

Evaluating Xception with Multitask Self-Supervised Learning for AI-Generated Image and Deepfake Video Detection

Iksan Nur Rochim¹, R. Rizal Isnanto², Aris Sugiharto³

¹ Master of Information Systems, Postgraduate Program, Diponegoro University, Semarang, Indonesia

^{2,3} Doctor of Information System, Postgraduate Program, Diponegoro University, Semarang, Indonesia

Received:

June 27, 2026

Revised:

August 8, 2026

Accepted:

September 25, 2026

Published:

October 24, 2026

Corresponding Author:

Author Name*:

Iksan Nur Rochim

Email*:

iksan.nurrochim@gmail.com

DOI:

10.63158/journalisi.v8i5.1864

© 2026 Journal of Information Systems and Informatics. This open access article is distributed under a (CC-BY License)



Abstract. This study evaluates an Xception-based Multitask Self-Supervised Learning (MSSL) framework for detecting real and synthetic visual content across image and video modalities. Separate models were developed for CIFAKE images and FaceForensics++ videos. Beyond the primary binary-classification objective, MSSL integrates three auxiliary representation-learning tasks: rotation prediction, patch reconstruction, and artifact-map prediction. These tasks are intended to encourage the model to learn robust semantic and manipulation-related visual features. Experiments used class-balanced subsets of 5,000 CIFAKE images and 840 FaceForensics++ videos. For video classification, five frames were selected through uniform temporal sampling, and frame-level probabilities were averaged to obtain video-level predictions; a three-frame setting was also examined. A fixed decision threshold of 0.50 was defined a priori and applied consistently across both modalities. On CIFAKE, Xception+MSSL achieved 92.60% accuracy and 0.9797 ROC-AUC, outperforming the baseline Xception model, which obtained 79.60% accuracy and 0.8937 ROC-AUC. On FaceForensics++, five-frame MSSL achieved 68.25% video-level accuracy and 0.7365 ROC-AUC, compared with 62.70% and 0.6652 for Xception. The results indicate consistent within-dataset improvements, although cross-dataset and cross-generator generalization remain unverified.

Keywords: AI-generated, deepfake detection, Multitask Self-Supervised Learning, Xception, FaceForensics++, CIFAKE.

1. INTRODUCTION

The rapid development of artificial intelligence (AI)-generated content technologies has significantly transformed the production of visual media through advances in deep learning and generative models, including text-to-image, text-to-video, generative adversarial networks, and diffusion-based models. These technologies enable AI systems to generate highly realistic images and videos that are increasingly difficult to distinguish from authentic visual content. AI-generated visual content has been widely adopted in creative industries, digital marketing, entertainment, and media production because it can reduce production costs and facilitate rapid content creation [1][2]. However, the increasing accessibility and visual quality of synthetic content also introduce substantial risks, including misinformation, manipulation of public opinion, privacy violations, identity misuse, and declining public trust in digital media [3]. Consequently, reliable methods for distinguishing authentic visual content from AI-generated or manipulated content have become increasingly important.

Previous studies have investigated different approaches for detecting synthetic visual content. For AI-generated images, commonly explored approaches include low-level artifact analysis, convolutional image encoders, frequency-based representations, and hybrid methods that combine global semantic information with local visual features [4][5]. For manipulated or synthetic videos, detection methods have additionally exploited temporal inconsistencies and spatio-temporal patterns across consecutive frames [6]. Although these approaches have achieved promising performance on their respective benchmark datasets, their effectiveness may decrease when they are evaluated on data generated by previously unseen generative models or obtained from different data distributions [7]. Therefore, an important challenge in AI-generated content detection is developing visual representations that are less dependent on dataset-specific characteristics and can capture complementary characteristics of synthetic content within the evaluated modality.

A major factor contributing to limited generalization is the tendency of supervised detectors to learn generator-specific or dataset-specific artifacts rather than representations that characterize synthetic content more generally [8][9]. For example, a detector trained on outputs from a particular generative model may rely on distinctive

texture, frequency, compression, or rendering patterns that are not necessarily present in content generated by another model. Furthermore, post-processing operations such as JPEG compression, resizing, noise addition, and resolution degradation may suppress low-level artifacts that conventional detectors use for classification [10]. Existing methods also frequently rely on a single feature representation or optimization objective, which may limit their ability to capture the heterogeneous traces introduced by modern generative techniques [11]. Previous findings indicate that combining complementary local and global information can provide richer representations for synthetic-content detection [12].

Self-supervised learning (SSL) provides a potential approach for learning visual representations from the intrinsic structure of data rather than relying exclusively on class labels [13][14]. In the context of AI-generated content detection, SSL is particularly relevant because synthetic-content artifacts may appear at multiple visual levels and can vary considerably between generative and manipulation processes. By learning representations through auxiliary objectives before or together with supervised classification, a model can be encouraged to capture structural and spatial characteristics beyond the direct binary classification signal. Such representations may provide complementary information for distinguishing authentic and synthetic content [15][16]. In this study, SSL is therefore employed as a representation-learning mechanism to support the learning of manipulation-sensitive visual characteristics. In particular, the rotation-prediction task changes the spatial orientation of the input while largely retaining the underlying manipulation-related visual traces, as the same transformation is applied to the entire image or facial frame rather than introducing new synthetic manipulations. Predicting rotations of 0°, 90°, 180°, and 270° encourages the backbone to learn spatial and structural relationships within facial regions, providing complementary representations that may assist in distinguishing authentic from manipulated faces.

In addition to SSL, multitask learning provides a mechanism for optimizing complementary learning objectives within a shared feature extractor. Rather than relying solely on binary classification between authentic and synthetic samples, auxiliary objectives can guide the backbone toward representations that are sensitive to different characteristics of generated or manipulated content [17][18][19]. In the proposed framework, the auxiliary learning objectives are intended primarily to complement the

main classification task by emphasizing global structural information, local spatial characteristics, and manipulation-related artifact regions. The combination of these auxiliary objectives with the primary classification objective is intended to provide complementary learning signals during representation learning rather than relying solely on a single binary classification objective.

Despite these developments, the integration of self-supervised representation learning and multitask optimization for AI-generated visual-content detection remains relatively limited [20][21]. Furthermore, image and video detection are evaluated separately because video contains temporal information absent from still images. Separate Xception+MSSL models are therefore trained for each dataset, with images evaluated individually and video frame predictions aggregated for video-level evaluation. Thus, this study assesses MSSL within each modality and does not claim cross-modal generalization.

The proposed framework adopts Xception as its backbone architecture. Although more recent architectures such as EfficientNet, Vision Transformer (ViT), ConvNeXt, and Swin Transformer provide competitive visual-recognition capabilities, Xception was selected for three main reasons. First, its depthwise separable convolutions efficiently model spatial and channel-wise information, which is relevant for identifying subtle texture and manipulation artifacts. Second, Xception has been extensively adopted as a reference backbone in deepfake and synthetic-media detection studies, making it suitable for comparison with previous methods [22][23][24]. Third, using Xception allows this study to focus on evaluating the proposed MSSL strategy without the additional complexity of larger architectures. Thus, the contribution lies in the learning framework rather than a new backbone, and no superiority over EfficientNet, ViT, ConvNeXt, or Swin Transformer is claimed because these models were not experimentally compared.

Based on these research gaps, this study proposes an Xception-based MSSL framework for AI-generated image detection on CIFAKE and deepfake video detection on FaceForensics++. It investigates whether MSSL improves classification performance over the corresponding Xception baseline, with frame-level predictions aggregated for video-level evaluation. The two settings are evaluated separately, and conclusions are limited to the evaluated datasets and configurations. Potential applications include media

forensic verification and automated content moderation for screening authentic and manipulated visual content.

The main contributions of this study are as follows:

- 1) An Xception-based MSSL framework integrating rotation prediction, patch reconstruction, and artifact-map prediction as auxiliary representation-learning tasks.
- 2) Evaluation of separate models for AI-generated image detection on CIFAKE and deepfake video detection on FaceForensics++.
- 3) Experimental comparison of Xception+MSSL with the corresponding Xception baseline under the same experimental conditions.
- 4) Evaluation using accuracy, precision, recall, F1-score, and ROC-AUC, with frame-level predictions aggregated for video-level evaluation on FaceForensics++.

Table 1 summarizes previous studies in terms of datasets, detection tasks, methods, performance, and limitations, highlighting remaining challenges in generalization, dataset diversity, and computational efficiency.

Table 1. Related Work

No	Study	Dataset	Task	Method	Metrics	Limitation
1	Bird & Lofi (2024), ("CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images")	CIFAKE	Real vs AI-generated image classification	CNN with hyperparameter tuning; Grad-CAM	Accuracy: 92.98%	The study focuses on AI-generated image classification and does not address video detection or cross-modal evaluation.
2	Batagelj et al. (2025), ("Robust cross dataset deepfake detection with multitask self	FF++, DFD, Celeb-DF-v2, DFDC, DFDC Preview, FFIW, DFF	Deepfake detection and localization in videos	M-Task-SS; one-class SSL; self-blending; EfficientNet-B4 + EVA Transformer; multitask	AUC; IoU; PA; mPA	Robustness to unseen manipulations remains a challenge; segmentation is sensitive to

No	Study	Dataset	Task	Method	Metrics	Limitation
	supervised learning")			classification and segmentation		manipulation-mask type
3	Kusniadi & Setyanto (2021), ("Fake Video Detection using Modified XceptionNet")	FF++, Celeb-DF	Deepfake video detection	MTCNN + XceptionNet/Modified XceptionNet; transfer learning; FPV variation	Accuracy: 88.78% (FF++); 83.75% (Celeb-DF)	Overfitting; limited datasets; modified XceptionNet did not consistently improve performance
4	Sahib & AlAsady (2022), ("Deep fake Image Detection based on Modified Xception Net and DenseNet")	140k Real and Fake Faces: 70,000 real face images from Flickr/NVIDIA A and 70,000 fake face images generated using StyleGAN	Fake face image detection	Modified Xception + DenseNet concept; YCbCr preprocessing	Accuracy: 99.93%; F1: 99.93%; Precision: 99.95%; Recall: 99.91%	Limited to a StyleGAN-based dataset; cross-dataset generalization was not evaluated
5	Shivam et al. (2025), ("Self-Supervised Learning for CIFAR-10 Image Classification")	CIFAR-10: 60,000 color images (32x32 px), 10 classes	Self-supervised image classification	CNN + Self-Supervised Learning (SSL); multi-task rotation prediction + jigsaw puzzle with Color Channel Permutation (CCP); data augmentation	Accuracy: 88% (combined pretext tasks); 85% (Jigsaw + CCP); 82% (rotation prediction); baseline: 75%	Evaluated only on CIFAR-10; scalability to larger and higher-resolution datasets remains uncertain; higher computational requirements may limit deployment on resource-

No	Study	Dataset	Task	Method	Metrics	Limitation
						constrained devices.
						Performance varies across unseen datasets and depends on backbone/model capacity; the framework mainly evaluates face-based deepfake datasets, while broader cross-modal and real-time applications remain to be further explored.
6	Yan et al. (2025), ("Self-supervised Feature Disentanglement for Deepfake Detection")	FF++ (training); Celeb-DF v1/v2, DFD, DFDC, DFDCP (cross-domain testing)	Generalizable deepfake detection	Self-Supervised Feature Disentanglement (SSFD); self-supervised forged-image synthesis; content/forgery feature disentanglement; conditional decoder with AdaIN; EfficientNet-B4 backbone	Average cross-domain AUC: 80.46%; CDFv1: 82.96%; CDFv2: 78.72%; DFD: 82.43%; DFDC: 78.41%; DFDCP: 79.82%	

2. METHODS

2.1. Research Framework

This study proposes a Multitask Self-Supervised Learning (MSSL) approach based on the Xception architecture to classify real and AI-generated visual content in images and videos. The research stages consist of six stages, namely dataset collection, data preprocessing, data splitting, pretraining using Multitask Self-Supervised Learning, fine-tuning using supervised learning, and model evaluation. The overall research flow is shown in Figure 1.

2.2. Dataset

Two public datasets are used in this study: CIFAKE for AI-generated image detection and FaceForensics++ (C23) for deepfake video detection. Although the complete datasets

contain 120,000 images and 7,000 videos, respectively, this study uses fixed class-balanced subsamples rather than the complete datasets.

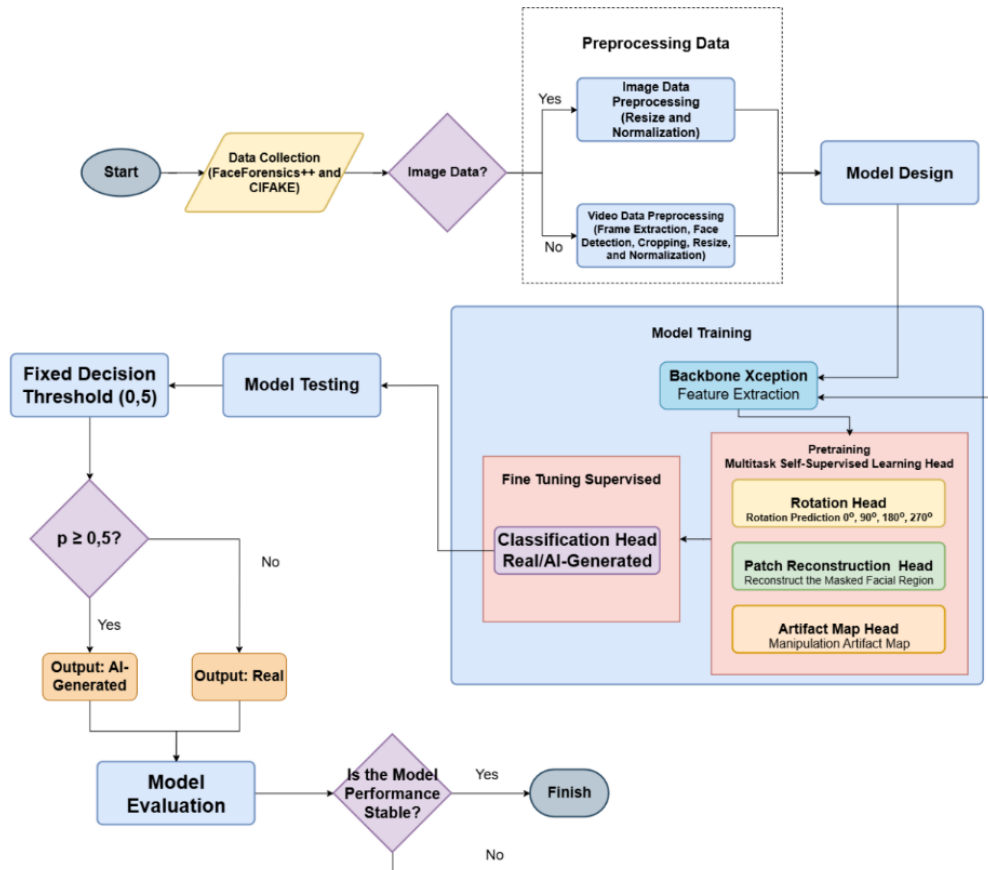


Figure 1. Research Procedure

2.2.1. CIFAKE

The CIFAKE dataset contains 120,000 images, consisting of 60,000 authentic images derived from CIFAR-10 and 60,000 AI-generated images generated using Stable Diffusion v1.4 [25]. From the official training portion of CIFAKE, 2,000 real and 2,000 AI-generated images are sampled to form the training-validation pool. In addition, 500 real and 500 AI-generated images are sampled from the official test portion and reserved as the held-out test set. Therefore, the total number of CIFAKE images used in this study is 5,000 images, with an equal 1:1 class ratio.

2.2.2. FaceForensics++

FaceForensics++ (C23) contains 7,000 videos: 1,000 authentic and 6,000 manipulated using six manipulation methods. To address class imbalance, 70 videos from each

manipulation method (420 manipulated) and 420 authentic videos were deterministically selected, producing a balanced subset of 840 videos. Table 2 summarizes the original and selected dataset compositions and class distributions.

Table 2. Dataset Composition

Dataset	Original Dataset	Selected Subset	Class Distribution
CIFAKE	120.000 images	5.000 images	2.500 real/2.500 fake
FaceForensics++	7.000 videos	840 videos	420 real/420 fake

The detailed train, validation, and test distributions are presented in Table 3.

Table 3. Detailed Train, Validation, and Test Split Distribution

Dataset	Unit	Training	Validation	Testing
CIFAKE	Image	3.200 (1.600/1.600)	800 (400/400)	1.000 (500/500)
FaceForensics++	Video	588 (294/294)	126 (63/63)	126 (63/63)
FaceForensics++ 5 frame	Face- frame	2.940 (1.470/1.470)	630 (315/315)	630 (315/315)
FaceForensics++ 3 frame	Face- frame	1.764 (882/882)	378 (189/189)	378 (189/189)

The values in Table 3 are obtained from the balanced subsets used in the experiments. For CIFAKE, the 4,000-image training-validation pool is divided into 3,200 training images and 800 validation images using an 80:20 stratified split, while the 1,000-image test set is kept independent. For FaceForensics++, the 840 balanced videos are divided into 588 training, 126 validation, and 126 test videos using a 70:15:15 stratified split.

2.3. Preprocessing Data

2.3.1. CIFAKE

For CIFAKE, the 4,000-image training-validation pool is stratified into 3,200 training (1,600 real and 1,600 AI-generated) and 800 validation images (400 per class) using `random_state = 42`. A separate balanced set of 1,000 images (500 per class) is sampled

6689 | Evaluating Xception with Multitask Self-Supervised Learning for AI-Generated

Learning multiple tasks simultaneously is expected to be able to produce more discriminatory feature representations than the single-task learning approach. The design of this research model can be seen in Figure 3.

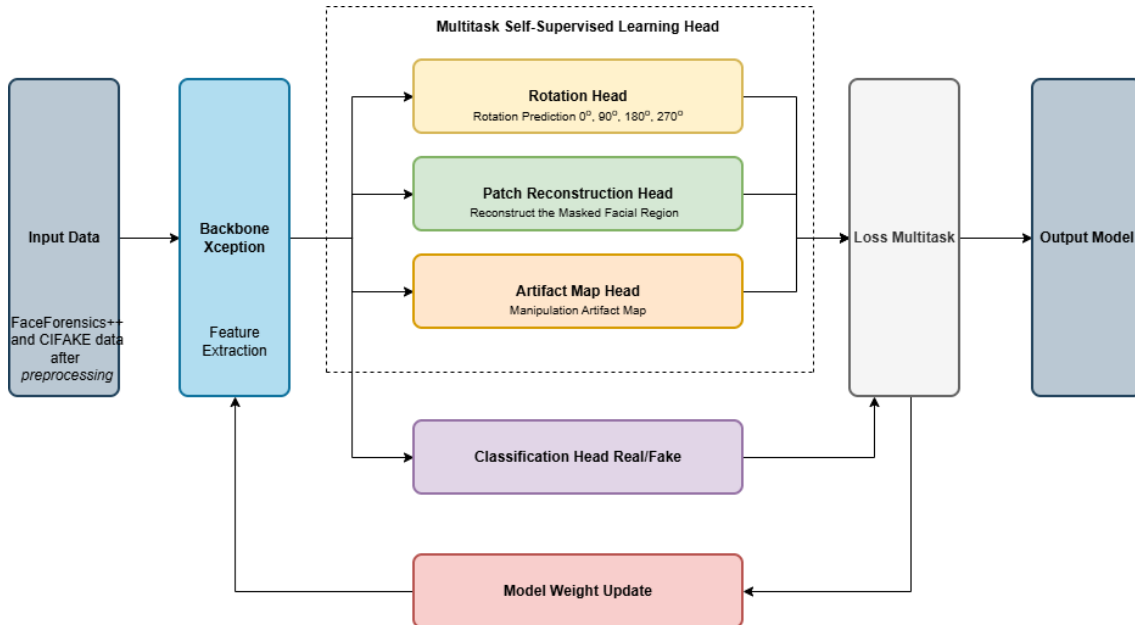


Figure 3. Design the Xception Model with Multitask Self-Supervised Learning

The proposed model consists of an Xception-based shared encoder followed by task-specific heads. Both the MSSL model and the conventional Xception baseline are initialized using ImageNet-pretrained weights, allowing the comparison to begin from identical backbone initialization. During MSSL pretraining, three auxiliary tasks are employed: rotation prediction, masked patch reconstruction, and visual artifact learning.

2.4.1. Rotation Prediction

Following the self-supervised rotation-prediction approach introduced by Gidaris et al [27]. For each input image x , a rotation angle is defined in Equation (1):

$$k \in \{0, 1, 2, 3\} \quad (1)$$

The image is then rotated as defined in Equation (2):

$$\theta = 90^\circ k \quad (2)$$

The rotation head predicts the corresponding four-class rotation label. The rotation loss is defined in Equation (3):

$$\mathcal{L}_{rot} = -\log P(\hat{k} = k | x_{rot}) \quad (3)$$

Where x_{rot} denotes the rotated input and k is the corresponding ground-truth rotation class. This task provides a spatial-structure learning signal to the shared encoder.

2.4.2. Masked Patch Reconstruction

Inspired by the masked-image reconstruction principle used in self-supervised representation learning [28], a portion of the input image is masked and the model is trained to reconstruct the original visual information. After rotation, a random square region is masked from the input image. For the FaceForensics++ pipeline, the side length of the masked square is $h/4$, where $h=299$, and the location of the masked region is randomly selected. The masked image is processed by the encoder, and a transposed-convolution decoder reconstructs the original rotated image. The reconstruction loss is defined in Equation (4):

$$\mathcal{L}_{patch} = \frac{1}{N} \|x_{rot} - \hat{x}_{rot}\|^2 \quad (4)$$

where x_{rot} is the original rotated image, $\hat{x}_{rot}$ is the reconstructed image, and N is the number of pixels.

For CIFAKE, approximately 25% of pixels are randomly masked, whereas FaceForensics++ uses a contiguous square mask. This modality-specific strategy encourages broader local dependency learning in low-resolution CIFAKE images and localized spatial reconstruction in facial frames, and is explicitly reported for methodological transparency.

2.4.3. Visual Artifact Learning

The visual artifact learning task does not use manipulation masks, manipulation labels, or manually annotated artifact regions. Instead, an artifact target is generated from the residual between the rotated image and a locally smoothed representation. In the implemented pipeline, the smoothing operation uses a 5×5 average-pooling kernel with a stride of 1 and SAME padding, the residual target is defined as shown in Equation (5):

$$r = |x_{rot} - AvgPool_{5 \times 5}(x_{rot})| \quad (5)$$

The artifact-learning decoder predicts an artifact representation $\hat{r}$, and the artifact loss is calculated using mean squared error, as defined in Equation (6):

$$\mathcal{L}_{art} = \frac{1}{N} \|r - \hat{r}\|^2 \quad (6)$$

Thus, the auxiliary artifact task provides a generic high-frequency and edge-related learning signal, rather than a manipulation-specific ground-truth mask or pseudo-label.

2.4.4. Multitask Loss

Following the general principle of combining task-specific objectives in multitask learning [29], the total MSSL objective in this study is formulated as a weighted combination of the three auxiliary losses, as defined in Equation (7):

$$\mathcal{L}_{MSSL} = \lambda_{rot}\mathcal{L}_{rot} + \lambda_{patch}\mathcal{L}_{patch} + \lambda_{art}\mathcal{L}_{art} \quad (7)$$

For CIFAKE, no explicit loss_weights are specified in Keras, resulting in an unweighted sum of the three task losses. For FaceForensics++, explicit weights of 1.0, 0.5, and 0.5 are used for rotation, patch reconstruction, and artifact learning, respectively.

2.5. Model Training

Model training consists of MSSL pretraining followed by supervised fine-tuning. MSSL uses only the training subset for three auxiliary tasks, while the baseline is fine-tuned directly from the same ImageNet-pretrained Xception backbone. During fine-tuning, the last 100 Xception layers are unfrozen for CIFAKE and 50 for FaceForensics++, with both FaceForensics++ models using the same 10-epoch budget for fair comparison. The complete configurations are summarized in Table 4.

Table 4. Training Configuration

Parameter	CIFAKE	FaceForensics++
Optimizer	Adam	Adam
Bcakbone initialization	ImageNet	ImageNet

MSSL pretraining learning rate	1×10^{-5}	1×10^{-4}
MSSL pretraining epochs	20	5
Observed MSSL Pretraining runtime	$\approx 35,7$ min	$\approx 9,7$ min
MSSL batch size	8	4
Unfrozen Xception layers	Last 100/32	Last 50/132
Fine-tuning epochs	Grid: 15; best model retraining: 20–30	10
Observed final fine-tuning runtime	$\approx 6,5$ min	$\approx 8,6$ min
Fine-tuning batch size	8	4
Classification loss	Binary cross-entropy	Label-smoothed binary cross-entropy
Label smoothing	-	0,1
Early stopping	Patience 5	Patience 3
Learning rate scheduler	ReduceLROnPlateau, factor 0.5, patience 3	ReduceLROnPlateau, factor 0.5, patience 2, min LR 10^{-7}
Decision threshold	0,5	0,5

2.5.1. Hyperparameter Selection

Hyperparameter tuning is performed using the validation set. The parameters investigated include learning rate, dropout rate, and the number of dense units. For CIFAKE, model selection is based on validation accuracy. For FaceForensics++, model selection is based primarily on video-level validation accuracy, with video-level ROC-AUC used as a tie-break criterion. This ensures that the hyperparameter selection is consistent with the primary evaluation target of the video experiment.

2.5.2. Random Seed and Reproducibility

A global random seed of 42 is used for Python's random module, NumPy, and TensorFlow. FaceForensics++ video sampling additionally uses deterministic random generators for each manipulation method using SEED + method_index, while authentic videos are sampled using the global seed. For CIFAKE, train_test_split uses random_state=42. However, the initial CIFAKE file-list shuffle in the current implementation is not explicitly seeded; therefore, the exact composition of the initially sampled CIFAKE subset may vary between runs.

2.6. Video-Level Prediction

For FaceForensics++, the model initially produces a fake-class probability for each detected face frame. These frame-level predictions are then grouped according to their source video ID. For a video containing N valid face frames, the video-level probability is calculated according to Equation (8):

$$P_{video} = \frac{1}{N} \sum_{i=1}^N P_i \quad (8)$$

where P_i is the predicted fake probability of frame i .

The resulting video-level probability is converted into a binary prediction using a fixed threshold of 0,50. Thus, the primary evaluation of FaceForensics++ is performed at the video level, while frame-level predictions are retained only as supplementary diagnostic information.

2.7. Model Evaluation

The models are evaluated on held-out test data that are excluded from MSSL pretraining, fine-tuning, and hyperparameter selection. CIFAKE is evaluated at the image level, whereas FaceForensics++ uses video-level predictions obtained by aggregating frame-level scores from each source video. Performance is measured using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and the confusion matrix shown in Table 5. These metrics are used to evaluate the effectiveness of Xception-based MSSL in detecting AI-generated visual content.

Table 5. Confusion Matrix

		Prediction Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

True Positive (TP) and True Negative (TN) represent correctly classified positive and negative samples, respectively, whereas False Positive (FP) and False Negative (FN) represent negative and positive samples that are incorrectly classified, respectively.

The evaluation metrics include Accuracy, Precision, Recall, F1-score, and ROC-AUC [30], with Accuracy, Precision, Recall, and F1-score calculated using Equations (9)–(12), respectively:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

$$F1 = 2 \frac{Precision \times Recall}{Precision + Recall} \quad (12)$$

ROC-AUC is calculated as the area under the Receiver Operating Characteristic curve and provides a threshold-independent measure of discriminative ability. In contrast, Accuracy, Precision, Recall, and F1-score are calculated using the fixed classification threshold of 0.50.

2.8. Hardware and Software

All experiments were conducted using a Kaggle Notebook environment with one NVIDIA Tesla T4 GPU with 16 GB memory. The primary software components were in Table 6.

Table 6. Hardware and Software Configuration

Component	Configuration
Computing platform	Kaggle Notebook
GPU	NVIDIA Tesla T4 12.8, 16 GB
Deep learning framework	TensorFlow 2.20.0/Keras 3.13.2
Video processing	OpenCV 4.13.0
Face detection	MTCNN via facenet-pytorch
Video/data bookkeeping	Pandas 2.3.3
GPU execution for MTCNN	PyTorch cuda

3. RESULTS AND DISCUSSION

This section presents the performance of Xception+MSSL and the Xception baseline on CIFAKE image detection and FaceForensics++ video detection. Video performance is evaluated by averaging frame-level probabilities for each source video, with conclusions

limited to the evaluated datasets. Before evaluation, learning rate, dropout rate, and dense units are optimized, as summarized in Table 7.

Table 7. Hyperparameter

Learning Rate	Dropout Rate	Dense Units
1×10^{-4}	0,3	256
1×10^{-5}	0,5	512

3.1. Experimental Dataset Summary

The experiments use class-balanced subsets. For CIFAKE, 4,000 training images are divided into training and validation sets, while 1,000 balanced test images are retained independently. For FaceForensics++, 840 balanced videos (420 authentic and 420 manipulated) are split at the video-ID level to prevent data leakage across subsets. The final training, validation, and test distributions are summarized in Table 8.

Table 8. Dataset Summary Used in the Experiments

Dataset	Original Size	Selected Subset	Training	Validation	Testing
CIFAKE	120.000	5.000	3.200	800	1.000
FaceForensics++	7.000	840	588	126	126
FF++ 5 frame	7.000	840	2.940	630	630
FF++ 3 frame	7.000	840	1.764	378	378

The video-level test set therefore contains 126 unique source videos for both the five-frame and three-frame experiments. The difference between the two scenarios is the number of frames extracted from each source video, not the number of test videos. This corrected protocol replaces the earlier frame-level splitting procedure that produced inconsistent test-set sizes.

3.2. CIFAKE Image Classification

3.2.1. Hyperparameter Tuning

Hyperparameter tuning was performed separately for both models' using validation accuracy, without involving the test set. The best Xception+MSSL configuration used a

learning rate of 1×10^{-5} , dropout of 0.5, and 512 dense units, achieving 95.25% validation accuracy, while the baseline achieved 87.25% with 1×10^{-4} , 0.3 dropout, and 256 dense units. The selected configurations are summarized in Table 9. These results indicate that the two learning strategies favor different optimization settings. However, validation accuracy is used only for model selection and is not interpreted as evidence of generalization beyond the evaluated dataset.

Table 9. Best Hyperparameter Configuration for CIFAKE

Model	Learning Rate	Dropout	Dense Units	Best Validation Accuracy
Xception + MSLL	1×10^{-5}	0,5	512	95,25%
Xception	1×10^{-4}	0,3	256	87,25%

3.2.2. Overall CIFAKE Performance

The final test results are presented in Table 9. Xception+MSLL achieved a testing accuracy of 92.60% and ROC-AUC of 0.9797, whereas the Xception baseline obtained 79.60% accuracy and ROC-AUC of 0.8937. The corresponding testing losses were 0.2039 and 0.4506, respectively. Table 10 presents the overall performance comparison between Xception + MSLL and Xception on the CIFAKE dataset.

Table 10. Overall Performance on CIFAKE

Metric	Xception + MSLL	Xception
Training Accuracy	99,91%	88,72%
Validation Accuracy	95,00%	85,25%
Testing Accuracy	92,60%	79,60%
Validation Loss	0,1485	0,3585
Testing Loss	0,2039	0,4506
ROC-AUC	0,9797	0,8937

Xception+MSLL achieved a 13.00-percentage-point higher test accuracy than the baseline on CIFAKE, with a higher ROC-AUC indicating better class separation. However, this is

reported as an empirical improvement rather than a statistically significant difference because repeated runs and statistical hypothesis tests were not conducted.

3.2.3. Per-Class Performance on CIFAKE

Table 11 reports class-specific precision, recall, F1-score, and support. Table 10 presents the class-wise performance of Xception + MSSL and Xception on the CIFAKE dataset. Table 11 presents the class-wise classification performance of Xception + MSSL and Xception on the CIFAKE dataset.

Table 11. Per-Class Classification Performance on CIFAKE

Model	Class	Precision	Recall	F1-Score	Support	ROC-AUC
Xception+MSSL	REAL	0,93	0,92	0,93	500	0,9797
Xception+MSSL	FAKE	0,92	0,93	0,93	500	0,9797
Xception	REAL	0,80	0,79	0,79	500	0,8937
Xception	FAKE	0,79	0,81	0,80	500	0,8937

Figure 4 shows the confusion matrix of the Xception + MSSL model on the CIFAKE dataset. Xception+MSSL produced 74 misclassifications, compared with 204 errors from the baseline. Its relatively balanced precision and recall indicate that the performance improvement was not driven by a strong bias toward either class.

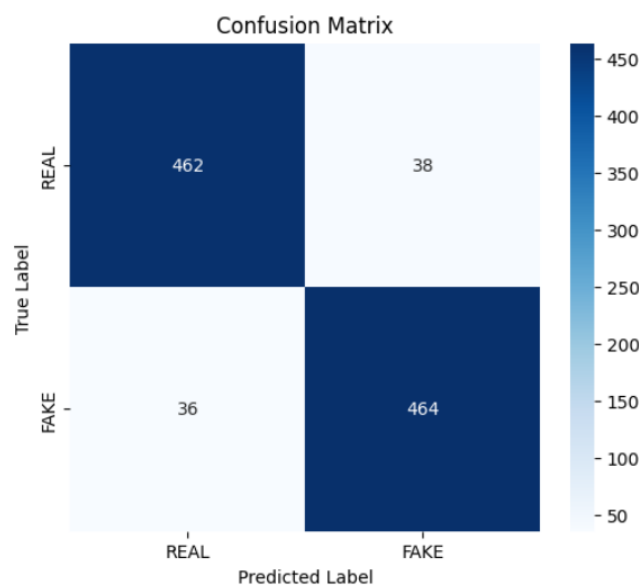


Figure 4. Confusion Matrix CIFAKE Image Classification Xception+MSSL

3.2.4. Training-Test Gap and Overfitting

Despite its higher CIFAKE performance, Xception+MSSL showed a 7.31-percentage-point training–test gap (99.91% vs. 92.60%), indicating some remaining overfitting. Thus, MSSL improved held-out performance but did not eliminate overfitting. Moreover, since CIFAKE uses Stable Diffusion v1.4 and no other generators were tested, these results do not demonstrate robustness to unseen generators.

3.2.5. Statistical Analysis

A post-hoc bootstrap analysis with 2,000 resamples was performed on the held-out CIFAKE test predictions at a fixed threshold of 0.50. The analysis yielded 95.10% accuracy and a ROC-AUC of 0.9899, with the 95% confidence intervals summarized in Table 12.

Table 12. Bootstrap 95% Confidence Intervals for Xception+MSSL on CIFAKE

Metric	Point Estimate	Bootstrap Mean	95% CI Lower	95% CI Upper
Accuracy	0,9510	0,9507	0,9370	0,9630
Precision	0,9519	0,9514	0,9321	0,9693
Recall	0,9500	0,9498	0,9304	0,9688
F1-Score	0,9510	0,9506	0,9368	0,9635
ROC-AUC	0,9899	0,9898	0,9852	0,9939

The relatively narrow bootstrap intervals indicate stable performance estimates within the evaluated test samples, although they do not capture variability from repeated model training or random initialization. Bootstrap resampling was performed at the image level for CIFAKE and at the source-video level for FaceForensics++, consistent with their respective evaluation units. However, paired statistical comparisons with the Xception baseline were not performed because baseline predictions for the same ordered test samples were unavailable; therefore, statistical significance between the models cannot be established.

3.2.6. Feature Embedding Visualization

To examine the learned representations, 256-dimensional feature vectors were extracted from the penultimate dense layer of Xception+MSSL for all 1,000 CIFAKE test images.

PCA and t-SNE were applied for two-dimensional qualitative visualization, with the PCA projection shown in Figure 5.

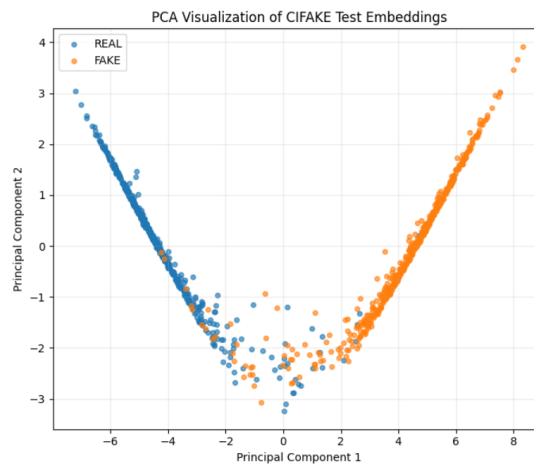


Figure 5. PCA Visualization of CIFAKE Test Embeddings. The silhouette score of 0.6328 indicates substantially stronger REAL-FAKE feature separation than that observed for FaceForensics++ (0.0493).

The first two principal components explained 89.88% and 6.78% of the variance (96.66% cumulative). PCA shows substantial separation between real and AI-generated samples along the first component, although some overlap remains, indicating a strong class-related representation with several ambiguous samples. Figure 6 presents the corresponding t-SNE visualization.

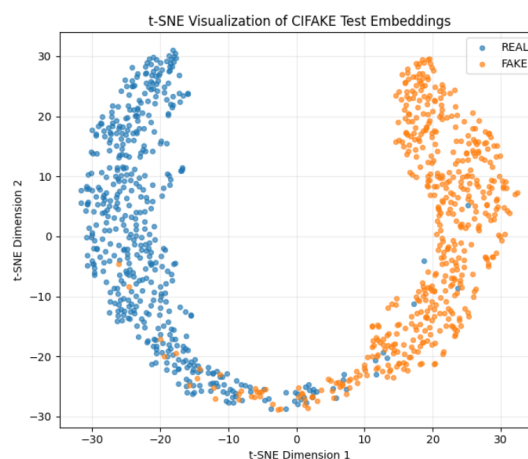


Figure 6. t-SNE Visualization of CIFAKE Test Embeddings. The silhouette score of 0.6328 reflects stronger class separation than the FaceForensics++ video-level embeddings (0.0493).

The t-SNE projection shows two predominantly class-specific regions with limited overlap, consistent with a silhouette score of 0.6328 in the original 256-dimensional embedding space. This indicates moderately strong class separation, however, these descriptive analyses do not demonstrate cross-dataset or unseen-generator generalization.

3.3. FaceForensics++ Video Classification Result

FaceForensics++ was split at the video-ID level to prevent data leakage, with both five- and three-frame scenarios evaluated on the same 126-video test set (63 real and 63 fake). Although predictions were generated per frame, frame probabilities from each source video were averaged and classified using the predefined threshold. Thus, all reported metrics represent video-level performance.

3.3.1. Five-Frame Xception+MSSL

The five-frame Xception+MSSL model achieved a video-level accuracy of 68.25% and a ROC-AUC of 0.7365 on the 126 held-out test videos. The classification results for the five-frame Xception + MSSL model are further illustrated by the confusion matrix in Figure 7.

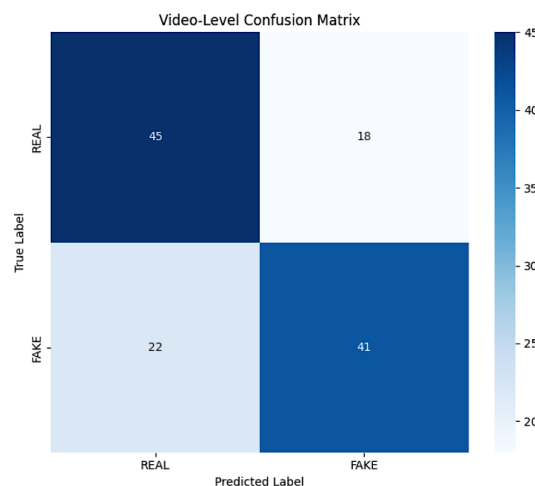


Figure 7. Confusion Matrix FaceForensics++ Five-Frame Xception+MSSL

Where the first row corresponds to real videos and the second row corresponds to fake videos. Of the 63 real videos, 45 were correctly classified as real, whereas 18 were incorrectly classified as fake. For the fake class, 41 of the 63 manipulated videos were

correctly detected, while 22 were incorrectly classified as real. Table 13 summarizes the class-wise video-level performance of the five-frame Xception + MSSL model.

Table 13. Five-Frame Xception+MSSL Video-Level Performance

Class	Precision	Recall	F1-Score	Support
REAL	0,6250	0,6349	0,6299	63
FAKE	0,6290	0,6190	0,6240	63
Macro Average	0,6270	0,6270	0,6270	126

The five-frame Xception+MSSL model improved video-level accuracy from 62.70% to 68.25% (+5.55 percentage points) and ROC-AUC from 0.6652 to 0.7365, with improvements also observed in macro-averaged metrics. However, these results do not establish statistical superiority because they are based on a single experimental split and training configuration. The revised results differ from the original manuscript because video-ID-based splitting was applied before frame extraction, preventing data leakage and providing a more appropriate estimate of performance on unseen source videos.

3.3.2. Five-Frame Xception Baseline

Using the same corrected video-ID split and the same five-frame representation, the Xception baseline obtained a video-level accuracy of 62.70% and ROC-AUC of 0.6652. Figure 8 presents the confusion matrix obtained from the five-frame Xception baseline model.

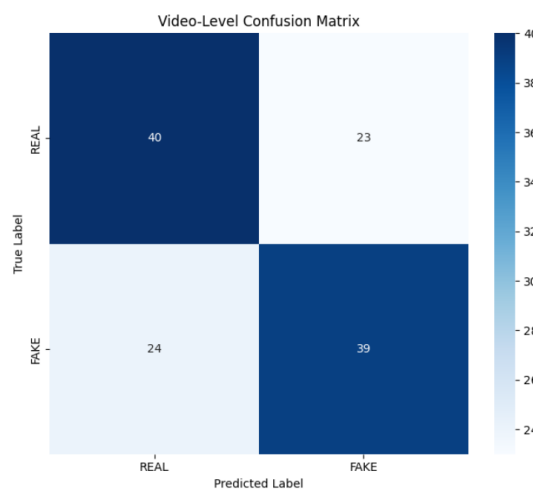


Figure 8. Confusion Matrix FaceForensics++ Five-Frame Xception Baseline

Of the 63 REAL and 63 FAKE videos, the baseline correctly classified 40 REAL and 39 FAKE videos, with 23 false positives and 24 false negatives, indicating relatively balanced errors between the two classes. The corresponding classification results are shown in Table .14

Table 14. Five-Frame Xception Baseline Video-Level Performance

Class	Precision	Recall	F1-Score	Support
REAL	0,6250	0,6349	0,6299	63
FAKE	0,6290	0,6190	0,6240	63
Macro Average	0,6270	0,6270	0,6270	126

The baseline showed comparable performance across classes, although its lower FAKE recall (0.6190) resulted in 24 of 63 manipulated videos (38.1%) being misclassified as REAL, which is particularly relevant for forensic and content-moderation applications. Compared with the baseline, Xception+MSSL improved accuracy from 62.70% to 68.25% and ROC-AUC from 0.6652 to 0.7365, while reducing false positives from 23 to 18 and false negatives from 24 to 22. These corrected results replace the earlier frame-level evaluation, as the revised video-ID-based splitting prevents frames from the same source video from appearing across different subsets and provides a more appropriate estimate of performance on unseen videos.

3.3.3. Three-Frame Xception+MSSL

The three-frame Xception+MSSL experiment was conducted using the same 126 test videos employed in the five-frame experiment, while reducing the number of representative face frames extracted from each video from five to three. This experimental setting allows the effect of the number of sampled frames to be examined while maintaining the same video-level test composition. The three-frame model achieved a video-level accuracy of 65.87% and a ROC-AUC of 0.6979. Figure 9 shows the confusion matrix of the three-frame Xception + MSSL model. Of the 63 real videos, 42 were correctly classified and 21 were misclassified as fake. For the fake class, 41 videos were correctly detected, whereas 22 were incorrectly classified as real. The detailed class-specific results are presented in Table 15.

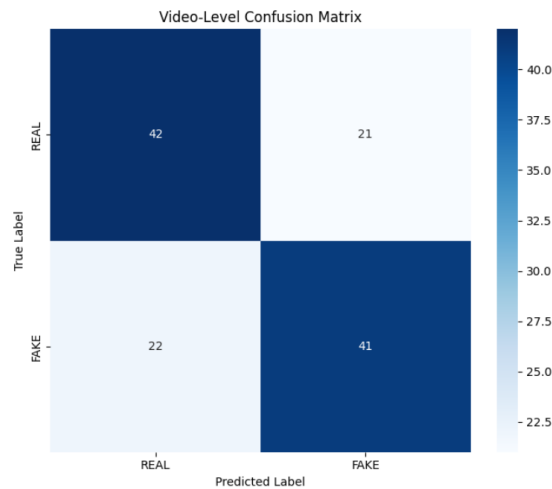


Figure 9. Confusion Matrix of three-frame Xception+MSSL

Table 15. Three-Frame Xception+MSSL Video-Level Performance

Class	Precision	Recall	F1-Score	Support
REAL	0,6562	0,6667	0,6614	63
FAKE	0,6613	0,6508	0,6560	63
Macro Average	0,6588	0,6587	0,6587	126

The five-frame Xception+MSSL model outperformed the three-frame configuration in accuracy (68.25% vs. 65.87%), ROC-AUC (0.7365 vs. 0.6979), and macro F1-score (0.6822 vs. 0.6587). Both configurations produced 22 false negatives, while five-frame sampling reduced false positives from 21 to 18, indicating that the improvement mainly resulted from better recognition of REAL videos. However, these results represent empirical differences under the evaluated setting and do not establish the general superiority of five-frame sampling without repeated experiments or statistical testing.

3.3.4. Feature Embedding Visualization

For the five-frame Xception+MSSL model, 128-dimensional frame-level embeddings from the penultimate dense layer were averaged by source video ID, producing a 126×128 video-level representation consistent with the evaluation protocol. PCA and t-SNE were applied for visualization, with the first two PCA components explaining 28.48% and 19.41% of the variance (47.89% cumulative). The PCA projection is shown in Figure 10.

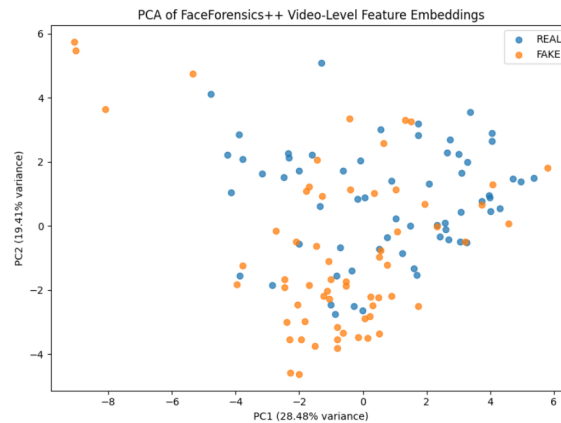


Figure 10. PCA Visualization of FaceForensics++ Video-Level Feature Embeddings. The lower silhouette score (0.0493) compared with CIFAKE (0.6328) reflects greater overlap between REAL and FAKE representations.

The PCA projection shows substantial overlap between REAL and FAKE representations, indicating limited linear separation in the first two components. However, these components explain only 47.89% of the total variance and therefore do not represent the complete 128-dimensional feature space. The corresponding t-SNE visualization is shown in Figure 11.

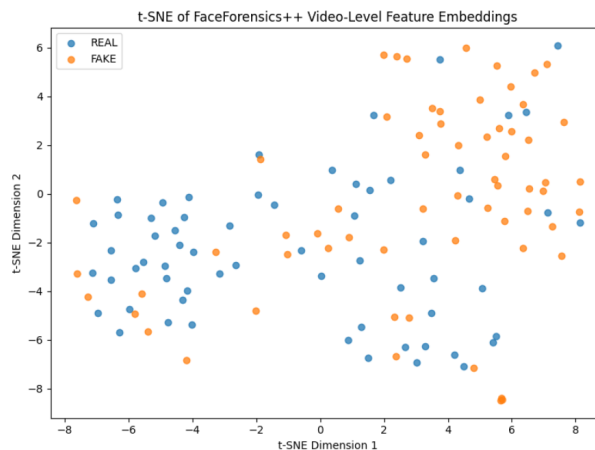


Figure 11. t-SNE Visualization of FaceForensics++ Video-Level Feature Embeddings. The low silhouette score (0.0493), compared with CIFAKE (0.6328), indicates weaker global REAL-FAKE feature separation.

The t-SNE projection shows substantial overlap between REAL and FAKE samples, consistent with the low silhouette score of 0.0493, indicating weak global separation in the 128-dimensional embedding space. Compared with CIFAKE, FaceForensics++ exhibits

greater class overlap, suggesting more challenging feature separation under the current Xception+MSSL representation. However, the silhouette score describes embedding organization rather than classification performance and may not reflect nonlinear decision boundaries used by the classifier.

3.4. Comparison of All Experimental Scenarios

To provide a direct comparison across all experiments, Table 16 summarizes the primary evaluation metrics.

Table 16. Overall Performance Comparison Across All Experimental Scenarios

Dataset	Model	Frames	Evaluation Units	Accuracy	Macro Precision	Macro Recall	Macro F1	ROC-AUC
CIFAKE	Xception	-	Image	0,7960	0,80	0,80	0,80	0,8937
CIFAKE	Xception+MSSL	-	Image	0,9260	0,93	0,93	0,93	0,9797
FaceForensics	Xception	5	Video	0,6270	0,6270	0,6270	0,6270	0,6652
FaceForensics++	Xception+MSSL	5	video	0,6825	0,6833	0,6825	0,6822	0,7365
FaceForensics++	Xception+MSSL	3	Video	0,6587	0,6588	0,6587	0,6587	0,6979

MSSL showed its largest empirical improvement on CIFAKE and a more modest gain on FaceForensics++, where overall performance remained lower. Thus, the observed benefit of MSSL is dataset-dependent rather than uniform across datasets or modalities.

3.5. Discussion

The results indicate that the contribution of Multitask Self-Supervised Learning (MSSL) is dependent on the evaluated dataset and experimental configuration rather than being uniform across visual-content detection tasks. On CIFAKE, Xception+MSSL achieved an accuracy of 92.60% and a ROC-AUC of 0.9797, compared with 79.60% and 0.8937 for the Xception baseline. On the corrected FaceForensics++ video-level evaluation, the five-frame Xception+MSSL model achieved 68.25% accuracy and a ROC-AUC of 0.7365, compared with 62.70% and 0.6652 for the corresponding Xception baseline. These findings therefore support an observed within-dataset benefit under the evaluated configurations, but they do not demonstrate cross-dataset, cross-generator, or cross-modal generalization.

3.5.1. The Effect of Multitask Self-Supervised Learning on Image Classification

For CIFAKE, the three MSSL objectives provide complementary learning signals for global structure, local dependencies, and artifact-related features, contributing to better held-out performance than the baseline. However, the 7.31-percentage-point training–test gap indicates remaining overfitting, and the use of Stable Diffusion v1.4 samples limits conclusions about unseen generators. The embedding analysis further showed strong REAL–FAKE separation, with a silhouette score of 0.6328, supporting the observed within-dataset performance but not external generalization.

3.5.2. The Effect of MSSL on Deepfake Video Detection

The corrected FaceForensics++ experiment represents a more challenging setting involving multiple manipulation methods and frame-based face crops. Xception is a general image-classification architecture rather than a dedicated temporal deepfake model, although it has been widely adopted for deepfake detection. Under the corrected video-ID split, Xception+MSSL improved video-level accuracy from 62.70% to 68.25% and ROC-AUC from 0.6652 to 0.7365, indicating a modest improvement over the baseline, although the gain was smaller than on CIFAKE.

The smaller improvement may reflect the image-oriented nature of the MSSL tasks, which emphasize spatial and high-frequency features rather than temporal manipulation cues. Task conflicts and fixed loss weights may also affect multitask optimization; however, without formal ablation and loss-weight analyses, these remain plausible explanations rather than demonstrated causes. A key limitation is that the video pipeline averages predictions from uniformly sampled frames without explicitly modeling temporal information such as motion or frame-to-frame inconsistencies. Future work should therefore investigate temporal architectures such as CNN-LSTM, 3D CNN, TimeSformer, ViViT, or temporal-attention models.

3.5.3. Error Characteristics and Practical Detection Trade-offs

In the five-frame experiment, Xception+MSSL produced 22 false negatives and 18 false positives, compared with 24 and 23 for the baseline. Although MSSL reduced both error types, its fake-class recall of 0.6508 indicates that approximately one-third of manipulated videos were still missed. In forensic or content-moderation applications, false negatives may allow manipulated content to pass as authentic, while reducing them

by lowering the threshold may increase false positives and manual-review workload. Therefore, deployment thresholds should consider the operational costs of both error types rather than accuracy alone.

3.5.4. Effect of Frame Count on Detection Performance

Five-frame sampling achieved higher performance than three-frame sampling, improving accuracy from 65.87% to 68.25%, macro F1-score from 0.6587 to 0.6822, and ROC-AUC from 0.6979 to 0.7365. Both configurations produced 22 false negatives, while false positives decreased from 21 to 18, indicating that the improvement mainly resulted from better recognition of authentic videos.

Five frames provide more spatial observations for probability averaging, potentially reducing sensitivity to ambiguous individual frames, but they do not constitute explicit temporal modeling. Therefore, the observed advantage is specific to the evaluated setting and does not establish the general superiority of five-frame sampling without repeated-run statistical comparisons.

3.5.5. Feature Representation in FaceForensics++

The FaceForensics++ embeddings showed substantial REAL-FAKE overlap in PCA and t-SNE, with the first two PCA components explaining 47.89% of the variance and a low silhouette score of 0.0493. This weak separation is consistent with the moderate classification performance, although the embedding analysis remains descriptive and does not directly represent classifier performance.

3.5.6. Hyperparameter and Multitask Optimization Considerations

The optimal hyperparameters varied across models and datasets, indicating that the additional MSSL objectives alter the optimization requirements rather than providing a universally optimal configuration. CIFAKE and FaceForensics++ required different MSSL settings, supporting the need for dataset- and objective-specific tuning. Future work should systematically evaluate the loss weights and individual contributions of rotation prediction, reconstruction, and artifact learning to determine whether each task contributes positively or introduces negative transfer.

3.5.7. Research Implications and Limitations

Overall, MSSL provides complementary representation-learning signals, but its benefits remain dataset- and modality-dependent. The current results provide within-dataset evidence of improvement over the Xception baseline, with a larger gain on CIFAKE than on FaceForensics++. However, broader robustness and generalization cannot be claimed because cross-dataset, unseen-generator, degradation, stronger-baseline, and ablation evaluations were not performed. Future work should therefore include external datasets and unseen manipulations, repeated runs, robustness testing, stronger baselines, temporal video models, and systematic analysis of auxiliary tasks and loss weights.

4. CONCLUSION

This study evaluated Xception-based MSSL for AI-generated image detection on CIFAKE and deepfake video detection on FaceForensics++. Xception+MSSL achieved 92.60% accuracy and 0.9797 ROC-AUC on CIFAKE, outperforming the Xception baseline (79.60%, 0.8937), while the five-frame configuration achieved 68.25% accuracy and 0.7365 ROC-AUC on FaceForensics++, compared with 62.70% and 0.6652 for the baseline and 65.87% and 0.6979 for the three-frame configuration. These findings demonstrate within-dataset improvements but do not establish cross-dataset, cross-generator, or robustness generalization. Key limitations include the absence of external and degradation testing, formal MSSL-task ablation, stronger baseline comparisons, repeated-run analysis, and explicit temporal modeling. Future work should therefore evaluate unseen generators and manipulation methods, robustness under realistic degradations, the individual contributions and loss weights of MSSL tasks, stronger convolutional and transformer-based baselines, and temporal architectures such as CNN-LSTM, 3D CNN, TimeSformer, ViViT, or temporal-attention models.

REFERENCES

- [1] D. S. Chinta, S. Kamineni, R. P. Chatragadda, and S. Kamepalli, "Analyzing Image Classification on AI-Generated Art Vs Human Created Art Using Deep Learning Models," in *2024 3rd International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICEEICT61591.2024.10718485.

- [2] M. Z. Hossain, F. Uz Zaman, and M. R. Islam, "Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights," in *2023 26th International Conference on Computer and Information Technology, ICCIT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCIT60459.2023.10440990.
- [3] A. Singh, "A Survey of AI Text-to-Image and AI Text-to-Video Generators," in *2023 4th International Conference on Artificial Intelligence, Robotics and Control, AIRC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 32–36. doi: 10.1109/AIRC57904.2023.10303174.
- [4] D. Park, H. Na, and D. Choi, "Performance Comparison and Visualization of AI-Generated-Image Detection Methods," *IEEE Access*, vol. 12, pp. 62609–62627, 2024, doi: 10.1109/ACCESS.2024.3394250.
- [5] J. Hall, J. Tang, and T. Cao, "Machine Learning Classification of AI and Real Images," in *2025 Intermountain Engineering, Technology and Computing, IETC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/IETC64455.2025.11039474.
- [6] L. Ma, Z. Yan, Q. Guo, Y. Liao, H. Yu, and P. Zhou, "Detecting AI-Generated Video via Frame Consistency," in *Proceedings - IEEE International Conference on Multimedia and Expo*, IEEE Computer Society, 2025. doi: 10.1109/ICME59968.2025.11210049.
- [7] J. Lin, Q. Zhou, and J. Li, "Dual-path AI-generated image detection: Leveraging texture-rich and texture-poor patches with global semantic features," *Digital Signal Processing: A Review Journal*, vol. 171, Mar. 2025, doi: 10.1016/j.dsp.2025.105840.
- [8] H. Yu, Y. Ye, J. Rong, Q. Xuan, and C. Ma, "RealHD: A High-Quality Dataset for Robust Detection of State-of-the-Art AI-Generated Images," *Association For Computing Machinery (ACM)*, Oct. 2025, pp. 11394–11403. doi: 10.1145/3746027.3755207.
- [9] J. Park and A. Owens, "Community Forensics: Using Thousands of Generators to Train Fake Image Detectors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2025, pp. 8245–8257. doi: 10.1109/CVPR52734.2025.00772.
- [10] S. Gye, J. Ko, H. Shon, M. Kwon, and J. Kim, "Reducing the Content Bias for AI-generated Image Detection," in *Proceedings - 2025 IEEE Winter Conference on Applications of Computer Vision, WACV 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 399–408. doi: 10.1109/WACV61041.2025.00049.

- [11] J. Zou, J. Wang, K. Lu, Y. Lai, K. Luo, and Z. Yu, "E2GenF: Universal AIGC image detection based on edge enhanced generalizable features," *Pattern Recognit. Lett.*, vol. 200, pp. 74–80, Feb. 2025, doi: 10.1016/j.patrec.2025.12.001.
- [12] A. H. Javaheri, H. Motamednia, and A. Mahmoudi-Azanveh, "Enhancing the Generalization of Synthetic Image Detection Models through the Exploration of Features in Deep Detection Models," in *Iranian Conference on Machine Vision and Image Processing, MVIP*, IEEE Computer Society, 2024. doi: 10.1109/MVIP62238.2024.10491148.
- [13] T. Sun, "Computer vision object detection and image recognition algorithm optimization based on self-supervised learning," *IET Conference Proceedings*, vol. 2025, no. 25, pp. 636–641, Oct. 2025, doi: 10.1049/icp.2025.3123.
- [14] N. G. Eapen and J. George, "Exploring Self-Supervised Learning Architectures for Image Processing: Milestones and Challenges," in *2024 IEEE 4th International Conference on ICT in Business Industry and Government, ICTBIG 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICTBIG64922.2024.10911752.
- [15] L. Wittscher, J. Diers, and C. Pigorsch, "Improving image classification robustness using self-supervision," in *Stat*, John Wiley and Sons Inc, Dec. 2022. doi: 10.1002/sta4.455.
- [16] C. Tang, Y. Zhou, W. Qin, Y. Zhang, R. Wu, and W. Wang, "Transformer-Driven Self-Supervised Learning for Visual Understanding: Methods and Applications," in *2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology, AIoT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 155–159. doi: 10.1109/AIoT66747.2025.11198691.
- [17] M. Zou, B. Yu, Y. Zhan, S. Lyu, and K. Ma, "Semantics-Oriented Multitask Learning for DeepFake Detection: A Joint Embedding Approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 10, pp. 9950–9963, 2025, doi: 10.1109/TCSVT.2025.3572508.
- [18] Y. Zheng, H. Ahn, and E. Park, "Art or Artifact? Segmenting AI-Generated Images for Deeper Detection," in *ACM WDC 2025 - Proceedings of the 4th Workshop on the security implications of Deepfakes and Cheapfakes*, Association for Computing Machinery, Inc, Aug. 2025, pp. 9–14. doi: 10.1145/3709022.3736544.

- [19] Y. Yao, N. Desai, and M. Palaniswami, "Masked Contrastive Representation Learning For Self-Supervised Visual Pre-Training," in *2024 IEEE 11th International Conference on Data Science and Advanced Analytics, DSAA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/DSAA61799.2024.10722789.
- [20] Y. Jiang and R. A. Frick, "Same Same, But Different?: Detecting AI-Generated Videos Using Knowledge Transfer," in *ACM WDC 2025 - Proceedings of the 4th Workshop on the security implications of Deepfakes and Cheapfakes*, Association for Computing Machinery, Inc, Aug. 2025, pp. 38–43. doi: 10.1145/3709022.3736547.
- [21] J. Park, L. H. Park, H. E. Ahn, and T. Kwon, "Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge," *IEEE Access*, vol. 12, pp. 11674–11687, 2024, doi: 10.1109/ACCESS.2024.3353785.
- [22] P. Joshi and V. Nivethitha, "Deep Fake Image Detection using Xception Architecture," in *5th International Conference on Recent Trends in Computer Science and Technology, ICRTCST 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 533–537. doi: 10.1109/ICRTCST61793.2024.10578398.
- [23] K. Yesugade and R. Jadhav, "Comparative Evaluation of Xception and DenseNet121 for Robust Deepfake Detection," in *2025 12th International Conference on Computing For Sustainable Global Development*, 2025. doi: 10.23919/INDIACom66777.2025.11115841.
- [24] L. K. Gautam, "Extreme Inception Model and Its Application in Indian Food Image Classification," in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry, IDICAIEI 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IDICAIEI61867.2024.10842726.
- [25] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024, doi: 10.1109/ACCESS.2024.3356122.
- [26] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.195.
- [27] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," *conference paper at ICLR*, Mar. 2018, doi: 10.48550/arXiv.1803.07728.

- [28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," *Conference on Computer Vision and Pattern Recognition*, 2022, doi: 10.1109/CVPR52688.2022.01553.
- [29] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2018, pp. 7482–7491. doi: 10.1109/CVPR.2018.00781.
- [30] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.