata, 700032, India

²Asutosh College, Department of Computer Science, Kolkata, 700026, India

³St. Petersburg Electrotechnical University "LETI", Dep. of Automation and Control Processes, St. Petersburg, 197022, Russia

⁴China University of Mining and Technology, Artificial Intelligence Research Institute, Xuzhou, 221116, China

⁵Jadavpur University, Department of Computer Science and Engineering, Kolkata, 700032, India
{aroy80321, myselfmohiuddin}@gmail.com, {mvminenko, dikaplun}@etu.ru, ramjucse@gmail.com

Keywords: Deepfake Detection, Subspace Optimization, Attention Mechanism.


Abstract: The recent advancements in deepfake technology pose significant challenges in detecting manipulated media content and preventing its malicious use in different areas. Using ConvNets feature spaces and fine-tuning them for deepfake classification can lead to unwanted modifications and artifacts in the feature space. To address this, we propose a model that uses Xception as the backbone and a Spatial Attention Module (SAM) to leverage spatial information using shallower features like texture, color, and shape, as well as deeper fine-grained features. We also create a task-specific subspace for projecting spatially enriched features, which boosts the overall model performance. To do this, we utilize Gram-Smith orthogonalization on the flattened features of real and fake images to produce the basis vectors for our subspace. We evaluate the proposed method using two widely used and standard deepfake video datasets: FaceForensics++ and Celeb-DF (V2). We conduct experiments following two different setups: intra-dataset (trained and tested on the same dataset) and inter-dataset (trained and tested on separate datasets). The performance of the proposed model is comparable to that of state-of-the-art methods, confirming its robustness and generalization ability. The code is made available at <https://github.com/AyushRoy2001/DeepSpace>.


1 INTRODUCTION


The proliferation of deepfake technology, driven by sophisticated machine learning algorithms, has ushered in a new era of digital manipulation, where individuals can be convincingly portrayed saying or doing things they never said or did. Face forgery, a subset of deepfakes, involves the seamless alteration of facial features in videos or images, presenting a formidable challenge to the authenticity of visual content. As this technology evolves, the need for robust detection mechanisms (Mohiuddin et al., 2023a) becomes increasingly a necessity to safeguard against poten-


tial misuse, misinformation, and the erosion of trust. Hence, improving deepfake detection can reduce misinformation and protect individuals from harm, enhancing trust in digital content. However, it raises ethical concerns, such as potential misuse for censorship or unjust accusations, requiring careful regulation.


Deepfakes, created with machine learning like Generative Adversarial Networks (GANs), produce convincing synthetic media and raise misinformation concerns. Detecting them is essential, with AI-based Counter-GANs distinguishing real from synthetic content. Deep Convolutional Neural Networks (CNNs), known for their ability to extract valuable features from images, are effective in computer vision and have led to various models for detecting deepfakes. For instance, the SiamNet model (Kingra et al., 2023) leverages inconsistencies in source camera noise patterns to identify artifacts in manipulated videos. New methods (Raza and Malik, 2023; Lewis et al., 2020) combine audiovisual learning to improve

^a <https://orcid.org/0000-0002-9330-6839>

^b <https://orcid.org/0000-0001-6411-4072>

^c <https://orcid.org/0000-0002-7334-7668>

^d <https://orcid.org/0000-0003-2765-4509>

^e <https://orcid.org/0000-0001-8813-4086>

*Corresponding author

deepfake detection, and the Xception network with depth separable convolution layers has shown state-of-the-art performance in this area. This is particularly evident when employing a single CNN model, as demonstrated by Li et al. (Li et al., 2020).

While deep CNN models excel at efficiently identifying local artifacts, contemporary deepfake generation techniques can produce a spectrum of artifacts, ranging from localized distortions to those that encompass the entire image. In addition, the diverse generation techniques contribute to a significant diversity in the types of artifacts generated. Existing deep learning methods occasionally struggle to effectively address this diversity, as evident in Li et al. (Li et al., 2020). To address the extensive diversity observed in the generation of fake content, numerous researchers, exemplified by Yu et al. (Yu et al., 2022), have integrated videos with distinctively manipulated faces into their assessment procedures. Nevertheless, this approach might not be sustainable given the broad spectrum of faking methods. Consequently, there is a pressing need to devise methods that can adeptly handle the vast array of fraudulent artifacts.

CNNs with spatial attention focus on crucial regions in an input image, improving feature recognition, performance, and interoperability. This prioritization aids robust feature extraction and enhances the model's decision-making. Some methods, reported in (Mohiuddin et al., 2023b; Naskar et al., 2024), strive to extract optimal features and strategically reduce computations in deepfake detection. However, unnecessary modifications in the feature subspace during deep learning training can alter features and affect performance. Robust models in other domains have been achieved through proper feature subspace generation and optimization (Yin et al., 2023; Li et al., 2023). A task-specific subspace aligns features in an explainable manner while enhancing model robustness. Limited research has explored the direction of a task-specific subspace for deepfake detection, despite its potential to address challenges in this field. Therefore, the proposed model is designed to handle the deepfake classification task.

The major **contributions** of our work are:

- We propose a methodology that incorporates the projection of the spatially enriched feature from Xception onto a task-specific subspace.
- A pivotal component of our model is the Spatial Attention Module (SAM). SAM amalgamates the shape, color, texture, etc. information from the shallow features, and the fine-grained information from the deep features to spatially enrich the output features of the Xception model.
- The lack of exploration of task-specific subspace for deepfake classification encourages us to formulate one. We have utilized features of 'fake' and 'real' images to create the orthonormal basis vectors of our task-aware subspace using the Gram-Smith orthogonalization. The features are projected onto this subspace for an interpretable boost in the model's performance.
- Extensive experiments conducted on two challenging and popular deepfake datasets: FaceForensics++ (Rossler et al., 2019) and CelebDF (Li et al., 2020), demonstrate that our model achieves state-of-the-art (SOTA) results in terms of both efficiency and resilience against known attacks. Our approach substantially reduces false positives, enhancing the accuracy and reliability of the proposed model.

The paper is structured as follows: Section 2 provides an overview of past methods employed for detecting deepfakes. The working principle of the proposed model is detailed in Section 3. In 4, we assess our method in extracting global inconsistencies, presenting related datasets and the experimental protocol. Finally, 5 reports conclusive remarks on our work.

2 RELATED WORK

In recent years, researchers have proposed deep learning-based methods for detecting deepfakes, aiming to enhance the robustness and accuracy of identifying manipulated media content. The advanced techniques, such as those utilizing GANs, challenge current detection methods, making typical identification methods less effective.

Earlier approaches primarily relied on deep learning architectures, especially CNNs, for detecting deepfakes. For instance, pre-trained Xception and Capsule Network models were used in Tolosana et al. (Tolosana et al., 2021) to analyze full-face and specific facial components. Studies such as those by Rossler et al. (Rossler et al., 2019) demonstrated the superior performance of XceptionNet across different datasets. Meanwhile, Afchar et al. (Afchar et al., 2018) utilized mesoscopic details with their Meso-4 and MesoInception-4 CNN architectures for forgery detection. Amerini et al. (Amerini et al., 2019) explored unusual facial motion, using PWC-Net and VGG16-based models to filter out authentic videos. Despite their success, these deep CNN models face limitations in capturing both local and global features simultaneously, leading to challenges in accurately identifying manipulation artifacts.

In addition to the common challenges associated with deep learning models, various studies (Ganguly et al., 2022; Lu et al., 2023) propose incorporating attention mechanisms into CNN models to broaden the focus on facial image regions. For example, Nguyen et al. (Nguyen et al., 2024) proposed an explicit attention mechanism in a multi-task learning framework (LAA-Net). By combining heatmap-based and self-consistency attention, it focuses on artifact-prone regions. Then, an Enhanced Feature Pyramid Network (E-FPN) efficiently spreads low-level features, limiting redundancy used for detection of deepfake. Xia et al. (Xia et al., 2024) introduced the Multi-Collaboration and Multi-Supervision Network (MM-Net), addressing diverse spatial scales and sequential permutations in manipulated face images, and achieving recovery without requiring knowledge of the specific manipulation method. Furthermore, these attention models primarily utilized convolution operations across the entire image to generate the necessary attention map, potentially emphasizing unimportant regions. However, the presence of irrelevant features in existing models can mislead classifiers during tasks, leading to longer training times. Utilizing an effective feature selection method may be helpful to overcome this limitation by eliminating non-important features.

Many recent approaches emphasize ensemble techniques, feature selection, and feature engineering. Some methods (Zhang et al., 2022; Hooda et al., 2024) prioritize the selection of relevant features through deep learning, while others (Mohiuddin et al., 2023b; Naskar et al., 2024) rely on feature engineering methods for deepfake detection. An alternative approach involves leveraging multiple modalities, such as audio, video, and text, to enhance detection accuracy and robustness against adversarial forgeries. By combining complementary features from these modalities, multimodal methods (Liz-Lopez et al., 2024; Yu et al., 2023; Raza and Malik, 2023) can effectively capture inconsistencies across different information streams, which are often difficult for unimodal approaches to detect. As explored in other domains, task-aware subspace learning offers several advantages, capturing task-specific information to reduce unwanted changes and modifications of the features (Zhou et al., 2023). Enhancing both local and global correlation structures improves data affinity for robust and applicable subspace clustering, preserves global and local data structure, and extracts discriminative features, thus improving classification performance (Kou et al., 2023). Overall, the enhancement in task performances by capturing task-specific information, reducing interference, and improving the robustness of subspace clustering and feature extraction

can be seen in (Srirangarajan et al., 2022). These facts motivate us to apply task-aware subspace learning for deepfake detection.

3 METHODOLOGY

The method being proposed uses the features of Xception to classify deepfake images. For effective deep feature extraction, we utilized the fine-tuned Xception due to its proven success in detecting face manipulation (Li et al., 2020). The features from Xception are spatially enriched by using the SAM and then flattened using the Global Average Pooling layer (GAP), followed by a dense layer consisting of 512 units and Rectified Linear Unit (ReLU) activation. These flattened features are then projected onto a task-specific subspace to provide a better feature representation, thus enhancing the overall performance of the model. Finally, the classification layer utilizes these projected flattened features to classify the deepfake images. A block diagram illustrating the overall architecture of the proposed method is shown in Fig. 1.

3.1 Xception

Xception (Chollet, 2017) is a type of CNN architecture that is well-known for its efficient and effective feature extraction. It is achieved by using depth-wise separable convolutions and residual connections. By using depth-wise separable convolutions, Xception reduces computational complexity while still maintaining expressive power. This is done by applying separate filters to each input channel followed by a point-wise convolution that mixes and transforms the output channels. This architectural choice enables efficient computation by reducing the number of parameters. Additionally, Xception incorporates residual connections that help in the learning of complex features and promote a better training of deep networks. This has been demonstrated in the case of deepfakes (Sahib and AlAsady, 2022).

3.2 Spatial Attention Module

Deepfake images can be identified by inconsistencies in their spatial features such as lighting, texture, shading, and object relationships within the image. Some common indicators of deepfakes are blurry edges, unnatural skin tones, and misplaced shadows. However, these signs may not always be easy to detect. To better detect deepfakes, we use both shallow and deep features of images. The shallow features capture

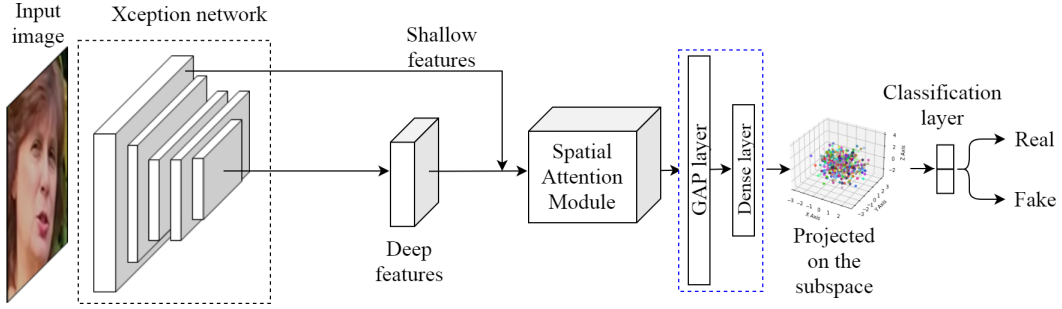


Figure 1: Overall workflow of the proposed deepfake detection method.

shape, edge, texture, and other information, while the deeper features capture the fine-grain relationships of the image. We use F_{enc1} , the output feature from the last layer of Xception, to capture these deep features, and F_{enc2} , the output feature from one of the initial layers of Xception, to capture the shallow features. The dimensions of F_{enc1} and F_{enc2} are $H \times W \times C$ and $H \times W \times C'$, respectively. Using F_{enc2} , we perform average and max pooling across the channel dimension to capture spatial relationships. These features are then concatenated and convoluted, followed by a sigmoid activation function to produce F_{attn} of dimension $H \times W \times 1$. Eq. 1 shows the formation of F_{attn} .

$$F_{attn} = \sigma(f^{1 \times 1}[\text{MaxPool}(F_{enc2}); \text{AvgPool}(F_{enc2})]) \quad (1)$$

Here, MaxPool and AvgPool are average pooling and max pooling across the channel dimension, respectively, $[\]$ denotes concatenation, σ denotes sigmoid activation, and $f^{1 \times 1}$ is a convolution layer with a kernel size of 1×1 . F_{attn} now consists of the spatially enriched information based on texture, shape, etc.

F_{enc1} is processed by a convolution layer, followed by a ReLU activation function. This produces a feature vector with dimensions $H \times W \times C'/2$. To enhance the spatial information of this feature, it is point-wise multiplied with F_{attn} , which already contains valuable spatial information. The resulting tensor, denoted as F_{eattn} , also has dimensions $H \times W \times C'/2$ and is calculated as shown in Eq. 2.

$$F_{eattn} = \text{relu}f^{1 \times 1}(F_{enc1}) \otimes F_{attn} \quad (2)$$

Here, \otimes denotes point-wise multiplication.

To provide attention weights across the channel dimension, a learnable weight α is multiplied across the channel dimension of F_{eattn} to produce F_{SAM} . A block diagram representation of SAM is shown in Fig. 2

3.3 Task-Aware Subspace

Deep learning often requires a fine-tuning of existing feature spaces for classification tasks. However,

this can sometimes cause unnecessary and undesired modifications to feature spaces, leading to a less explainable model that performs poorly. To address this issue, it is important to customize an optimal feature space for a particular classification task, which can result in a more explainable and high-performing model.

To classify deepfakes, we have created a task-specific subspace by collecting flattened features from a trained Xception model (without the SAM) on deepfake datasets. We have used the features from the Dense layer, which consists of 512 units and 'relu' activation (as shown in Fig. 1), and randomly selected features of 512 images (256 'fake' and 256 'real') from all the collected features. With these selected flattened features, we have created a subspace oriented specifically towards classifying deepfakes. To calculate the basis vectors for this subspace, we have employed Gram-Schmidt orthogonalization, which is a method frequently used in linear algebra and signal processing to find an orthogonal basis for a subspace spanned by a set of vectors in an inner product space.

Given a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ (here, $\mathbf{v}_1, \mathbf{v}_2, \dots$ are the flattened features of the selected 512 images that we collected), the Gram-Schmidt orthogonalization process computes an orthogonal set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ as follows:

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{v}_1 \\ \mathbf{u}_2 &= \mathbf{v}_2 - \frac{\langle \mathbf{v}_2, \mathbf{u}_1 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle} \mathbf{u}_1 \\ \mathbf{u}_3 &= \mathbf{v}_3 - \frac{\langle \mathbf{v}_3, \mathbf{u}_1 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle} \mathbf{u}_1 - \frac{\langle \mathbf{v}_3, \mathbf{u}_2 \rangle}{\langle \mathbf{u}_2, \mathbf{u}_2 \rangle} \mathbf{u}_2 \\ &\vdots \\ \mathbf{u}_n &= \mathbf{v}_n - \sum_{k=1}^{n-1} \frac{\langle \mathbf{v}_n, \mathbf{u}_k \rangle}{\langle \mathbf{u}_k, \mathbf{u}_k \rangle} \mathbf{u}_k \end{aligned}$$

For our task, we normalize $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ ($n=512$) to create an orthonormal basis. We then project the flattened features onto this subspace and utilize the projected feature in the classification layer.

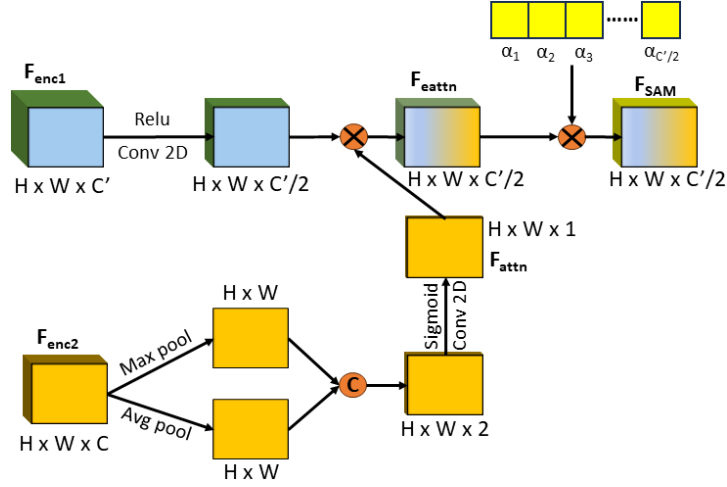


Figure 2: An illustration of the Spatial Attention Module (SAM).

For a detailed understanding, refer to the step-by-step explanation outlined in Algorithm 1.

Algorithm 1: Constructing Task-aware Subspace.

- 1: **Input:** Flattened features $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, where $\mathbf{v}_i \in \mathbb{R}^d$ (collected from Dense layer of Xception model)
 - 2: **Output:** Orthonormal basis $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$
 - 3: Initialize $\mathbf{U} \leftarrow \{\}$ (empty set for orthogonal vectors)
 - 4: **for** $i = 1$ to n **do**
 - 5: Compute $\mathbf{u}_i \leftarrow \mathbf{v}_i$
 - 6: **for** $j = 1$ to $i - 1$ **do**
 - 7: $\mathbf{u}_i \leftarrow \mathbf{u}_i - \frac{\langle \mathbf{v}_i, \mathbf{u}_j \rangle}{\langle \mathbf{u}_j, \mathbf{u}_j \rangle} \mathbf{u}_j$
 - 8: **end for**
 - 9: Normalize $\mathbf{u}_i \leftarrow \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|}$
 - 10: Add \mathbf{u}_i to \mathbf{U}
 - 11: **end for**
 - 12: **Project Features:** For each feature $\mathbf{v}_k \in \mathbf{V}$, project onto the subspace:
 - 13: $\mathbf{p}_k \leftarrow \sum_{j=1}^n \langle \mathbf{v}_k, \mathbf{u}_j \rangle \mathbf{u}_j$
 - 14: Use the projected features $\{\mathbf{p}_k\}$ for classification tasks.
-

4 RESULTS

In this section, we discuss the effectiveness of our proposed method for detecting deepfake videos. We provide a detailed analysis of the results obtained from our experiments. We have tested our method against various challenges and perturbations to assess its ability to maintain effectiveness in adverse conditions. Furthermore, we have evaluated the performance of our method on different datasets to en-

sure its robustness. The performance measure metrics used in our evaluation are the Area Under Curve (AUC) score and accuracy.

4.1 Dataset Description

We have performed experiments to evaluate the effectiveness of our proposed deepfake detection methodology. For this purpose, we have used two commonly used public benchmark datasets: Celeb-DF (V2) (Li et al., 2020), also known as "CeDF", and FaceForensics++ (Rossler et al., 2019), abbreviated as "FF++". **CeDF Dataset.** The dataset contains 5,639 high-quality videos showcasing various celebrities. These videos have been created from 590 original videos that have been collected from YouTube and then filtered to produce variations in age, gender, and background. The dataset includes a total of 518 videos, consisting of 178 real videos and 340 fake ones. It is an ideal resource for evaluating the effectiveness of new deepfake detection techniques.

FF++ Dataset. The dataset is made up of four types of mixed videos: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Each category contains 1,000 videos that have been created using unique source videos. For our experiments, we have focused on the deepfake category, which we have considered as the fake video class. Meanwhile, we have regarded the original category videos as real videos. The FF++ dataset is available in different compression rates, but we have used the c23 version.

Data Preparation for Experiments. To conduct experiments, we have split both datasets into three sets: train, validation, and test. For the training sets, we have selected frames equidistant in the time domain to extract frames from the train video sets. This has

been done to include the maximum variations in the training image dataset. For the test and validation sets, only the first I-frame from each video is considered. This means that the first I-frame in a video decides the video’s authentication. While generating cropped face images using the MTCNN algorithm, we have extracted only one cropped face image with the highest confidence score. Table 1 shows the detailed distribution of video and image levels used in our research.

4.2 Experimental Protocols

We have employed two evaluation setups to assess the performance of our model on the two datasets:

Intra-Dataset Evaluation. In this particular setup, the model has undergone training and testing on the images collected from the same dataset. This means that the same artifacts left by the generation methods of deepfakes are present in both the train and test samples. Essentially, the model learns these artifacts during training, which are present in the test samples as well. We represent the evaluation protocols as FF++_FF++ and CeDF_CeDF when the model is trained and tested on the FF++ dataset and the CeDF dataset, respectively. This evaluation setup allows us to analyze the model’s performance within a single dataset and assess its ability to differentiate between real and fake videos within the same modality.

Inter-Dataset Evaluation. In this particular setup, the model is trained on one dataset category and later tested on the other. For instance, we have first trained the model on the FF++ dataset and then validated and tested it on the CeDF dataset, referred to as FF++_CeDF. Similarly, we have trained the model on the CeDF dataset and then validated and tested it on the FF++ dataset, denoted as CeDF_FF++. This evaluation helps us to understand how well the model performs on different datasets with varying modalities and characteristics, and how robust it is across them.

For all our experiments, we have used a learning rate of 0.001, the Adam optimizer, and a batch size of 32 for all of our experiments. We have applied the cross-entropy loss function for training the model over 50 epochs and evaluated it using standard metrics in TensorFlow on an NVIDIA TESLA P100 GPU. The training and validation curves of the proposed model are shown in Fig. 3.

4.3 Ablation Studies

We have conducted a series of ablation experiments on the CeDF dataset (CeDF_CeDF setup) to determine the optimal architectural configuration and iden-

tify the improvements that each component brings. All of these experiments have been conducted under the same setup, including learning rate and epochs, as mentioned in section 4.1.

In our first ablation experiment, we examine the performance enhancement of a model using SAM. We have used the output of the Xception layer with the dimensions of $19 \times 19 \times 728$ as F_{enc2} for SAM v-1. For SAM v-2, we have used the output of the Xception layer with the dimensions of $10 \times 10 \times 1024$ as F_{enc2} (as seen in Fig. 2). The results, shown in Table 2, indicate that SAM v-2 performs significantly better than SAM v-1, resulting in improved accuracy (+0.58%) and AUC (+2.05%) scores for Xception.

We have conducted additional experiments to demonstrate the significance of a task-specific subspace in projecting flattened features before classification. By utilizing the best architectural configuration from Table 2 and running ablation tests, we have determined that the task-oriented subspace created using orthonormal basis vectors and normalizing the orthogonal vectors using the L2 norm yielded the best results, as shown in Table 3. Additionally, we have visualized the subspace representation of features before and after projection onto the task-aware subspace in Fig. 4. To generate the subspace plots, we have reduced the features from the Global Average Pooling (GAP) and the Dense layer using PCA to three dimensions. The GAP layer features are the features before projection onto the subspace, whereas the Dense features are the projected features onto the subspace. It can be seen in the plots that the organized orientation after projecting onto the subspace helps in classifying the real and fake images.

4.4 SOTA Comparison

We have conducted evaluations of different methods to compare them fairly and meaningfully. To achieve this, we have followed our established experimental protocols as described earlier in subsection 4.2. This approach has allowed us to assess and compare different methods based on important factors such as their detection performance, robustness, and generalization ability. In this regard, we present the performance results of the deepfake detection methods on our experimental setup, specifically on the two image datasets.

The corresponding performance metrics considered are the test accuracy and AUC score. We have reported these results in Table 4, which shows the performance against intra-dataset experiments and Table 5, which exhibits the performance against inter-dataset experiments. By examining the results, it becomes evident that our method outperforms most of

Table 1: The datasets used here exhibit the distribution of classes. The labels “Re” and “Fa” correspond to Real and Fake, respectively.

Dataset	#Video						#image					
	Train		Validation		Test		Train		Validation		Test	
	Re	Fa	Re	Fa	Re	Fa	Re	Fa	Re	Fa	Re	Fa
Celeb-DF	612	4399	100	900	178	340	1130	8022	100	900	178	340
FF++	700	700	200	200	100	100	2930	2946	200	200	100	100

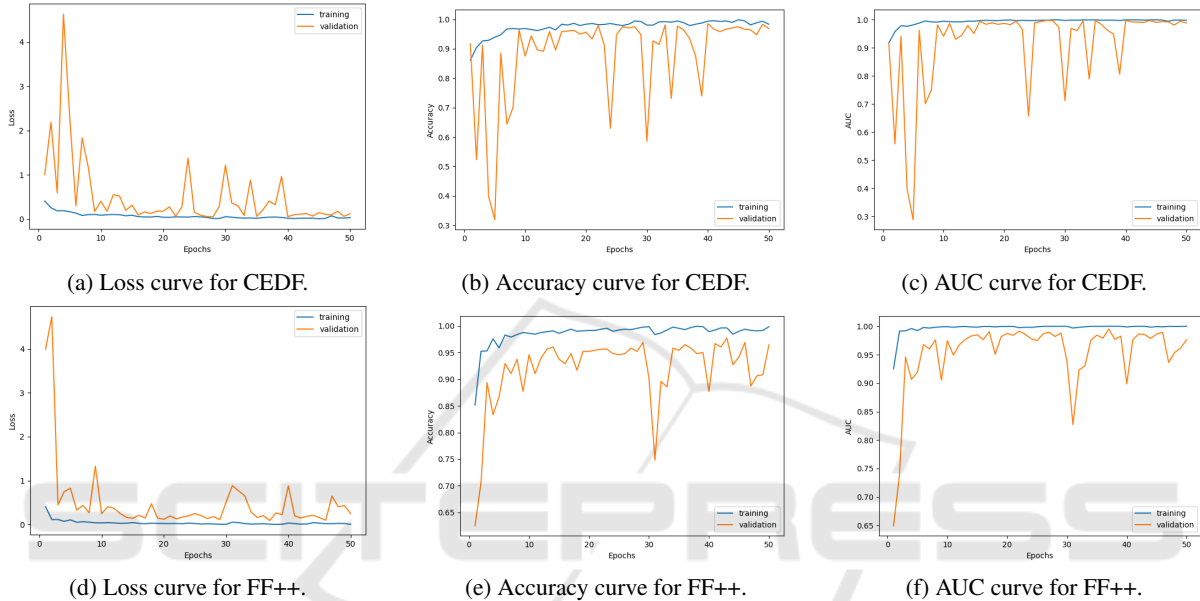


Figure 3: The loss, accuracy, and AUC curves of the proposed model for FF++ (bottom row) and CeDF (top row).

Table 2: Performance of Xception with and without the use of SAM. All values are in %.

Model	Accuracy	AUC Score
Xception	93.63	90.73
Xception + SAM v-1	93.82	92.22
Xception + SAM v-2	94.21	92.78

Table 3: Analysis of the task-aware subspace using various norms for the basis vectors. All values are in %.

Model	Accuracy	AUC Score
Orthogonal	93.63	91.40
Orthonormal (L1 norm)	94.21	91.71
Orthonormal (L2 norm)	95.17	93.25

the SOTA methods used in this comparison, particularly in the context of intra-dataset experimental setups. In other words, the proposed method demonstrates superior performance when tested within the same dataset.

Furthermore, when considering inter-dataset experiments, which essentially evaluate the generaliza-

tion ability of a model, once again the current method outperforms most of the other methods and achieves competent results. This indicates that our method exhibits high adaptability and effectiveness when faced with different datasets. Overall, these findings underscore the impressive performance of the current method compared to other methods, both within and across different datasets, confirming its superiority in deepfake detection.

The confusion matrices obtained by evaluating the proposed model using the intra-dataset and inter-dataset experimental setup are shown in Fig. 5.

5 CONCLUSION

In this paper, we have developed a deep learning-based approach to detect deepfake videos. Our approach initially employs Xception as the backbone to extract deep learning features. We then use SAM to spatially enrich the extracted features by leveraging information from deeper features (fine-grained de-

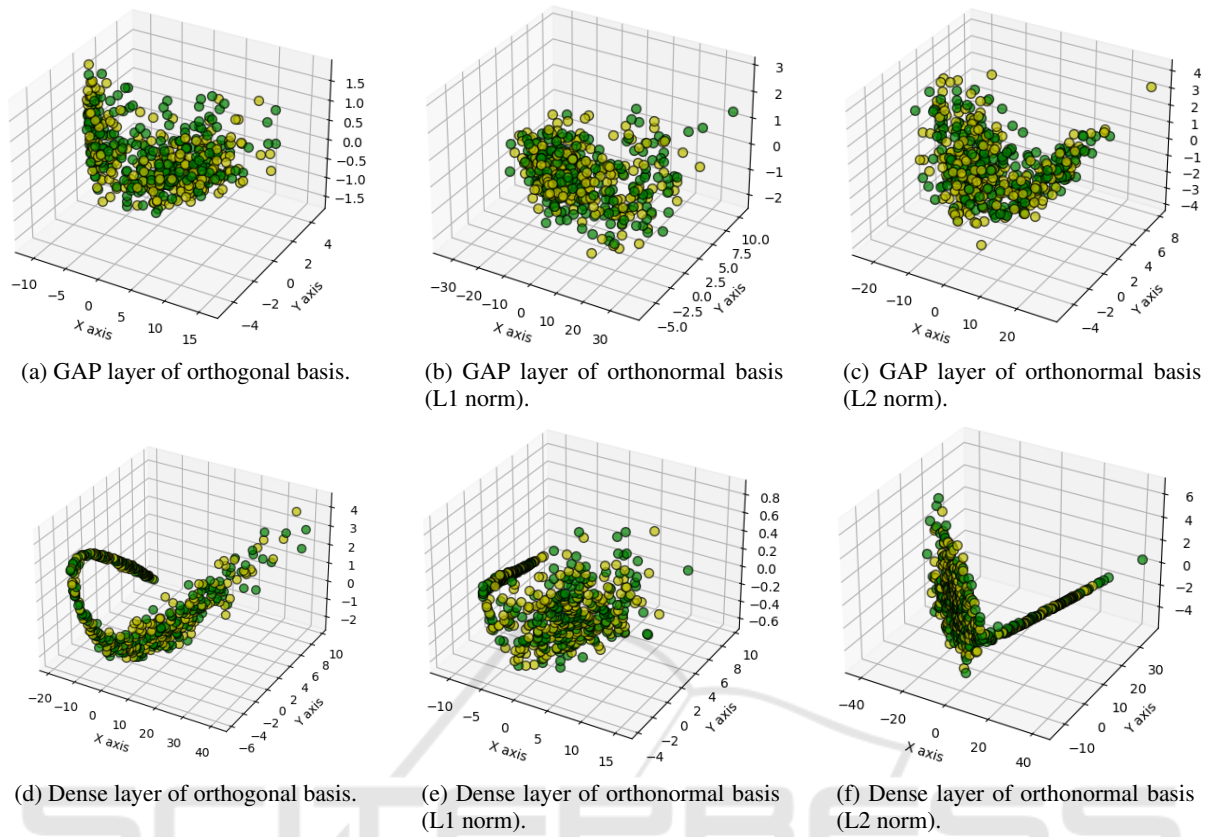


Figure 4: An illustration of the subspace before projection (upper row) and after projection (lower row) onto the subspace. Green is for real image features and yellow is for fake image features.

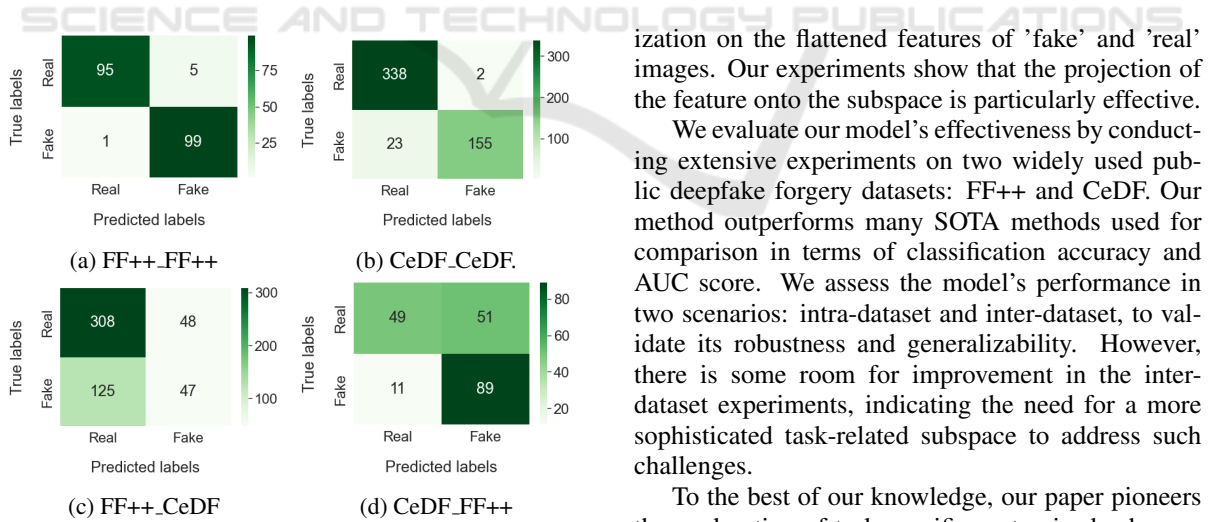


Figure 5: Confusion matrices obtained using intra-dataset (a & b) and inter-dataset (c & d) experimental setups.

tailed features) and shallower features (shape, color, texture, etc.). Additionally, we produce a task-specific subspace for projecting the spatially enriched features. We generate the basis vectors of this customized subspace using the Gram-Smith orthogonal-

ization on the flattened features of 'fake' and 'real' images. Our experiments show that the projection of the feature onto the subspace is particularly effective.

We evaluate our model's effectiveness by conducting extensive experiments on two widely used public deepfake forgery datasets: FF++ and CeDF. Our method outperforms many SOTA methods used for comparison in terms of classification accuracy and AUC score. We assess the model's performance in two scenarios: intra-dataset and inter-dataset, to validate its robustness and generalizability. However, there is some room for improvement in the inter-dataset experiments, indicating the need for a more sophisticated task-related subspace to address such challenges.

To the best of our knowledge, our paper pioneers the exploration of task-specific customized subspace for deepfake classification, with no prior research in this area. Future research needs to be focused on interpretable and customized subspace optimization to achieve enhanced results.

Table 4: Comparison with SOTA methods based on intra-dataset experiment.

Experiment	Method	Test accuracy	AUC score
CeDF	Li et al. (Li et al., 2020)	95.37	98.88
	Afchar et al. (Afchar et al., 2018)	65.83	64.80
	Qian et al. (Qian et al., 2020)	87.06	81.48
	Guo et al. (Guo et al., 2021)	68.33	78.04
	Wang et al. (Wang and Chow, 2023)	70.10	75.89
	Proposed	95.17	93.25
FF++	Li et al. (Li et al., 2020)	96.00	98.34
	Afchar et al. (Afchar et al., 2018)	65.00	66.93
	Qian et al. (Qian et al., 2020)	95.50	96.52
	Guo et al. (Guo et al., 2021)	78.50	87.62
	Zhao et al. (Zhao et al., 2021)	96.00	98.97
	Wang et al. (Wang and Chow, 2023)	84.36	93.99
	Wang et al. (Wang et al., 2023)	92.11	97.66
	Proposed	97.50	97.50

Table 5: Comparison with SOTA methods based on inter-dataset experiment.

Experiment	Method	Test accuracy	AUC score
CeDF_FF++	Li et al. (Li et al., 2020)	64.50	75.19
	Afchar et al. (Afchar et al., 2018)	50.50	51.51
	Qian et al. (Qian et al., 2020)	54.50	54.04
	Ganguly et al. (Ganguly et al., 2022)	65.00	63.80
	Mohiuddin et al. (Mohiuddin et al., 2021)	60.00	59.93
	Mohiuddin et al. (Mohiuddin et al., 2023b)	66.50	76.72
	Proposed	69.00	68.94
FF++.CeDF	Li et al. (Li et al., 2020)	58.06	55.60
	Afchar et al. (Afchar et al., 2018)	66.41	65.58
	Qian et al. (Qian et al., 2020)	63.89	53.89
	Ganguly et al. (Ganguly et al., 2022)	68.04	66.12
	Mohiuddin et al. (Mohiuddin et al., 2021)	63.71	56.43
	Zhao et al. (Zhao et al., 2021)	-	67.44
	Miao et al. (Miao et al., 2021)	-	66.12
	Wang et al. (Wang et al., 2023)	63.27	72.43
	Proposed	68.53	65.06

ACKNOWLEDGEMENTS

This work was supported by the Russian Science Foundation (project No. 22-76-10042).

REFERENCES

- Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE.
- Amerini, I., Galteri, L., Caldelli, R., and Del Bimbo, A. (2019). Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.
- Ganguly, S., Mohiuddin, S., Malakar, S., Cuevas, E., and Sarkar, R. (2022). Visual attention-based deepfake video forgery detection. *Pattern Analysis and Applications*, 25(4):981–992.
- Guo, Z., Yang, G., Chen, J., and Sun, X. (2021). Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding*, 204:103170.
- Hooda, A., Mangaokar, N., Feng, R., Fawaz, K., Jha, S., and Prakash, A. (2024). D4: Detection of adversarial diffusion deepfakes using disjoint ensembles. In *Proceedings of the IEEE/CVF Winter Conference on*

- Applications of Computer Vision*, pages 3812–3822.
- Kingra, S., Aggarwal, N., and Kaur, N. (2023). Siamnet: Exploiting source camera noise discrepancies using siamese network for deepfake detection. *Information Sciences*, page 119341.
- Kou, S., Yin, X., Wang, Y., Chen, S., Chen, T., and Wu, Z. (2023). Structure-aware subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*.
- Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., Prasad, C., and Palaniappan, K. (2020). Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE.
- Li, T., Wang, Y., Liu, L., Chen, L., and Chen, C. P. (2023). Subspace-based minority oversampling for imbalance classification. *Information Sciences*, 621:371–388.
- Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.
- Liz-Lopez, H., Keita, M., Taleb-Ahmed, A., Hadid, A., Huertas-Tato, J., and Camacho, D. (2024). Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Information Fusion*, 103:102103.
- Lu, W., Liu, L., Zhang, B., Luo, J., Zhao, X., Zhou, Y., and Huang, J. (2023). Detection of deepfake videos using long-distance attention. *IEEE Transactions on Neural Networks and Learning Systems*.
- Miao, C., Chu, Q., Li, W., Li, S., Tan, Z., Zhuang, W., and Yu, N. (2021). Learning forgery region-aware and id-independent features for face manipulation detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):71–84.
- Mohiuddin, S., Ganguly, S., Malakar, S., Kaplun, D., and Sarkar, R. (2021). A feature fusion based deep learning model for deepfake video detection. In *International conference on mathematics and its applications in new computer systems*, pages 197–206. Springer.
- Mohiuddin, S., Malakar, S., Kumar, M., and Sarkar, R. (2023a). A comprehensive survey on state-of-the-art video forgery detection techniques. *Multimedia Tools and Applications*, pages 1–41.
- Mohiuddin, S., Sheikh, K. H., Malakar, S., Velásquez, J. D., and Sarkar, R. (2023b). A hierarchical feature selection strategy for deepfake video detection. *Neural Computing and Applications*, 35(13):9363–9380.
- Naskar, G., Mohiuddin, S., Malakar, S., Cuevas, E., and Sarkar, R. (2024). Deepfake detection using deep feature stacking and meta-learning. *Heliyon*.
- Nguyen, D., Mejri, N., Singh, I. P., Kuleshova, P., Astrid, M., Kacem, A., Ghorbel, E., and Aouada, D. (2024). Laa-net: Localized artifact attention network for high-quality deepfakes detection. *arXiv preprint arXiv:2401.13856*.
- Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer.
- Raza, M. A. and Malik, K. M. (2023). Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000.
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.
- Sahib, I. and AlAsady, T. A. A. (2022). Deep fake image detection based on modified minimized xception net and densenet. In *2022 5th International Conference on Engineering Technology and its Applications (IIC-ETA)*, pages 355–360. IEEE.
- Srirangarajan, S. et al. (2022). Locality-aware discriminative subspace learning for image classification. *IEEE Transactions on Instrumentation and Measurement*, 71:1–14.
- Tolosana, R., Romero-Tapiador, S., Fierrez, J., and Vera-Rodriguez, R. (2021). Deepfakes evolution: Analysis of facial regions and fake detection performance. In *international conference on pattern recognition*, pages 442–456. Springer.
- Wang, T., Cheng, H., Chow, K. P., and Nie, L. (2023). Deep convolutional pooling transformer for deepfake detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–20.
- Wang, T. and Chow, K. P. (2023). Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14548–14556.
- Xia, R., Liu, D., Li, J., Yuan, L., Wang, N., and Gao, X. (2024). Mmnet: Multi-collaboration and multi-supervision network for sequential deepfake detection. *IEEE Transactions on Information Forensics and Security*.
- Yin, W., Ma, Z., and Liu, Q. (2023). Discriminative subspace learning via optimization on riemannian manifold. *Pattern Recognition*, 139:109450.
- Yu, C.-M., Chen, K.-C., Chang, C.-T., and Ti, Y.-W. (2022). Segnet: a network for detecting deepfake facial videos. *Multimedia Systems*, 28(3):793–814.
- Yu, Y., Liu, X., Ni, R., Yang, S., Zhao, Y., and Kot, A. C. (2023). Pvass-mdd: predictive visual-audio alignment self-supervision for multimodal deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Zhang, D., Wu, P., Li, F., Zhu, W., and Sheng, V. S. (2022). Cascaded-hop for deepfake videos detection. *KSII Transactions on Internet & Information Systems*, 16(5).
- Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194.
- Zhou, C., Zhong, F., and Öztireli, C. (2023). Clip-pae: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9.