# Sharp Multiple Instance Learning for DeepFake Video Detection

Xiaodan Li*
Alibaba Group, China

Yining Lang*
Alibaba Group, China

Yuefeng Chen†
Alibaba Group, China

Xiaofeng Mao
Alibaba Group, China

Yuan He
Alibaba Group, China

Shuhui Wang
Inst. of Comput. Tech., CAS, China

Hui Xue
Alibaba Group, China

Quan Lu
Alibaba Group, China

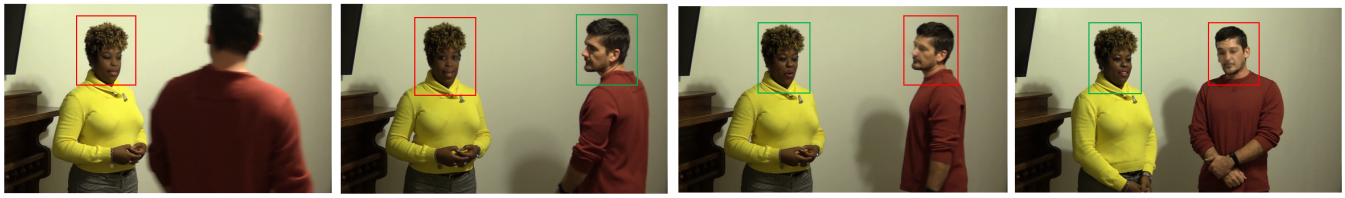


**Figure 1: Example of partially attacked DeepFake video. The green and red boxes represent real and fake faces respectively. This figure illustrates that not all faces in a fake video are manipulated. Real and fake faces may appear in the same frame. Face labels of one person in nearby frames may also be different.**

## ABSTRACT

With the rapid development of facial manipulation techniques, face forgery has received considerable attention in multimedia and computer vision community due to security concerns. Existing methods are mostly designed for single-frame detection trained with precise image-level labels or for video-level prediction by only modeling the inter-frame inconsistency, leaving potential high risks for DeepFake attackers. In this paper, we introduce a new problem of partial face attack in DeepFake video, where only video-level labels are provided but not all the faces in the fake videos are manipulated. We address this problem by multiple instance learning framework, treating faces and input video as instances and bag respectively. A sharp MIL (S-MIL) is proposed which builds direct mapping from instance embeddings to bag prediction, rather than from instance embeddings to instance prediction and then to bag prediction in traditional MIL. Theoretical analysis proves that the gradient vanishing in traditional MIL is relieved in S-MIL. To generate instances that can accurately incorporate the partially manipulated faces, spatial-temporal encoded instance is designed to fully model the intra-frame and inter-frame inconsistency, which further helps to promote the detection performance. We also construct a new dataset FFPMS for partially attacked DeepFake video detection, which can benefit the evaluation of different methods at both frame and video levels. Experiments on FFPMS and the widely used DFDC dataset verify that S-MIL is superior to other counterparts for partially attacked DeepFake video detection. In addition, S-MIL can also be adapted to traditional DeepFake image detection tasks and achieve state-of-the-art performance on single-frame datasets.

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision problem**.

## KEYWORDS

DeepFake; Multi-Instance Learning; Weighting; Temporal

**ACM Reference Format:**
Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. 2020. Sharp Multiple Instance Learning for DeepFake Video Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414034

## 1 INTRODUCTION

The rise of face manipulation techniques has fueled the advent of forgery face images and videos [9, 10, 47, 48]. As the swapped faces become more and more lifelike with advanced deep learning models such as Generative Adversarial Networks (GAN), these techniques, *e.g.*, DeepFake [9], are abused for malicious purposes, *e.g.*, face recognition attack or even political persecution. To enhance the content security, it is crucial to develop more advanced forgery detection technology.

*Both authors contributed equally to this research.
†Yuefeng Chen is the corresponding author (yuefeng.chenyf@alibaba-inc.com).


*MM '20, October 12–16, 2020, Seattle, WA, USA*

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414034

In previous study, the DeepFake video detection is either treated as a forgery image detection problem, or addressed as a video level fake prediction problem with video feature learning. Different from traditional setting, we address a new problem that exists more ubiquitously in real situations, *i.e.*, partial faces attack in a video where not all faces in a fake video are manipulated. This problem can be illustrated from two aspects, as shown in Fig. 1. Firstly, real and fake faces may occur simultaneously in one frame. Secondly, faces of one target person may be partially manipulated in one DeepFake video due to arbitrary content attack behaviors.

Existing DeepFake video detection methods can be divided into frame-based [12, 32, 41, 42, 46] and video-based methods [43]. Frame-based methods meet with three major difficulties when applied to fake video detection with partial faces attack. First, they are generally data hungry and require a remarkable number of labeled frames for training [20, 26]. Their performances may drop significantly when detecting partially attacked DeepFake videos since they simply treat video labels as frame labels. Besides, every sampled frame needs to be predicted for final prediction during inference, which is very computationally exhaustive. The video-level decision making based on frame-level representation is another key issue for frame-based methods. Using the maximum fake scores over faces or averaging them may lead to unexpected false alarm or false negative predictions. For video-based methods [43], more attention is paid on the temporal feature modeling, which is less effective for DeepFake detection under partial attack situation since it is also highly dependent on appearance modeling. Thus, traditional solutions leave potential high risks for DeepFake attackers.

Considering that videos are only given the video-level annotation in many practical situations, we address DeepFake video detection based on multiple instance learning (MIL) framework by treating faces and input video as instances and bag respectively. One bag consists of multiple instance and only bag labels are available. When one instance is positive, the corresponding bag will be labeled as positive. This setting is naturally suitable for DeepFake video detection task. However, in traditional MIL, instance predictions have to be learned based on bag labels, and bag prediction has to be made based upon instance predictions. Thus, gradient vanishing occurs constantly when the fake score of one instance in a bag is high. This issue leads to partial fake instance detection results [22]. Research endeavor has been dedicated to more careful network design [22, 51]. Nevertheless, theoretic insight on how the gradient vanishing problem can be alleviated has not been provided. The research challenge can be further analysed from the attacker side where DeepFake [9] tampers video on specific frames, which inevitably causes spatial-temporal inconsistency in nearby frames, evidenced as the content jitters. However, in traditional MIL, instances are generated by considering appearance cues only. This issue has to be tackled with a careful design of instance structure by considering both the rich spatial and temporal cues in videos.

In this paper, we propose Sharp Multiple Instance Learning (S-MIL) for DeepFake video detection. Different from traditional MIL where bag prediction has to be made upon instance prediction built on the multi-dimensional instance feature space, we apply Sigmoid on a weighted sum operation on a bag of instance embeddings to directly produce the bag prediction. Consequently, the loss on bag prediction can be directly propagated to instance embeddings, which can help to ease the gradient vanishing problem. Besides, the proposed S-MIL can still be explained in traditional MIL probabilistic framework, which benefits for comprehensive theoretic analysis. Accordingly, it can be theoretically derived that the gradient surface of the loss with respect to the instance prediction of S-MIL tends to be much sharper than traditional MIL, so the gradient vanishing problem during the back propagation from bag prediction to instance embeddings can be effectively alleviated. Based on the proposed S-MIL, we propose to produce multiple spatial-temporal instances to fully model the intra-frame and inter-frame inconsistency caused by independent frame attacking behavior to generate forgery faces.

To inspire research on video-level DeepFake detection under partial faces manipulation situations, we construct a new dataset named FaceForensics Plus with Mixing samples (FFPMS) for partially attacked DeepFake video detection. The dataset contains both frame-level and video-level annotations for more comprehensive evaluation while models are allowed to be trained using only the video-level annotations. Experiments on FFPMS dataset verify that S-MIL is superior to other competitors for partially attacked DeepFake video detection. Also, S-MIL can make a more accurate judgement on whether a video is fake or not than other methods on DFDC benchmark. S-MIL can also be adapted to traditional single-frame detection task and achieve state-of-the-art performance on the FF++ datasets [42], showing a promising result for face information protection.

The major contributions of our paper include:

- We introduce a new problem named partial faces attack in DeepFake video detection and propose S-MIL to address this problem. To the best of our knowledge, this is the first MIL-based solution for DeepFake video detection. Theoretical analysis shows that S-MIL can alleviate the gradient vanishing problem effectively.
- We design a new spatial-temporal instance to capture the inconsistency between faces, which can help to improve the accuracy of DeepFake detection.
- We propose FFPMS dataset with both frame-level and video-level annotations for video-based DeepFake face detection.
- We verify that our approach is superior to other counterparts for video face forgery detection and our S-MIL can also be adapted to traditional single-frame detection.

## 2 RELATED WORK

DeepFake detection obtains considerable attention due to the security concerns caused by the abuse of face swapping techniques. In this section, we briefly review typical detection methods and works related to multi-instance learning.

### 2.1 DeepFake Detection

According to the detection target, DeepFake detection can be divided into image-level detection and video-level detection. Early works [17, 18, 34] achieve the forgery face detection by handcraft features, as the face manipulation techniques are limited at that time. These traditional manipulation methods typically swap the face area based on the facial landmarks and then utilize some post-processing techniques to make the boundary of swapped face inconspicuous.

With the development of generative adversarial network (GAN) [19, 39], the forgery faces become more and more realistic. Therefore, some works [5, 8, 40, 49] begin to utilize deep networks to learn discriminative features or find manipulation traces for DeepFake detection. For instance, Rossler *et al.* [42] introduce a simple but effective Xception Net as a binary DeepFake image detector and Li *et al.* [27] as well as Lingzhi *et al.* [26] aim at the finding traces such as blending boundaries left by the DeepFake generation methods with deep neural networks. They are trained with frame-level labels and can achieve high accuracies for DeepFake image detection.

Compared to the forgery images, video-level forgeries [9, 47, 48] are more harmful, since they look more convincing with real audio. The above frame-based methods [1, 20, 26, 27] can also be used to detect video forgeries. The typical solution is choosing some frames randomly from the video and taking the max or average score as the final score of this video for classification (real or fake). However, the maximum operation may lead to a false alarm while averaging may cause false negative prediction because the scores on manipulated faces are overwhelmed by those of normal faces. Previous researchers have also explored to address DeepFake video detection as a video level prediction problem with video feature learning. Likewise, Sabir *et al.* [43] propose to use recurrent convolutional models [21] to exploit the temporal information in DeepFake videos. However, they are limited in some cases since face forgery detection is highly dependent on appearance modeling.

## 2.2 Multiple Instance Learning

For a typical Multiple-instance learning (MIL) method, the model receives a group of labeled bags which consist of many instances, rather than requiring instances that are labeled individually. MIL aims to learn a model that can predict the bag label [2, 23, 50].

In the early work of MIL, pre-computed features or hand-craft features are utilized to represent the instances [11, 31]. The boom of deep learning-based approaches further enhances the ability of multi-instance learning by a large margin [33, 50], especially for medical image scenarios [38, 45]. Generally speaking, MIL algorithms can be divided into two fold: instance-space paradigm and embedded-space paradigm [51]. Traditional MIL generally follows instance-space paradigm, which aggregates instances on output layer. Some works claim that it is inflexible and propose to do mean-pooling or max-pooling in embedded-space [16, 37, 54]. But both operators are non-trainable and susceptible to extreme values, which potentially limits their applicability.

The attention mechanism can relive the above weakness by focusing on key instances, which is widely used in recent works [4, 30, 52]. Inspired by the idea of lp-norm pooling [15], Zagoruyko *et al.* [53] propose a novel attention mechanism that requires a student model to mimic the feature map of an attention model (teacher model), which enhances the performance of the student model significantly. Recently, the attention mechanism has been also used in some MIL works [22, 35, 36]. For example, Ilse *et al.* [22] claim that traditional MIL may face the gradient vanishing problem and they propose an attention-based MIL network that achieves the attention weights training by a small neural network. However, theoretic insight on how the gradient vanishing problem can be alleviated has not been provided.

# 3 OUR APPROACH

In this section, we illustrate the proposed algorithm in detail. As shown in Fig. 2, our algorithm consists of three key components. First, face detection is conducted on the sampled frames in input videos. The extracted faces are then fed into a CNN to obtain features as instances. Second, spatial and temporal instances are extracted with corresponding encoding branches to form spatial-temporal bags with different temporal kernel sizes. These bags are used together to represent a video. Last, S-MIL is performed on these bags to get the final fake scores of all bags, which can derive a final fake score for the whole video.

## 3.1 Problem Formulation

Formally, for $N$ videos in an input batch, the goal for binary classification is to minimize the following objective function:

$$\mathcal{L} = -\sum_{i=1}^{N}(y_i \log p_i + (1-y_i)\log(1-p_i)), \quad p_i = F(W, x_i), \quad (1)$$

where $y_i$, $p_i$ are the label, prediction of the $i$-th video $x_i$. $F$ and $W$ are the prediction model and trainable parameters respectively.

For one single video, traditional DeepFake detection methods tend to convert the above video level prediction task to frame level task by training a supervised per-frame binary classifier [1, 26, 42]. During inference, averaging or maximizing is performed to get the final prediction of input video:

$$p_i = \begin{cases} \frac{1}{M}\sum_{j=1}^{M} p_i^j = \frac{1}{M}\sum_{j=1}^{M} F(W, x_i^j), & \text{Average} \\ \max\limits_{j\in[1,M]} p_i^j = \max\limits_{j\in[1,M]} F(W, x_i^j), & \text{Maximum} \end{cases} \quad (2)$$

where $p_i^j$ and $M$ are the fake score of the $j$-th frame and the total frame number in video $x_i$ respectively.

The above formulation will force every frame in fake videos to be predicted as fake. However, it is inaccurate since fake videos may only contain part of the manipulated target faces. Simply applying the video label to frame label may result in noises. Furthermore, when generating video-level label with face labels, using the maximum fake scores may lead to false alarms while averaging may lead to false negative predictions. Under MIL framework, we treat videos as an integral instead of analyzing frame by frame.

For a typical MIL work, the model receives labeled bags which consist of different number of instances, rather than requiring instances labeled individually. MIL classifier aims to predict the labels of testing bags. A bag is labeled as negative when all the instances in this bag are negative, while a bag is labeled as positive when there is at least one instance in this bag that is positive [50]. Formally, for a given bag with a bag label $y_i \in \{0, 1\}$, it consists of $M$ instances. For the $i$-th bag in MIL,

$$y_i = \begin{cases} 0, & \text{if } \sum_{j=1}^{M} y_i^j = 0 \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

We fit MIL into DeepFake video detection. Following the traditional MIL definition, the probability for a bag is:

$$p_i = 1 - \prod_{j=1}^{M}(1 - p_i^j). \quad (4)$$

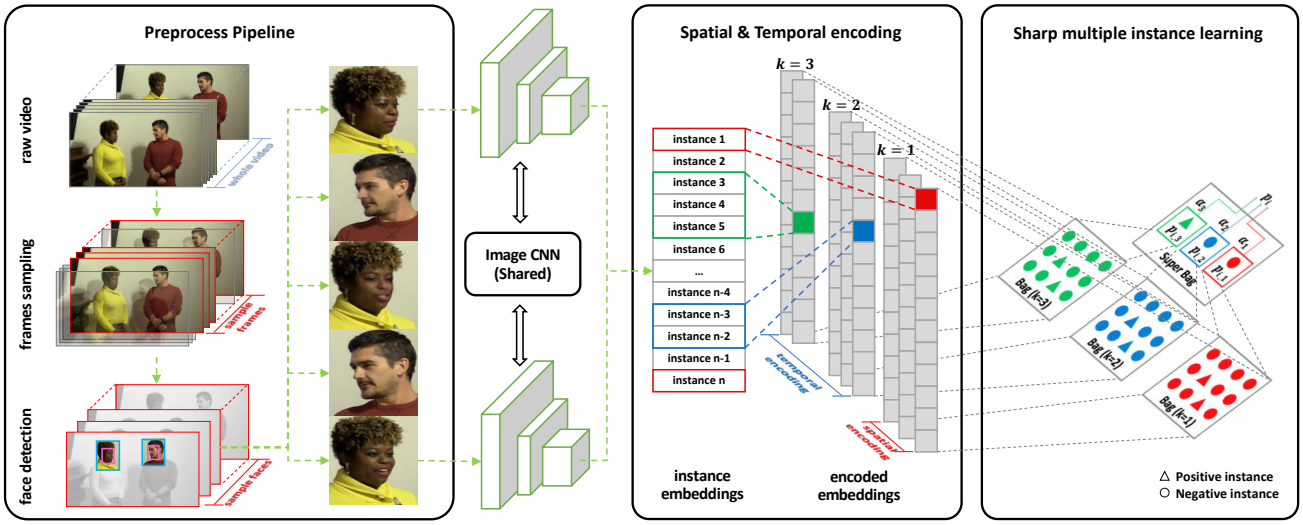


**Figure 2: Structure of the proposed algorithm which consists of three modules. Given a frame sequence from a video, our approach firstly detects faces and extracts individual feature maps by CNN. Then, the feature maps of the face sequences are encoded as spatial-temporal instances to get multiple spatial-temporal bags. After that, the resulted features of instances in different bags are integrated by the proposed sharp multi-instance learning method. Finally, S-MIL is performed on these bags to get the final video prediction (real or fake).**

Then the objective function of input bags can be denoted as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(1-\prod_{j=1}^{M}(1-p_i^j)) + (1-y_i)\log(\prod_{j=1}^{M}(1-p_i^j))). \quad (5)$$

For instances in positive bags, the gradient is:

$$\frac{\partial \mathcal{L}}{\partial p_i^j} = \frac{\partial \mathcal{L}}{\partial p_i}\frac{\partial p_i}{\partial p_i^j} = \frac{\partial \mathcal{L}}{\partial p_i}\Big(\prod_{k=1,k\neq j}^{M}(1-p_i^k)\Big). \quad (6)$$

When there is one instance predicted as positive, other positive instances will face the gradient vanishing issue. This will limit its applications such as judging one frame is fake or not.

## 3.2 Sharp Multi-Instance Learning

Generally speaking, MIL algorithms can be divided into two folds: instance-space paradigm and embedded-space paradigm [51]. Traditional MIL generally follows instance-space paradigm, which aggregates instances on label output layer. Since the individual labels are unknown, there is a bottleneck for the prediction accuracy, caused by the insufficient training of instance-level model [51]. To integrate instances in embedded-space, we propose to use a simple sum operator to fuse features of different instances. Let $H_i = [h_i^1, ..., h_i^M]$ be a bag of $M$ embeddings extracted by the backbone network, and the classifier layer be $W$. Denote the fake scores for the $i$-th video and the $j$-th frame in the $i$-th video as $p_i$ and $p_i^j$, respectively, then:

$$p_i = \frac{1}{1+e^{-W\sum_{j=1}^{M}(h_i^j)}}, \quad p_i^j = \frac{1}{1+e^{-Wh_i^j}}. \quad (7)$$

Actually, our proposed MIL can also be formulated in tradition MIL manner as in Eq. 8, which is helpful in analyzing its theoretic merits.

$$p_i = \frac{1}{1+e^{-W\sum_{j=1}^{M}(h_i^j)}} = \frac{1}{1+\prod_{j=1}^{M}(\frac{1}{p_i^j}-1)}. \quad (8)$$

**The sharp characteristic** With Eq. 8, it can be found that the proposed MIL can ease the gradient vanishing problem, which is proved in the supplementary. Besides, for an intuitive explanation, we set $M$ to 2 to show the gradient surface $\partial\mathcal{L}/\partial p_i^j$ in 3D space with respect to different $p_i^1$ and $p_i^2$. As shown in the supplementary, the gradient surface of the proposed MIL looks sharper than traditional MIL, resulting in a smaller area with vanished gradients. This validates that the proposed formulation in S-MIL can relieve the gradient vanishing problem.

**Weighting mechanism** Eq.8 simply treats instances equally without any focus, which tends to be sub-optimal. For example, if $[p_i^1, p_i^2, p_i^3] = [0.1, 0.1, 0.9]$, with Eq. 8, $p_i < 0.5$, which is not consistent of the definition of MIL.

Referring to the ideas of boosting [7] and focal loss [29], which are designed to emphasizing the learning on hard examples, we propose to "boost" the informative instances in bags. To the end, the proposed S-MIL is defined as:

$$p_i = \frac{1}{1+\prod_{j=1}^{M}(\frac{1}{p_i^j}-1)^{\alpha_i^j}}, \quad a_i^j = \frac{\exp\{\mathbf{w}^\top h_i^j\}}{\sum_{j=1}^{M}\exp\{\mathbf{w}^\top h_i^j\}}, \quad (9)$$

where $\mathbf{w}$ is the learnable parameter of a neural network. It can adjust the weights of different instances within one bag. The proposed

**Figure 3: Visualization for manipulated faces across frame sequences. Faces vary a lot along the temporal dimension.**

S-MIL can also be converted to embedded-space and it is a weighted sum of instances represented by low-dimensional embeddings $h_i^j$:

$$p_i = \frac{1}{1+\prod_{j=1}^{M}(\frac{1}{p_i^j}-1)^{\alpha_i^j}} = \frac{1}{1+e^{-W\sum_{j=1}^{M}(\alpha_i^j h_i^j)}}. \tag{10}$$

There are some advantages of the weighting mechanism. First, it is designed to enhance the weights of informative instances to extract discriminative representations for input bags. Second, for positive bags, the key instances are mostly positive ones, which should be assigned with high weights. Thus, it can help to interpret the final decision.

In similar research routine, attention-based multi-instance learning has been studied by previous work [22], which claims traditional MIL may face the gradient vanishing problem. Nevertheless, the attention-based approaches explore the instance-bag relation in a data-driven manner, and theoretic insight on how the gradient vanishing problem can be alleviated has not been provided. In comparison, our method can be naturally explained with traditional MIL probabilistic framework, and also has the advantage of attention-based models where the gradients can be back-propagated to instance embeddings without serious vanishing.

## 3.3 Spatial-temporal Instances

The S-MIL and traditional MIL are developed based on instances or frames that are independent to each other and they are organized at the spatial-level. It is true that faces are manipulated individually and they are irrelevant to each other. However, from another point of view, since they are made separately, the sequences along the temporal dimension may not be as smooth as real face sequences.

As described in Fig. 3, the forgery faces vary drastically along the temporal dimension. This important cue can be utilized for DeepFake video detection. To model inter-frame consistencies between faces extracted from nearby frames, based on the proposed S-MIL framework, we design a new spatial-temporal instance by adding a spatial-temporal branch with multiple temporal kernels. Inspired by the previous sentence classification method [24], we utilize the 1-d CNN to encoding input frames in the temporal level.

Let $Conv1d_{k,r}$, be a 1-d convolutional block, where $r$ and $k$ are the number and size of filters respectively. Feeding feature map $H_i = [h_i^1, h_i^2, ..., h_i^M]$ generated by the CNN backbone, after zero padding, into $Conv1d_{k,r}$, we produce a $M \times r$ feature map. Then ReLU activation function is performed on this feature map for non-linearity. More formally, we describe the above process as

$$c_i^k = ReLU(Conv1d_{k,r}(H_i)), \tag{11}$$

where $c_i^k$ is the spatial-temporal encoding bag for $H_i$ with kernel size $k$.

When we set the kernel size $k$ as 2, it allows two adjacent rows in $h_i^j$ to interact with each other. As the kernel size grows, more nearby rows are exploited simultaneously. We employ kernel size $k$ of 2 and 3 for temporal encoding and at the same time, keep $k = 1$ to retain the spatial encoding instances as well. The number of filters for each kernel size is set to 512.

Finally, several spatial-temporal encoded bags with different kernel sizes compose a super bag to represent a video, which is processed by S-MIL to get the final fake score for the input video. Denote $p_{i\_k}$ as the fake score of bag encoded with kernel size $k$, the final prediction of the input video is:

$$p_i = \text{S-MIL}(p_{i\_1}, p_{i\_2}, ..., p_{i\_k}), \quad \text{where} \quad p_{i\_k} = \text{S-MIL}(c_i^k). \tag{12}$$

## 3.4 Loss Function

Since we define the MIL DeepFake video detection as a binary classification task, we choose the Binary Cross Entropy (BCE) as our loss function. Given the positive probability $p_i$ of the $i$-th video, the loss is calculated as follows and the entire network is trained towards minimizing the following loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)). \tag{13}$$

# 4 EXPERIMENTS

In this section, we experimentally evaluate the proposed algorithm for DeepFake video detection and compare it with other counterparts in terms of partially attacked datasets evaluation, as well as evaluations on fully attacked benchmarks under video-level and frame-level settings. Also, we conduct an ablation study to explore the influence of proposed components as well as the number of sampled frames during the inference. For clarify, we denote S-MIL as the version of method with $k = 1$, and S-MIL-T as the version with $k = 1, 2, 3$ in subsequent experiments. We also provide a demo video in supplementary.

## 4.1 DeepFake Video Detection Datasets

**FaceForensics ++(FF++)** [42]: It is a recent released benchmark dataset and has been widely used for evaluation in different works [26, 42]. It consists of 1000 original videos and corresponding fake videos which are generated by four typical manipulation methods: DeepFakes (DF) [9], Face2Face (F2F) [48], FaceSwap (FS) [10], NeuralTexures (NT) [47]. It provides frame-level labels and nearly every frame and face in fake videos are manipulated.

**Celeb-DF** [28]: Celeb-DF dataset is composed of 5639 YouTube videos based on 59 celebrities. These videos are tampered by the DeepFake method and contain more than 2 million frames in total.

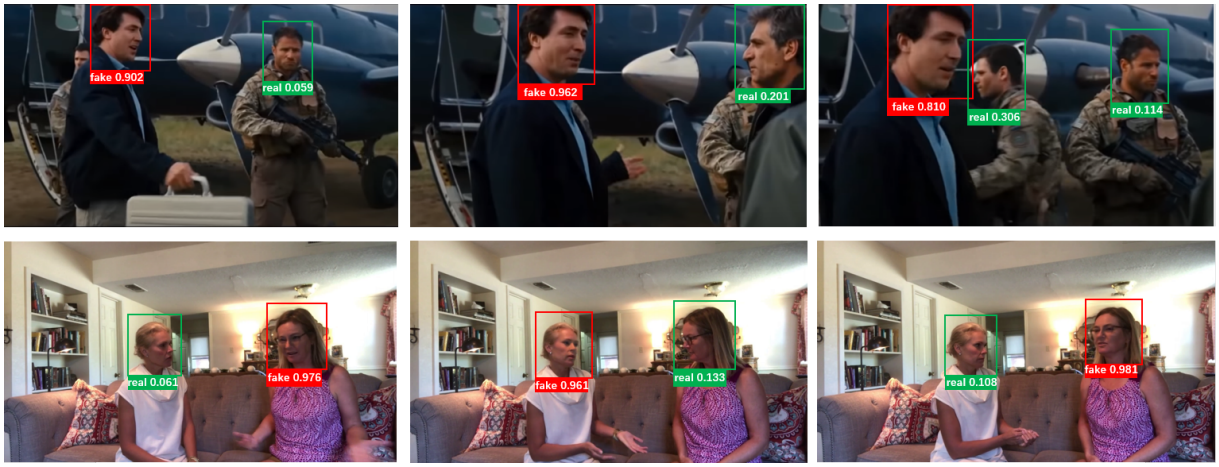


**Figure 4: Visualization results with S-MIL for DeepFake video detection of two partially attacked example frame sequences [3, 14]. Faces in green and red boxes are real and fake respectively and the scores associated with these boxes are the corresponding fake scores predicted by the proposed method.**

The quality of the tampered videos looks great with little notable visual artifacts. Nearly every frame in the video is manipulated.

**Deepfake Detection Challenge (DFDC)** [13]: DFDC is a preview dataset released by DeepFake detection challenge[1]. This dataset is generated by two kinds of unknown synthesis methods on 1131 original videos. Totally, it creates 4113 forgery videos based on humans of various ethnic, ages and genders. It only provides video-level labels and some fake videos contain many un-manipulated frames and faces.

**FaceForensics Plus with Mixing samples (FFPMS)**: Since existing datasets with frame labels have few video samples with mixed real/fake instances, which is not beneficial for analyzing performances in partial face attack scenario. We construct a new benchmark called FaceForensics Plus with Mixing samples (FFPMS).

Specially, we sample 20 frames from each video in the FF++ testing set whose compress rate is c40 and replace several fake frames with real frames from corresponding videos. The number of replaced fake frames ranges from 1 to 19. The resulting dataset has 14000 frames, containing four kinds of fakes (DF, F2F, FS, NT) and the original ones. With both frame-level labels and video-level labels, this dataset is designed to model fake videos with un-manipulated frames and faces, which is very common in Youtube videos but rare in existing datasets except for DFDC. However, DFDC does not contain frame-level labels.

## 4.2 Implementation Detail

We adopt XceptionNet as a default backbone network as other works for fairness [26, 42]. For spatial and temporal encoding, we set the kernel sizes to 1, 2, 3. The number of filters of each kernel size is set to 512. All frame-based and video-based models are trained with only 20 uniformly sampled frames in each video. For video-based models, we utilize a batch size of 32 for training and finish the training process after 30 epochs. During each training epoch, 8 frames are randomly extracted from the sampled 20 frames. Faces are detected and cropped from the resulting frames as the video input. We finetune on the ImageNet pretrained xception model with a learning rate of 0.0002, which is reduced to half every 5 epochs. During training, we adopt the Adam method [25] as the optimizer. Only random crop and horizontal flip are employed during training since we focus more on network designing. By default, 20 frames of each video are uniformly extracted as network inputs for efficiency during testing. More details about face detectors can be found in supplementary.

[1]https://deepfakedetectionchallenge.ai/

## 4.3 Evaluation on Partially Attacked Datasets

We conduct the experiment on DFDC, Celeb, and FFPMS datasets to evaluate the generalization ability of S-MIL. We compare the proposed S-MIL with frame-based model such as XceptionNet [42], D-FWA [27] and some video-based models such as LSTM [21] and I3D [6]. Since the FF++ dataset, from which FFPMS is extracted, contains few unmodified faces in fake training videos, we replace 25% of the whole fake frames with their corresponding real ones in the fake videos for training the S-MIL model. The experiment is evaluated by the average accuracy of fake and real testing videos.

For baseline XceptionNet, which is a frame-based method, there are two ways for converting it to a video-level model, *i.e.*, averaging fake score of each frame or treating the maximum fake score as the video-level prediction. For thorough analysis, we experiment on both cases. Table 1 summarizes the accuracy of different detectors with respect to each type of manipulated video in different datasets. As shown in Table 1, the proposed S-MIL performs better than both frame-based and video-based methods in most cases. Besides, either in max or average ways, XceptionNet performs much worse than the proposed S-MIL. When equipped with spatial-temporal encoding as in S-MIL-T, the proposed method gets a better performance. This validates the effectiveness of the proposed S-MIL and spatial-temporal instances. We also compare the accuracies under different noise levels on FFPMS dataset. As shown in Fig. 6, the proposed S-MIL-T performs well even when there are only 10%


**Figure 5: Visualization of a frame sequence and the corresponding heatmaps extracted with gradcam [44]. Faces in red bounding boxes are fake. The last row shows the weights of each face. Forgery faces have higher weights than real faces.**

fake faces in input instances, which outperforms frame-based XceptionNet and video-based LSTM significantly. Besides, as shown in Fig. 5, the weights get higher on fake faces, which demonstrates the interpretability of the proposed method.

| Methods | DFDC | Celeb | FFPMS | | | |
|---|---|---|---|---|---|---|
| | | | DF | F2F | FS | NT |
| XN-avg [42] | 0.8458 | **0.9944** | 0.8036 | 0.7714 | 0.8036 | 0.7000 |
| XN-max [42] | 0.7687 | 0.8989 | 0.8536 | 0.7821 | 0.8571 | 0.6571 |
| D-FWA [27] | 0.8511 | 0.9858 | 0.7964 | 0.7571 | 0.8036 | 0.7214 |
| LSTM [21] | 0.7902 | 0.9573 | 0.8500 | 0.7564 | 0.7750 | 0.7393 |
| I3D [6] | 0.8082 | 0.9923 | 0.6214 | 0.6857 | 0.7071 | 0.6679 |
| **S-MIL** | 0.8378 | 0.9923 | 0.9036 | 0.8107 | 0.8609 | **0.7857** |
| **S-MIL-T** | **0.8511** | 0.9884 | **0.9071** | **0.8250** | **0.8857** | 0.7535 |

**Table 1: Video-level accuracies of the proposed method and other state-of-the-art methods on DFDC, Celeb and FFPMS datasets. Results in bold text indicate the best results while results in pink and blue indicate the best and the second-best results appeared in proposed method.**

## 4.4 Evaluation on Fully Attacked Datasets

**Video-level benchmark results.** We test the proposed S-MIL on fully attacked datasets such as Celeb and FF++ to evaluate its generalization ability. For a comprehensive discussion, we compare the proposed method with a frame-level detector such as XceptionNet (XN), as well as video-level detectors, *i.e.*, LSTM [21] and I3D [6]. As shown in Table 1 and 2, the proposed multi-instance method outperforms frame-based detectors and video-based detectors in most cases.

**Frame-level benchmark results.** Since recent works focus more on frame-level detection, we also evaluate our method to validate the generalization in frame-level detection case.

| | Methods | CR | FaceForensics++ | | | |
|---|---|---|---|---|---|---|
| | | | DF | F2F | FS | NT |
| Video level accuracy | | | | | | |
| Fb | XN-avg [42] | c0 | 1.0000 | 0.9964 | 0.9964 | 0.9964 |
| Vb | LSTM [21] | | 1.0000 | **1.0000** | 1.0000 | 0.9929 |
| | I3D [6] | | 0.9857 | 0.9500 | 0.9714 | 0.9429 |
| | **S-MIL** | | 1.0000 | 0.9964 | 1.0000 | 0.9786 |
| | **S-MIL-T** | | **1.0000** | 0.9964 | **1.0000** | **0.9964** |
| Fb | XN-avg [42] | c23 | **1.0000** | 0.9964 | 0.9964 | 0.9393 |
| Vb | LSTM [21] | | 0.9964 | 0.9929 | 0.9821 | 0.9393 |
| | I3D [6] | | 0.9286 | 0.9286 | 0.9643 | 0.9036 |
| | **S-MIL** | | 0.9857 | 0.9929 | 0.9929 | **0.9571** |
| | **S-MIL-T** | | 0.9964 | **0.9964** | **1.0000** | 0.9429 |
| Fb | XN-avg [42] | c40 | 0.9714 | **0.9250** | 0.9607 | 0.8607 |
| Vb | LSTM [21] | | 0.9643 | 0.8821 | 0.9429 | 0.8821 |
| | I3D [6] | | 0.9107 | 0.8643 | 0.9143 | 0.7857 |
| | **S-MIL** | | 0.9679 | 0.9143 | 0.9464 | **0.8857** |
| | **S-MIL-T** | | **0.9714** | 0.9107 | **0.9607** | 0.8679 |
| Frame level AUC | | | | | | |
| Fb | XN-avg [42] | c0 | 0.9938 | **0.9953** | 0.9936 | 0.9729 |
| | Face X-ray [26] | | 0.9917 | 0.9906 | 0.9920 | **0.9893** |
| Vb | MIL [50] | | 0.9951 | 0.9859 | 0.9486 | 0.9796 |
| | **S-MIL** | | **0.9984** | 0.9934 | **0.9961** | 0.9885 |

**Table 2: Quantitative comparisons with the state-of-the-art methods including frame-based(Fb) and video-based(Vb) algorithms on the FF++ dataset under different compress rates(CR) evaluated by video-level accuracy and frame-level AUC. Results in bold text indicate the best results while results in pink and blue indicate the best and second-best results appeared in proposed method.**

For this experiment, we take the FaceForensics++ (FF++)[42] dataset as our training data. Only the spatial-level encoding branch

| Methods | CR | FaceForensics++ | | | |
|---|---|---|---|---|---|
| | | DF | F2F | FS | NT |
| MIL [50] | c0 | 1.0000 | 0.9750 | 0.9679 | 0.9357 |
| **S-MIL** | | 1.0000 | 0.9964 | 1.0000 | 0.9786 |
| **S-MIL-T** | | **1.0000** | **0.9964** | **1.0000** | **0.9964** |
| MIL [50] | c23 | 0.9857 | 0.9857 | 0.9786 | 0.8714 |
| **S-MIL** | | 0.9857 | 0.9929 | 0.9929 | **0.9571** |
| **S-MIL-T** | | **0.9964** | **0.9964** | **1.0000** | 0.9429 |
| MIL [50] | c40 | 0.9500 | 0.8786 | 0.9071 | 0.8036 |
| **S-MIL** | | 0.9679 | **0.9143** | 0.9464 | **0.8857** |
| **S-MIL-T** | | **0.9714** | 0.9107 | **0.9607** | 0.8679 |

**Table 3: Ablation study on FF++ dataset under different compress rates evaluated by video-level accuracy in terms of S-MIL and spatial-temporal instances.**

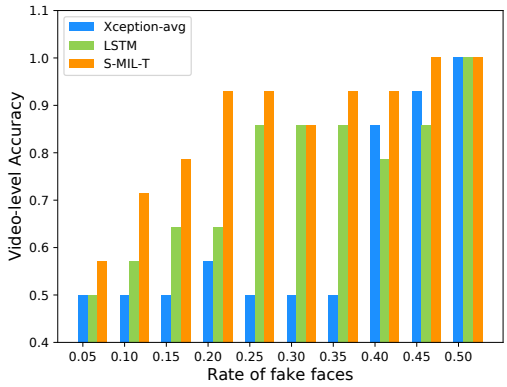


**Figure 6: Comparison for S-MIL-T, XceptionNet-avg, LSTM on different rate of fake faces in input sequences of FFPMS dataset. The S-MIL-T can get a good performance even when the fake rate is low while others can not.**

is employed for frame-level evaluation since there is no temporal information during testing. Specifically, we sample 20 frames (less than 10% of training data) from each video as training data and extract 8 frames from samples randomly to form the video input for S-MIL. During testing, each frame extracted from testing videos is fed into S-MIL individually for frame-level evaluations. We evaluate the resulting four models on this dataset using Area Under Curve (AUC) metric as frame-level methods [26, 42].

As shown in Table 2, the proposed S-MIL has achieved state-of-the-art performance with only 10% training data. This shows that our method can reliably judge input faces.

## 4.5 Ablation Study

**Component analysis.** As shown in Table 2 and 3, the proposed S-MIL outperforms the traditional MIL significantly under both frame-level and video-level settings. First, the proposed S-MIL can relieve the gradient vanishing problem, which can help to force every fake instance to be predicted as fake. Besides, the proposed weighting mechanism can help to focus on more important instances instead of treating instances uniformly. This can help to gain a reliable prediction. Tables 1, 2 and 3 show that with the

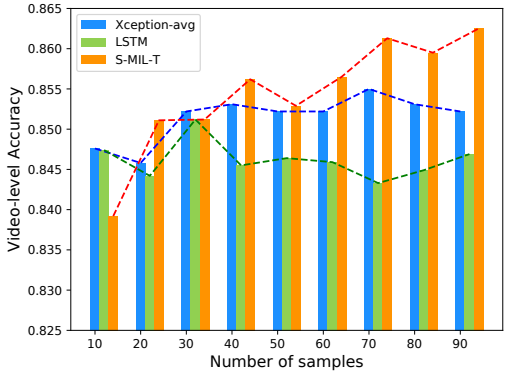


**Figure 7: Comparison results for DeepFake video detection with different numbers of input instances on DFDC dataset. With more samples, the proposed method can get a better performance while others have little improvements.**

spatial-temporal encoding module, the proposed method can get a better performance, which validates the effectiveness of spatial-temporal instances.

**Inference on different number of frames.** In video-based settings, for efficiency, sampling is commonly used for less computational and time costs. Thus, the number of sampling frames is an important factor in videos with noises. We explore the influences of different number of frames on frame-based method XceptionNet, video-based method LSTM, and the proposed S-MIL-T with DFDC dataset. As shown in Fig. 7, the accuracy of proposed S-MIL-T raises when more testing frames are sampled. However, with the growth of sample numbers, the accuracies of XceptionNet and LSTM have little changes, which in turn, shows the superiority of the proposed S-MIL-T.

## 5 CONCLUSION

In this paper, we introduce a new problem of partially attacked DeepFake video detection which is somewhat ignored in previous study but widely exists in real applications. We address this problem with MIL paradigm by treating faces and each input video as instances and bag respectively. Accordingly, we propose S-MIL by building direct connection between bag label prediction and instance embeddings, so that the gradient vanishing problem in traditional MIL can be alleviated. We also conduct theoretic analysis to verify the sharp property of the loss function of S-MIL. Second, a new spatial-temporal instance is designed to fully model the inconsistency between faces in nearby frames, which helps to improve the detection performance further. To inspire research on video-level DeepFake detection under partial faces manipulation situations, we construct a new dataset named FFPMS for evaluation of partially attacked DeepFake video detection task. Experiments on our FFPMS dataset and DFDC benchmark verify that our approach is superior to other counterparts for DeepFake video detection. Moreover, our approach can also be adapted to traditional single-frame detection task and achieves state-of-the-art performance on the FF++ datasets, showing a promising result for face information protection.

## REFERENCES


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

[2] Jaume Amores. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201 (2013), 81–105.

[3] Anonymous. [n.d.]. Example of Partially attacked DeepFake video on Youtube. [EB/OL]. https://www.youtube.com/watch?v=BU9YAHigNx8 Accessed April 4, 2020.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *arXiv preprint:1409.0473*.

[5] Belhassen Bayar and Matthew C Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In *the 4th ACM Workshop*. 5–10.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*. 6299–6308.

[7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*. 785–794.

[8] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2017. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection. (2017), 159–164.

[9] DeepFakes. 2019. www.github.com/deepfakes/ faceswap. *Accessed* (2019).

[10] DeepFakes. 2019. www.github.com/MarekKowalski/. *Accessed* (2019).

[11] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. 2001. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence* 89, 1-2 (2001), 31–71.

[12] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul S Krueger, and Michael Hahsler. 2019. Swapped Face Detection using Deep Learning and Subjective Assessment. *arXiv: Learning* (2019).

[13] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv preprint:1910.08854* (2019).

[14] Facebook. [n.d.]. DeepFake Detection Challenge. [EB/OL]. https://www.kaggle.com/c/deepfake-detection-challenge Accessed May 20, 2020.

[15] Jiashi Feng, Bingbing Ni, Qi Tian, and Shuicheng Yan. 2011. Geometric lp-norm feature pooling for image classification. In *CVPR*.

[16] Ji Feng and Zhihua Zhou. 2017. Deep MIML Network. In *AAAI*.

[17] Jessica Fridrich and Jan Kodovsky. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7, 3, 868–882.

[18] Miroslav Goljan and Jessica Fridrich. 2015. CFA-aware features for steganalysis of color images. *electronic imaging* 9409.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.

[20] David Guera and Edward J Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *AVSS*. 1–6.

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[22] Maximilian Ilse, Jakub M Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. *arXiv preprint:1802.04712* (2018).

[23] James D. Keeler, David E. Rumelhart, and Wee Kheng Leow. 1990. Integrated segmentation and recognition of hand-printed numerals. In *NIPS*.

[24] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint:1408.5882* (2014).

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint:1412.6980* (2014).

[26] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2019. Face X-ray for More General Face Forgery Detection. *CVPR*.

[27] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *CVPRW*.

[28] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2019. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint:1909.12962* (2019).

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*. 2980–2988.

[30] Zhouhan Lin, Minwei Feng, Cicero Nogueira Dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. *arXiv: Computation and Language*.

[31] Oded Maron and Tomas Lozano-Perez. 1998. A framework for multiple-instance learning. In *NIPS*.

[32] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. 2018. Detection of GAN-Generated Fake Images over Social Networks. In *IEEE MIPR*.

[33] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Weakly supervised object recognition with convolutional neural networks. In *NIPS*.

[34] Xunyu Pan, Xing Zhang, and Siwei Lyu. 2012. Exposing image splicing with inconsistent local noise variances. In *ICCP*. 1–10.

[35] Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspectbased sentiment analysis. In *EMNLP*.

[36] Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit Document Modeling through Weighted Multiple-Instance Learning. In *Journal of Artificial Intelligence Research*.

[37] Pedro O Pinheiro and Ronan Collobert. 2015. From image-level to pixel-level labeling with convolutional networks. In *CVPR*.

[38] Gwenole Quellec, Guy Cazuguel, Beatrice Cochener, and Mathieu Lamard. 2017. Multiple-Instance Learning for Medical Image and Video Analysis. *IEEE Reviews in Biomedical Engineering* 10, 213–234.

[39] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint:1511.06434* (2015).

[40] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2017. Distinguishing computer graphics from natural images using convolution neural networks. In *IEEE Workshop on Information Forensics and Security*.

[41] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niesner. 2018. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *CVPR*.

[42] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niesner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *arXiv preprint arXiv:1901.08971*.

[43] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)* 3 (2019), 1.

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*. 618–626.

[45] Korsuk Sirinukunwattana, Shan Raza, Yee Wah Tsang, David Snead, Ian Cree, and Nasir Rajpoot. 2016. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging* 35, 5 (2016), 1196–1206.

[46] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. 2018. Detecting both machine and human created fake face images in the wild. *In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security* (2018).

[47] Justus Thies, Michael Zollhofer, and Matthias Niesner. 2019. Deferred neural rendering: image synthesis using neural textures. *ACM Transactions on Graphics* 38, 4 (2019), 66.

[48] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niesner. 2018. Face2Face: real-time face capture and reenactment of RGB videos. *Communications of The ACM* 62, 1 (2018), 96–104.

[49] Run Wang, Lei Ma, Felix Juefeixu, Xiaofei Xie, Jian Wang, and Yang Liu. 2019. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. In *arXiv preprint arXiv:1909.06122*.

[50] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2016. Revisiting multiple instance neural networks. In *Pattern Recognition*.

[51] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. 2018. Revisiting multiple instance neural networks. *Pattern Recognition* 74 (2018), 15–24.

[52] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[53] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *arXiv preprint*.

[54] Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. 2017. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *MICCAI*.