UNMASKING DEEPFAKES: MASKED AUTOENCODING SPATIOTEMPORAL TRANSFORMERS FOR ENHANCED VIDEO FORGERY DETECTION

Sayantan Das[1], **Mojtaba Kolahdouzi**[1], **Levent Özparlak**[2], **Will Hickie**[2], **Ali Etemad**[1]
[1]Queen's University, Canada
[2] Irdeto BV

ABSTRACT

We present a novel approach for the detection of deepfake videos using a pair of vision transformers pre-trained by a self-supervised masked autoencoding setup. Our method consists of two distinct components, one of which focuses on learning spatial information from individual RGB frames of the video, while the other learns temporal consistency information from optical flow fields generated from consecutive frames. Unlike most approaches where pre-training is performed on a generic large corpus of images, we show that by pre-training on smaller face-related datasets, namely Celeb-A (for the spatial learning component) and YouTube Faces (for the temporal learning component), strong results can be obtained. We perform various experiments to evaluate the performance of our method on commonly used datasets namely FaceForensics++ (Low Quality and High Quality, along with a new highly compressed version named Very Low Quality) and Celeb-DFv2 datasets. Our experiments show that our method sets a new state-of-the-art on FaceForensics++ (LQ, HQ, and VLQ), and obtains competitive results on Celeb-DFv2. Moreover, our method outperforms other methods in the area in a cross-dataset setup where we fine-tune our model on FaceForensics++ and test on CelebDFv2, pointing to its strong cross-dataset generalization ability.

1 Introduction

Facial forgery detection, also known as deepfake detection, is a rapidly growing field with important real-world applications [44]. With the recent explosion in the success of sophisticated deep generative models [20, 48, 34], it has become increasingly easy to generate highly realistic fake images and videos (see Figure 1 for an example of a real image along with four manipulated versions). The advancements in artificial intelligence have made it possible to create deepfakes that are nearly indistinguishable from genuine content, making the detection process even more challenging. This has led to a growing concern about the potential for malicious actors to use these tools for nefarious purposes, such as spreading disinformation, manipulating public opinion, or even causing social unrest. Given the potential risks associated with deepfakes, the importance of developing effective detection methods cannot be overstated [28, 31, 35].

The field of deepfake detection has seen considerable progress in recent years with a number of sophisticated techniques being proposed in the area [50, 33]. However, in the context of detecting manipulated content in videos, many existing methods primarily focus on spatial features extracted from individual frames [5]. This approach can lead to the overlooking of temporal dynamics that evolve throughout video sequences. This strategy can result in limitations, as temporal artifacts such as flickering and motion discontinuities, are common indicators of deepfake manipulation. Furthermore, sophisticated deepfakes may exhibit subtle spatial inconsistencies that manifest over time, necessitating an integrated analysis of both spatial and temporal information. Moreover, we hypothesize that capturing subtle *spatiotemporal* inconsistencies that are often caused by different deepfake generation methods, could significantly enhance performance by learning representations that generalize to unseen forgery methods, which is often a challenging problem in this area.

In response to the challenges mentioned above, in this paper, we present a novel approach to deepfake detection that consists of two distinct components. One component learns spatial information from individual RGB frames of the

Figure 1: A facial image along with the four manipulated versions from the FaceForensics++ dataset.

videos, while the second component leverages optical flow fields to learn temporal consistency across the video. Both components utilize vision transformers [13], which we train in two steps. First, inspired by [16] we pre-train the models in an autoencoding setup using a self-supervised reconstruction scheme. Second, we discard the reconstruction decoder and add a new classification head to each encoder, where they are fine-tuned for deepfake detection, followed by score-level fusion of the results. We pre-train the spatial learning and temporal consistency learning encoders with CelebFaces-Attributes (Celeb-A) [26] and YouTube Faces [47] datasets respectively. For downstream deepfake detection, we evaluate our approach on the FaceForensics++ (High Quality) and FaceForensics++ (Low Quality) datasets [36] which employ compression factors of 23% and 40%, respectively, in addition to the CelebDFv2 dataset [24]. Additionally, we synthesize a more challenging variation of the FaceForensics++ dataset, which we call FaceForensics++ (Very Low Quality) by compressing the data with a rate of 65%. This dataset is then used to further evaluate our approach against existing techniques in the presence of extreme compression artifacts. Experimental results demonstrate that our method achieves state-of-the-art performance on FaceForensics++ (LQ and HQ) datasets and highlight the efficacy of our approach in detecting deepfakes across diverse compression levels. Ablation studies demonstrate the importance of different components of our method. Lastly, state-of-the-art results when fine-tuning our model on FaceForensics++ and testing it on CelebDFv2 (cross-dataset evaluation) demonstrates the strong generalization of our method.

Our contributions in this paper are summarized as follows: (**1**) We propose a new approach for effective facial forgery detection. Our method uses a vision autoencoding transformer and is pre-trained in a self-supervised masked reconstruction setup. Our solution consists of two main components which learn spatial (RGB) and temporal consistency information (optical flow fields) separately. (**2**) We leverage relatively small datasets, namely Celeb-A and YouTube Faces, for pretraining our transformers, and achieve state-of-the-art results in the downstream task of deepfake detection on FaceForensics++ (LQ, HQ, and VLQ) datasets and competitive results on Celeb-DFv2.

2 Related Work

Deepfake detection has traditionally been addressed as a binary classification task [33], where the objective is to discern between authentic and manipulated media. The application of deep learning models, particularly convolutional neural networks (CNNs), has been central to achieving this objective [8, 10, 1]. Authors of FaceForensics++ dataset [36] used Xception network, which was one of the best-performing architectures at the time, to perform deepfake detection via transfer learning [8].

Researchers proposed a method in [10] that utilizes residual-based descriptors in the form of a constrained CNN for image forgery detection. This approach aims to capture and analyze the residual noise present in manipulated images, which can be a strong indicator of forgery.

In contrast, another method introduced a deep learning approach that focuses on the mesoscopic properties of images [1]. Within the context of image analysis and deep learning, 'mesoscopic' refers to properties or features that fall between the small scale (microscopic) and the large scale (macroscopic). By concentrating on mesoscopic features, the model can capture subtle artifacts and inconsistencies in manipulated images, potentially making it more effective in detecting forgeries.

Attention mechanisms, as introduced in [46], have been combined with CNNs in various works [54, 40] to enhance interpretability and facilitate the identification of manipulated regions. These attention-based models generate attention maps that highlight regions contributing significantly to the detection decision. For example, the study in [54] utilized attention maps generated by deep semantic features to outline crucial regions that contributed to the classification result. These attention maps guide the aggregation of low-level textural features and high-level semantic features, which helps to capture more subtle artifacts in the image. Furthermore, a new attention mechanism was proposed in [40] that calculates the self-information from the input feature map and outputs a discriminative attention map. This attention

map emphasizes regions that contribute significantly to the detection decision, enhancing the model's ability to identify manipulated areas.

Various studies have utilized frequency analysis to detect inconsistencies that arise during deepfake creation, as evidenced in [25, 7]. In [25], the researchers employed the phase spectrum for forged face image detection, showing that CNNs can identify additional implicit phase spectrum features that are advantageous in detecting face forgeries. Concurrently, the study in [7] developed a multi-scale patch similarity module to specifically model second-order relationships between distinct local regions, forming a similarity pattern through pairwise cosine measurements. This pattern distinguishes real from forged regions by recognizing differences such as irregular textures and high-frequency noise.

Self-supervised learning (SSL) has been explored to address the issue of limited labeled data for deepfake detection [6, 55, 53, 21, 49]. For instance, self-supervised learning was employed in [6] with an auxiliary task specifically designed for deepfake detection, using a synthesizer and adversarial training framework to dynamically generate forgeries. This approach enriches diversity and strengthens sensitivity to produce strong results. In the method proposed in [55], mouth motion representations were learned by encouraging close-paired video and audio representations, while keeping unpaired ones diverse. The study in [53] proposed a decoupling strategy to separate facial authenticity and compression relevance, implemented through a joint self-supervised learning approach using compression ratios as self-supervised signals. Another study utilized a multi-modal backbone trained in a self-supervised manner and adapted it to the video deepfake domain [21]. These self-supervised models leverage unlabeled data to learn useful representations for detection tasks. Contrastive learning is another common pre-text learning approach often considered for deepfake detection [15, 49]. In the study by [15], two different transformed versions of a face image were generated using two distinct transformations. The agreement between these transformed images is maximized after they are passed through an encoder network and a projection head network, effectively training the model without supervision signals. On the other hand, another study employed supervised contrastive learning to learn common features between instances of the same class, while distinguishing between samples from different classes [49].

In recent studies, researchers have sought to combine multiple modalities, such as visual, audio, and temporal information, to improve detection performance in deepfake detection tasks [25, 4, 18]. These multi-modal approaches provide a comprehensive view of the media, making them more robust against limitations specific to individual modalities [19]. While frequency-based modalities have been employed in multi-modal deepfake detection solutions [25], we believe that optical flow has the potential to serve as an effective alternative source of information. Furthermore, we aim to address generalizability and robustness in representation learning by employing the MAE framework based on vision transformers.

3 Proposed Methodology

3.1 Overview

Our proposed approach titled Masked Autoencoding Spatiotemporal Deepfake Transformer (MASDT) consists of two components: spatial learning and temporal consistency learning. The spatial learning component has the objective of learning robust spatial features from the RGB images, while temporal consistency learning aims to extract temporal features from optical flow fields derived from the input images. We fuse the classification outputs derived from the spatial and temporal consistency learning components. Both these components follow a self-supervised autoencoding approach in a *two-step* process.

The first step involves a self-supervised pre-training strategy which involves both of the MASDT components in a data reconstruction task. We discuss this strategy in section 3.2. The second step is the downstream task of deepfake detection, wherein we re-purpose components trained in the previous step to perform the classification of deepfake data through a model fine-tuning process, followed by fusion of information from both components (spatial learning and temporal consistency learning). This is discussed in detail in Section 3.3. Before we discuss each of the two steps, we discuss the optical flow field generation strategy in Section 3.1.1. A general scheme of the MASDT strategy is presented in Figure 2.

3.1.1 Optical flow field estimation

We utilize a CNN model named PWC-Net for generating optical flow fields [39]. Let the model for estimating optical flow be F_θ, and two consecutive frames be f_t and f_{t+1}. Accordingly, the estimated optical flow Φ_t can be denoted by:

$$\Phi_t = F_\theta(f_t, f_{t+1}), \tag{1}$$

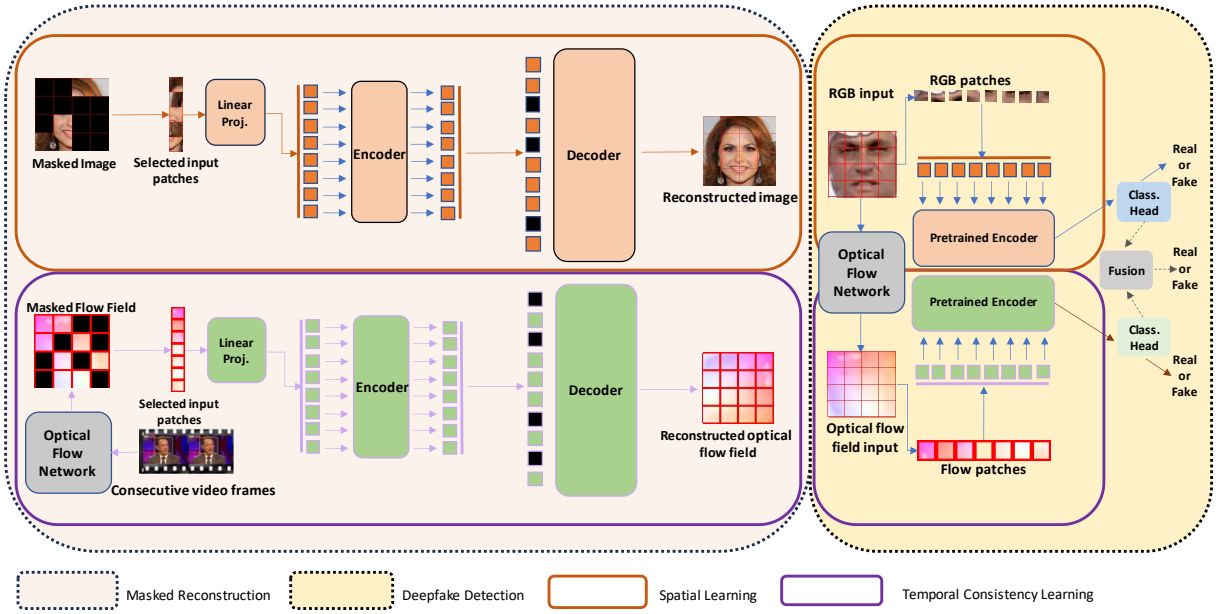

Figure 2: An overview of our method (MASDT), which includes the masked facial reconstruction and deepfake detection steps.

where Φ_t is a 3-channel optical flow matrix of size $H \times W \times 3$ representing the flow field between the consecutive frames.

3.2 Masked Facial Reconstruction

The first step of our approach utilizes a masked self-supervised auto-encoder which learns to reconstruct original facial images, given partial observations [16]. This auto-encoder reconstruction pipeline consists of two blocks: a reconstruction encoder, which captures a latent representation from the visible portions of each image, and a reconstruction decoder that aims to reconstruct the masked sections of the image using this latent representation. In this procedure, the encoder is trained to extract robust spatial features from masked facial images, eliminating noise and redundancy while transforming the reconstruction task into a challenging process that requires generalizing features to represent a small subset of available data [17]. Consequently, by masking portions of the facial image using random spatial pixels or patches, we can avoid a potential location bias toward image reconstruction, which can be critical for the detection of deepfake images.

The goal of the decoder is to use the features obtained from the latent space by the encoder to reconstruct the masked information from the original facial image. We train this reconstruction encoder-decoder pair using a simple mean squared error (MSE) reconstruction loss $\mathcal{L}_r$:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{2}$$

where N represents the number of sampled patches, and $\hat{y}_i$ and y_i are the ith output and expected ith output, respectively.

We perform the above masked reconstruction task for both the components independently where we employ the encoder-decoder pairs for reconstructing RGB images y and optical flow fields Φ for spatial learning and temporal consistency learning respectively, which is also referred to as pre-training in self-supervised learning literature. This prepares the encoders for the fine-tuning step mentioned in the next section.

3.3 Deepfake Detection

The second step of MASDT is aimed at the supervised training for the classification of deepfake images. For this purpose, we employ the encoders that learned to extract robust representations in the reconstruction pipeline. Thus, to

perform binary classification, a classification head consisting of a simple MLP is attached to each of the pre-trained encoders.

We adopt a dual-encoder setup for the fine-tuning process, utilizing the spatial learning encoder θ_s and the temporal consistency learning encoder θ_t. These encoders were previously trained in the initial step of our proposed solution. In the process of fine-tuning for a binary classification task, we employ a binary cross-entropy loss, denoted as $\mathcal{L}_b$. The formula for this loss is as follows:

$$\mathcal{L}_b = -\frac{1}{M}\sum_{j=1}^{M}\left[o_j \log(\hat{o}_j) + (1-o_j)\log(1-\hat{o}_j)\right], \tag{3}$$

Here, $\hat{o}_j$ is the predicted output from the network, o_j represents the actual or target class (either 0 or 1), and M denotes the total count of samples in the batch.

3.3.1 Dual Modality Fusion

To further harness the strengths of both θ_s and θ_t, we use a simple fusion mechanism. This method aims to exploit the complementary information that each encoder provides, thereby improving the overall classification performance. The fusion process begins with the individual outputs from θ_s and θ_t, denoted as $\hat{o}_s$ and $\hat{o}_t$ respectively, which are then combined to create a fused score-level prediction, $\hat{o}_f$. Mathematically, this can be expressed as:

$$\hat{o}_f = \alpha \cdot \hat{o}_s + (1-\alpha) \cdot \hat{o}_t, \tag{4}$$

where α is a fusion weight that determines the contribution of each encoder to the final output.

4 Experiments

In this section, we present the specifics and details of our method and experiments, describe the datasets used, and discuss the ablation studies conducted to validate the impact of different components of our proposed solution.

4.1 Implementation Details

In this section, we outline the implementation details of our deepfake detection method, which incorporates both RGB and optical flow modalities. Our experiments are conducted using the PyTorch framework [29] on 4 Nvidia A100 GPUs, each with 40 GB of vRAM. We generate optical flow fields using the PWC-Net present in the MMFlow toolbox [9].

Our method's performance is evaluated using the top-1 accuracy, which denotes the percentage of correctly classified deepfake and real videos out of the total number of videos in the test set. This metric is widely used in deepfake detection tasks as it provides a clear indication of a model's ability to distinguish between real and fake videos. Accuracy and area under the curve (AUC) are presented as the metrics for our experiments, following other publications in the area.

For evaluation purposes, we use the FaceForensics++ (LQ and HQ) and CelebDFv2 datasets (the details of these datasets are presented in the next Section) and divide them into training, validation, and test sets, ensuring an even distribution of deepfake and real videos across all sets following the instructions provided in the original dataset papers [36, 24]. Data augmentation techniques such as random cropping, horizontal flipping, color jittering, and MixUp augmentation, are employed to improve our model's robustness to input data variations. MixUp augmentation [52] involves generating new training samples by taking linear combinations of input data and their corresponding labels, which encourages the model to learn smooth and robust features. In addition to MixUp, the model employs CutMix [51] data augmentation technique with default settings (alpha set to 0, probability set to 1, and switch probability set to 0.5). Label smoothing is applied with a smoothing factor of 0.1. A drop path rate of 0.1 is used for stochastic depth regularization.

Input images are resized to 224×224, with patches of 16×16. We observe that a masking ratio of 90% is optimal for pre-training. We use the transformer architecture [13] with a default Vit-B configuration as our model. The model is trained using the AdamW optimizer, with a weight decay of 0.05, a base learning rate of 5×10^{-4}, and layer decay of 0.8. The learning rate is scaled according to an effective batch size of 64. We train the model for 300 epochs, using a gradient accumulation of 1 iteration. We utilize a distributed training approach with distributed evaluation. The CUDA benchmark is enabled, and the model is trained on available CUDA devices. For fine-tuning, the model is initialized with our pre-trained weights from the first step (self-supervised pre-training), and position embeddings are interpolated accordingly.

4.2 Datasets

We use the FaceForensics++ (LQ), FaceForensics++ (HQ) [36], Celeb-DFv2 [24], Celeb-A [26], and YouTube Faces [47] datasets. The first three datasets are employed for evaluating our proposed method, while the latter two are utilized for pre-training only. Below, we provide a detailed description of each dataset:

FaceForensics++ (LQ) [36] simulates various scenarios where manipulated videos appear in compressed formats. With a 40% compression factor using the H.264 video compression standard, the LQ version introduces artifacts that may be present in real-world cases. This dataset challenges researchers to develop techniques capable of detecting manipulations even when the video quality is degraded due to compression.

FaceForensics++ (HQ) [36] maintains a higher quality (compression factor of 23%) compared to the LQ version, enabling researchers to study deepfakes and other manipulations with greater detail and less information loss due to compression. Both FaceForensics++ versions contain over 1000 original videos, with manipulated videos created using various methods, such as DeepFakes [11], FaceSwap [22], Face2Face [43], and NeuralTextures [42]. These datasets cover a wide range of manipulation methods, allowing researchers to test their detection algorithms on diverse types of deepfakes.

In order to further push our method to the limit and challenge its detection ability in the presence of significant compression artifacts, we create an even more compressed version in comparison to FaceForensics++ (LQ), which we call **FaceForensics++ (VLQ)** where VLQ stands for very low quality. To generate this variant of the dataset, we take the original non-compressed videos of FaceForensics++ and compress them by a compression factor of 65%, which we will also use in our experiments besides the datasets with two standard compression ratios. For this purpose, we use the FFMPEG framework [45].

Celeb-DFv2 [24] includes 590 original videos collected from YouTube, featuring subjects of varying ages, ethnicities, and genders, as well as 5639 corresponding DeepFake videos. The Celeb-DF dataset's average video length is 13 seconds, and all videos have a standard 30 FPS frame rate.

Celeb-A (CelebFaces-Attributes) [26] is a large-scale collection of over 200,000 celebrity images, with 40 attribute labels annotated for each image. The dataset comprises diverse subjects and captures various facial expressions, poses, and lighting conditions.

YouTube Faces [47] is a comprehensive collection of videos from YouTube focusing on individuals' faces. It contains over 3,000 annotated videos of 1,595 people, offering diverse subjects with different ethnicities, ages, and genders. Each video in the dataset is labeled with the corresponding subjects' identities, and is often used for face recognition and verification tasks. It captures various poses, expressions, illuminations, and occlusions.

4.3 Pre-training Strategy

For pre-training the RGB modality in our proposed method, we utilize the Celeb-A dataset instead of the typically used ImageNet [12]. Celeb-A is considerably smaller than ImageNet, as Celeb-A contains 200,000 images whereas ImageNet contains over 14 million images. This reduced size allows for faster pre-training and lower computational requirements, making the process more efficient and accessible to a wider range of researchers and practitioners. Celeb-A is specifically tailored for facial tasks, consisting exclusively of human face images. In contrast, ImageNet covers many object categories and may not be as well-suited and efficient for facial analysis. By pre-training our model on Celeb-A, we ensure that the initial features learned by the model are more relevant to facial structures, expressions, and attributes, which can ultimately contribute to a more effective deepfake detection system.

For pre-training the optical flow modality in our method, we utilize the YouTube Faces dataset. This dataset provides video data, essential for optical flow calculation. Naturally, datasets of images such as ImageNet and Celeb-A cannot be used for optical flow generation. Moreover, the YouTube Faces dataset is specifically designed for facial analysis tasks as it consists exclusively of human face videos. By pre-training our model on this dataset, we ensure that the initial features learned by the temporal consistency encoder can better capture information such as facial structures, expressions, and attributes, ultimately contributing to a more effective deepfake detection system.

4.4 Results

In this section, we present the outcome of our experiments, which assess the performance of the proposed method for deepfake detection on the FaceForensics++ and Celeb-DFv2 datasets. Our evaluation concentrates on the effectiveness of integrating both RGB and optical flow modalities, as well as the impact of pre-training on the Celeb-A and YouTube Faces datasets. By contrasting our approach with existing methods and baseline models, we aim to evaluate the benefits of our technique in accurately identifying deepfakes under a range of conditions.

Table 1: Quantitative results for ACC and AUC on the FaceForensics++ dataset with both quality settings (LQ and HQ). The results are arranged in ascending order on the basis of ACC (LQ).

Methods	**ACC (HQ)**	**AUC (HQ)**	**ACC (LQ)**	**AUC (LQ)**
Steg. Features [14]	70.97%	-	55.98%	-
LD-CNN [10]	78.45%	-	58.69%	-
CP-CNN [32]	79.08%	-	61.18%	-
Face X-ray [23]	-	87.40%	-	61.60%
C-Conv [3]	82.97%	-	66.84%	-
MesoNet [1]	83.10%	-	70.47%	-
Xception [37]	95.73%	-	86.86%	-
Two-branch RN [27]	96.43%	88.70%	86.34%	86.59%
Self Info. Att. [40]	97.64%	99.35%	90.23%	93.45%
F3-Net [30]	97.52%	98.10%	90.43%	93.30%
E2E Learning [5]	97.06%	99.32%	91.03%	95.02%
Local Relation Learning [7]	97.59%	99.46%	91.47%	95.21%
Ours	**98.19%**	**99.67%**	**97.79%**	**98.45%**

Table 2: Quantitative results (ACC) on the FaceForensics++ (LQ) dataset with four manipulation methods: DeepFakes (DF), Face2Face (FF), FaceSwap (FS), and NeuralTextures (NT).

Methods	**DF** [11]	**FF** [43]	**FS** [22]	**NT** [42]
Steg. Features [14]	67.00%	48.00%	49.00%	56.00%
LD-CNN [10]	75.00%	56.00%	51.00%	62.00%
C-Conv [3]	87.00%	82.00%	74.00%	74.00%
CP-CNN [32]	80.00%	62.00%	59.00%	59.00%
MesoNet [1]	90.00%	83.00%	83.00%	75.00%
Xception [37]	96.01%	93.29%	94.71%	79.14%
F3-Net [30]	97.97%	95.32%	96.53%	83.32%
Local Relation Learning [7]	**98.84** %	95.53%	97.53%	**89.31%**
Ours	97.84%	**96.27%**	**97.89%**	78.23%

Table 3: Quantitative results in terms of ACC and AUC on the Celeb-DFv2 dataset.

Methods	**ACC**	**AUC**
F3-Net [30]	95.95%	98.93%
Xception [37]	97.90%	99.73%
E2E Learning [5]	**98.59%**	**99.94%**
Ours	98.00%	98.90%

In Table 1 we present the top-1 accuracy and AUC scores of our proposed method compared to the current state-of-the-art approaches. The table presents the quantitative results for various deepfake detection techniques available in the FaceForensics++ dataset with both high and low quality settings. It can be observed that our proposed method achieves the highest accuracy and AUC scores in both quality settings, surpassing the prior works and setting a new state-of-the-art.

In our experiments, we assess the performance of different deepfake generation methods in the FaceForensics++ (LQ) dataset, comprising four distinct techniques: DeepFakes (DF) [11], Face2Face (FF) [43], FaceSwap (FS) [22], and NeuralTextures (NT) [42], as illustrated in Table 2. In this table, we present a breakdown of the performance of our method and others across these four deepfake generation methods, and compare the accuracy with other state-of-the-art approaches. The results indicate that our method achieves strong results across all four manipulation techniques, particularly in the FF and FS methods, and generates competitive results for the other two. These findings demonstrate the effectiveness of our approach in detecting manipulated face images across different forgery approaches.

Next, we evaluate the performance of our method compared to other recent methods on the Celeb-DFv2 dataset and present the performance in Table 3. It can be observed that our method achieves results competitive to the current state-of-the-art [5].

To further explore the generalization capability of our model, we follow the cross-dataset scheme presented in [5], [37], and [7]. In this experiment, we train the model on the FaceForensics++ datasets and test its performance on the Celeb-DFv2 dataset. We present the results in Table 4, where we observe that our method outperforms prior works in

Table 4: Cross-dataset evaluation (AUC) by training on FaceForensics++ (LQ) and testing on the Celeb-DFv2 dataset.

Methods	AUC
Xception [37]	36.19%
E2E Learning [5]	68.71%
Local Relation Learning [7]	78.26%
Ours	**80.21%**

Table 5: Quantitative results on FaceForensics++ (VLQ) dataset which is constructed by applying a 65% compression ratio.

Methods	ACC
DCL [41]	65.20%
E2E Learning [5]	78.20%
Ours	**79.70%**

the area, indicating strong generalization ability in detecting deepfakes even when training is done on a different dataset and likely constitutes a different distribution (out-of-distribution).

To further push our approach to the limit, we explore its performance on the VLQ version of the FaceForensics++ dataset which we constructed for the first time by applying a 65% compression ratio (see Section 4.2). We also use this dataset on two leading methods, namely DCL [41] and E2E Reconstruction Learning [5]. The results are presented in Table 5 where we observe that our method outperforms both other solutions, highlighting the efficiency and resilience of our approach in detecting deepfakes, even in the presence of highly compressed data.

Lastly, we utilize Grad-CAM [38] visualization on our model and similar performing methods to demonstrate and investigate the attention patterns of each method. Grad-CAM is capable of pinpointing the areas that the network applies more attention to, and thus deems important. We present a sample image in Figure 3, where the red areas highlight parts of the image which are more salient for the models. We observe that our model considers broader areas of the face image as important toward detection of whether the input is a deepfake image or not. This is a noteworthy observation as it indicates that the proposed method is capable of capturing a more comprehensive set of features and artifacts, which might be overlooked by the other models. This ability to focus on multiple areas simultaneously could enable the proposed method to better discern subtle inconsistencies and artifacts that are characteristic of deepfakes or manipulated images. In contrast, the other two models, with their more concentrated attention patterns, may be less effective in capturing the full extent of these subtle cues, which might result in lower overall performance in detecting such forgeries. Another interesting pattern which can be observed is that prior methods seem to focus on select areas, namely the left eye and to some extent the right ear. However, in addition to these regions, our method considers the nose and mouth regions, which are critical areas for authentic face images.

4.5 Ablation Studies

In this section, we investigate the contributions of different components of our method toward facial forgery detection. As the first step, we remove the temporal consistency encoder and present the results in Tables 6, 7, and 8, for FaceForesnsics++ (LQ), FaceForesnsics++ (HQ), and Celeb-DFv2, respectively. When comparing these results to the performance of our original method (also presented in each table), we observe that removing the temporal consistency encoder results in performance drops of 1.2% to 2.9%. This indicates the importance of learning additional temporal information through optical flow which may be difficult for the model to learn without explicit supervision.

Next, we examine the impact using simple score-level fusion in our model. To this end, we adopt two strategies instead. First, we use the joint learning approach proposed in [2], where a single pre-trained encoder accepts patches from both the RGB and optical flow modalities simultaneously. Second, instead of score-level fusion, we use feature-level fusion immediately after the embeddings are obtained from the spatial and temporal consistency encoders. The results for both experiments are presented in Tables 6, 7, and 8, for the three datasets, respectively. We observe that while feature-level fusion achieves results closer to ours in comparison to joint learning, our method still obtains superior results to both these strategies.

Lastly we illustrate the Receiver Operating Characteristic (ROC) curves for our method (depicted in blue) and the three ablated variants discussed above, in Figure 4. These results are obtained on the FaceForesnsics++ (LQ), demonstrated in Table 6. We observe that the true positive rates are generally higher than the model variants across different false positive rate regions, except for the version where temporal consistency is not used, which shows comparable results

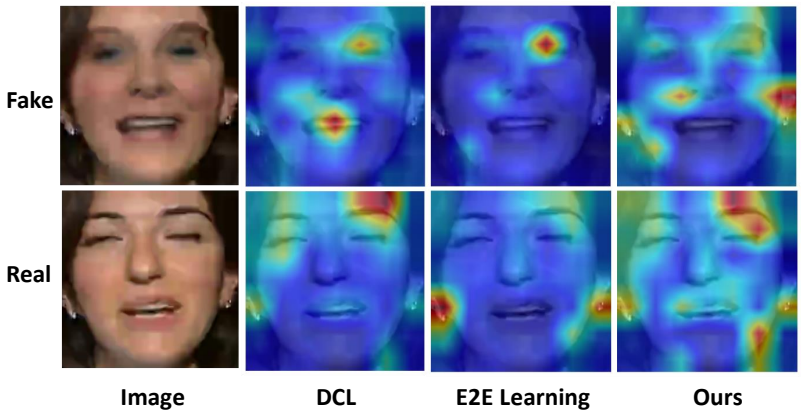

Figure 3: Comparison of Grad-CAM visualizations [38] for our method in comparison to two other recent works.

Table 6: Ablation experiments on FaceForensics++ (LQ). The ablated versions include the removal of the temporal consistency component, removal of score-level fusion and replacing it with MultiMAE joint learning [2], and removal of score-level fusion and replacing it with feature-level fusion.

Technique	ACC	AUC
Proposed	**97.79%**	**98.45%**
w/o temporal consistency	96.51%	97.03%
w/ joint learning [2]	95.02%	97.05%
w/ feature-level fusion	96.01%	97.10%

Table 7: Ablation experiments on FaceForensics++ (HQ). The ablated versions include the removal of the temporal consistency component, removal of score-level fusion and replacing it with MultiMAE joint learning [2], and removal of score-level fusion and replacing it with feature-level fusion.

Technique	ACC	AUC
Proposed	**98.19%**	**99.67%**
w/o temporal consistency	96.90%	97.35%
w/ joint learning [2]	95.81%	97.58%
w/ feature-level fusion	98.01%	99.09%

Table 8: Ablation experiments on CelebDFv2. The ablated versions include the removal of the temporal consistency component, removal of score-level fusion and replacing it with MultiMAE joint learning [2], and removal of score-level fusion and replacing it with feature-level fusion.

Technique	ACC	AUC
Proposed	**98.00%**	**98.90%**
w/o temporal consistency	95.08%	97.17%
w/ joint learning [2]	95.06%	96.55%
w/ feature-level fusion	96.81%	98.10%

in true positive rates for high false positive regions. This indicates that the temporal consistency component is highly effective in reducing the number of false alarms.

4.6 Limitations

We identify several limitations in our work. First, while the integration of temporal information through optical flow improves the detection performance of our method, it also increases the computational complexity of the system, potentially limiting its real-time applicability. Second, the proposed approach may not be robust to novel deepfake techniques or attacks targeting the identified limitations. Therefore, the effectiveness and generalizability of our proposed method will need to be validated further on new datasets and deepfake scenarios as they become available

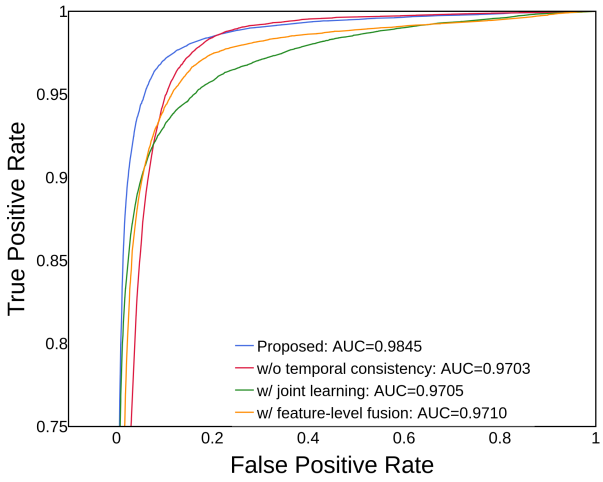

Figure 4: Receiver operating characteristic (ROC) curves for our proposed method (blue) and three ablations on the FaceForensics++ (LQ) dataset.

in the future. Lastly, we observe that the temporal consistency contributed mostly to the reduction of false positive detection. While this can indeed be valuable for practical applications, designing additional components to further enhance the true positive detection is also of critical importance.

5 Conclusion and Future Work

In this work, we introduce MASDT, a learning framework for enhanced deepfake detection. The proposed method, consists of two components, spatial and temporal consistency learning. The model follows a sequential two-step process. Initially, it employs a self-supervised pre-training strategy where both spatial learning and temporal consistency learning components engage in a data reconstruction task. Spatial learning makes use of a masked self-supervised auto-encoder to derive robust spatial features from partial facial images, while temporal consistency learning employs a similar auto-encoder to extract temporal features from partial optical flow fields. Subsequently, deepfake detection is executed through fine-tuning of the encoders of both learning components followed by simple score-level fusion. Various experiments on FaceForensics++ (LQ and HQ) and CelebDFv2 datasets demonstrate that our approach outperforms state-of-the-art methods by effectively learning spatial and temporal information, resulting in enhanced classification performance.

Several exciting avenues can be explored for future work. First, a lightweight version of our model, which could be achieved through distillation, could play a crucial role in extending the proposed method for real-time detection of facial forgeries in video streams for practical applications. Moreover, by integrating various modalities such as visual, audio, and text data and leveraging the strengths and complementary aspects of each modality, a unified framework could significantly enhance detection capabilities and overall performance through a holistic understanding of manipulated content.

6 Acknowledgements

This work was funded by Irdeto Canada Corporation and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. *arXiv preprint arXiv:2204.01678*, 2022.

[3] B. Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *IH&MMSec '16*, 2016.

[4] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization, 2022.

[5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.

[6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial examples: Towards good generalizations for deepfake detections. In *CVPR*, 2022.

[7] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1081–1088, May 2021.

[8] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[9] MMFlow Contributors. MMFlow: Openmmlab optical flow toolbox and benchmark. `https://github.com/open-mmlab/mmflow`, 2021.

[10] D. Cozzolino, G. Poggi, and L. Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, 2017.

[11] Deepfakes. Faceswap: Deepfakes Software. `https://github.com/deepfakes/faceswap/`, accessed 2023.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[14] J. Fridrich and Jan Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7:868–882, 2012.

[15] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021.

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[17] Damian Ibanez, Ruben Fernandez-Beltran, Filiberto Pla, and Naoto Yokoya. Masked auto-encoding spectral–spatial transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[18] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing*, 136:110124, 2023.

[19] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[20] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

[21] Gil Knafo and Ohad Fried. Fakeout: Leveraging out-of-domain self-supervision for multi-modal video deepfake detection, 2022.

[22] Marek Kowalski. FaceSwap: Deep Learning for Face Swapping. `https://github.com/MarekKowalski/FaceSwap`, accessed 2023.

[23] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and B. Guo. Face x-ray for more general face forgery detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5000–5009, 2020.

[24] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020.

[25] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[27] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch Recurrent Network for Isolating Deepfakes in Videos. *arXiv e-prints*, page arXiv:2008.03412, Aug. 2020.

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, dec 2021.

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[30] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. *ArXiv*, abs/2007.09355, 2020.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Nicolas Rahmouni, Vincent Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2017.

[33] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE Access*, 2022.

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[37] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, Justus Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 111–127. Springer, 2022.

[41] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022.

[42] Justus Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv: Computer Vision and Pattern Recognition*, 2019.

[43] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[44] Aniruddha Tiwari, Rushit Dave, and Mounika Vanamala. Leveraging deep learning approaches for deepfake detection: A review, 2023.

[45] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[47] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 529–534. IEEE, 2011.

[48] Jungang Xu, Hui Li, and Shilong Zhou. An overview of deep generative models. *IETE Technical Review*, 32(2):131–139, 2015.

[49] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389, 2022.

[50] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *IET Biometrics*, 10(6):607–624, 2021.

[51] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[52] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018.

[53] Jian Zhang, Jiangqun Ni, and Hao Xie. Deepfake videos detection using self-supervised decoupling network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.

[54] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[55] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. Self-supervised transformer for deepfake detection, 2022.