Face Forensics in the Wild

Tianfei Zhou[1] , Wenguan Wang[1]* , Zhiyuan Liang[2] , Jianbing Shen[3,2]
[1]ETH Zurich [2]Beijing Institute of Technology [3]Inception Institute of Artificial Intelligence
https://github.com/tfzhou/FFIW

Abstract

On existing public benchmarks, face forgery detection techniques have achieved great success. However, when used in multi-person videos, which often contain many people active in the scene with only a small subset having been manipulated, their performance remains far from being satisfactory. To take face forgery detection to a new level, we construct a novel large-scale dataset, called FFIW$_{10K}$, which comprises 10,000 high-quality forgery videos, with an average of three human faces in each frame. The manipulation procedure is fully automatic, controlled by a domain-adversarial quality assessment network, making our dataset highly scalable with low human cost. In addition, we propose a novel algorithm to tackle the task of multi-person face forgery detection. Supervised by only video-level label, the algorithm explores multiple instance learning and learns to automatically attend to tampered faces. Our algorithm outperforms representative approaches for both forgery classification and localization on FFIW$_{10K}$, and also shows high generalization ability on existing benchmarks. We hope that our dataset and study will help the community to explore this new field in more depth.

1. Introduction

The rise of synthetic audiovisual media is forcing us towards a critical and unsettling realization: our belief that video and audio recordings are reliable representations of reality is no longer tenable. In particular, since emerging in 2017, the deepfake phenomenon has grown rapidly, requiring *face forensics* to recognize potentially manipulated facial regions in images and videos. Accurate face forgery detection[1] would have an immediate and far-reaching impact in alleviating the malicious intents of deepfakes, such as, face recognition attacks [42] and fake news [36, 4].

To help with this, several benchmarks have been established. The pioneering large-scale dataset, *i.e.*, FaceForensics++ [63], has greatly contributed to spurring interest and progress in the area of face forgery detection. However, as algorithms evolve, there have been signs of performance saturation on this dataset [62]. More recent datasets (*e.g.*, DeeperForensics-1.0 [37], DFDC [22], Celeb-DF [50]) thus employ more advanced synthesis techniques to produce highly realistic tampered faces. Even so, all previous datasets are subject to a significant limitation: they have a strong selection bias [69] to favor trimmed videos, each of which involves only one person. Thus, they provide insufficient representation of true visual world, which makes them inappropriate and unreliable to evaluate face forgery detection models in real-world, multi-person circumstances.

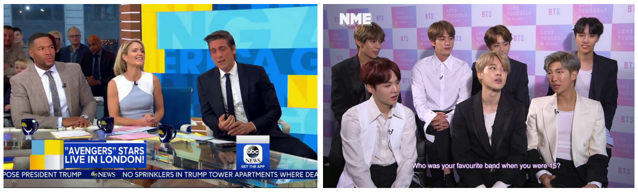

"AVENGERS" STARS LIVE IN LONDON!
NME
Who was your favourite band when you were 15?

Figure 1: **Representative examples from *FFIW*$_{10K}$, showing multi-person video frames with only a few faces being forged.** Can you recognize the manipulated ones?[2]

To take the research of face forensics into a new level, we introduce a new large-scale dataset, called *FFIW*$_{10K}$, promoting empirical study of face forgery detection in multi-person scenarios. In *FFIW*$_{10K}$, *each video involves multiple individuals but only some, not all, faces are manipulated* (see Fig. 1). This raises a significant challenge to current techniques: even for the fake videos, real faces are still in the majority. In particular, *FFIW*$_{10K}$ features 10,000 high-fidelity manipulated videos, with 12 seconds long on average, resulting in 33 hours of video in total. In comparison with existing datasets, *FFIW*$_{10K}$ has several distinguished features: **i)** *Real-world complexity.* The number of identities in each frame ranges from one to fifteen, with three on average, yielding a better representation of real visual scenes. **ii)** *High fidelity with low human cost.* The synthesis quality is controlled by a quality assessment network (Q-Net), which provides an effortless way to measure the realism of

*Corresponding author: *Wenguan Wang*.

[1] In this work, "forgery" refers to altering imagery by swapping faces.

[2] Answer: The middle person in the left image and the rightmost of the back row in the right image are fake.

forged faces. **iii)** *Large scale.* *FFIW*$_{10K}$ is comparable to the largest current dataset [37] in terms of the number of unique fake videos (see Table 1), and, more importantly, provides videos under multi-person settings.

FFIW$_{10K}$ provides both video- and face-level annotations, allowing benchmarking methods on both forgery classification and localization tasks. In addition, to bring the research into a more natural setting, *FFIW*$_{10K}$ provides two benchmark settings. In the first setting, benchmarking methods can make use of face-level supervision. However, in the second one, only video-level labels are allowed to be accessed during training. This makes the task more valuable from both academic and practical perspectives.

Along with *FFIW*$_{10K}$, we propose a discriminative attention model for face forgery classification and localization in multi-person scenarios. The model explores the idea of multiple instance learning [20, 52] and can be trained with video-level label only. It comprises three essential parts: i) a *multi-temporal-scale instance feature aggregation* module that summarizes short-term, long-term and global features of each face tracklet to obtain a robust and discriminative representation; ii) an *attention-based bag feature aggregation* module that adaptively aggregates the representations of all face tracklets into a video-level representation; and iii) a *sparse regularization loss* to enforce the sparsity of face selection for real/fake discrimination. The sparsity-regularized attention learning mechanism is tasked with automatically selecting possible tampered faces during classification, promoting the localization ability of the system.

In summary, our contributions are three-fold: **i)** To pave the avenue for face forgery detection in open world, we contribute *FFIW*$_{10K}$ dataset to the community, which is distinctive in its real-world complexity. As far we know, it is the first large-scale face forensics dataset for *fully unconstrained*, *multi-person* face forgery detection. **ii)** We propose a model-agnostic quality assessment model for synthesis quality management. Trained independently of deepfake methods, the model has high flexibility and accessibility, and can facilitate future dataset construction. **iii)** We propose a discriminative attention model for multi-person face forgery detection. By revisiting multiple instance learning and gathering diverse temporal context, it provides promising performance on both fake video classification and fake face localization tasks with only video-level supervision.

2. Related Work

Existing Face Forensics Datasets. Being the foundations of more advanced techniques, the pursuit of better datasets has attracted substantial research interest in the area of face forgery detection (see Table 1). Early attempts can be traced back to MICC-F2000 [5] and DSI-1 [19], in which faces were manipulated in still images under strictly constrained conditions. In recent years, great efforts have been devoted to establishing video-based datasets, such as UADFV [79], DF-TIMIT [42], FaceForensics++ [63], DFDC [21], Celeb-DF [50], VideoForensicsHQ [27] and DeeperForensics-1.0 [37]. With constantly upgraded face forgery techniques (*e.g.*, FaceSwap [2], NeuralTextures [65], FaceShifter [45], NVP [64]), videos of forged faces in some datasets seem deceptively real to the human eye. These datasets have undoubtedly advanced this field. Nonetheless, they are still limited in that most videos come from simple scenarios with only one or two identities. Therefore, benchmarking algorithms on these datasets is not sufficient for measuring their performance in practical scenarios.

Dataset	Year	Pub.	#Real	#Fake	#Synthetic Methods	#Face Per-frame
UADFV [79]	2018	ICASSP	49	49	1	1
DeepFake-TIMIT [42]	2018	arXiv	320	640	2	1
Deep Fake Detection [9]	2019	-	363	3,068	5	1
FaceForenscics++ [63]	2019	ICCV	1,000	4,000	4	1
DFDC Preview [22]	2019	arXiv	1,131	4,113	2	∼1
Celeb-DF [50]	2020	CVPR	590	5,639	1	1
DeeperForensics-1.0 [37]	2020	CVPR	50,000	10,000†	1	1
FFIW$_{10K}$ (Ours)	2020	-	10,000	10,000	3	3.15

Table 1: **Comparisons of *FFIW*$_{10K}$ with existing datasets.** As far as we know, *FFIW*$_{10K}$ is the first specializing in *unconstrained*, *multi-person* face forgery detection. †: in DeeperForensics-1.0 [37], each fake video is randomly perturbed for augmentation, and the perturbed videos are counted as new fake videos. Ignoring perturbation, DeeperForensics-1.0 only contains 1,000 unique fake videos, much fewer than the 10,000 in *FFIW*$_{10K}$.

To address the limitations of previous datasets, we introduce *FFIW*$_{10K}$ which targets at *multi-person* face forgery detection. *FFIW*$_{10K}$ is unique in its real-world complexity (*i.e.*, it covers multiple individuals), high-fidelity manipulation (*i.e.*, its quality is guaranteed by a model-agnostic quality assessment network, Q-Net) and scalability (*i.e.*, it is constructed in a unconstrained, automatic condition).

Neural Face Synthesis and Face Forensics Dataset Construction. Neural face manipulation has been a longstanding research topic in computer vision and computer graphics for over two decades [58]. The very first work, *i.e.*, Video Rewrite [11], automatically synthesizes human faces with proper lip sync to a given audio signal. Introduced later, face swapping systems [2, 66, 67] typically follow computer graphics pipelines, fitting a parametric 3D face model to target faces for manipulation. However, the performance of these methods relies heavily on the quality of the 3D model. Some alternatives [41, 65] thus alleviate this by combining graphics pipelines with learnable components, which can use imperfect 3D models for synthesis. More recently, GAN-based models [1, 78, 59, 45, 50, 37] have become popular due to their concise and flexible framework, which does not require expensive manual operation or acquisition hardware. This thus enables them to be frequently engaged in constructing face forensics datasets.

Although current deepfake generators can produce high-quality manipulations when source and target faces yield strong consistency (in terms of color, pose, illumination), they easily suffer from occlusions, glasses, profile faces or sudden motions. Therefore, during face forensics dataset construction, extensive human interventions are often included to guarantee the quality of collected data, by manually filtering out those suboptimal tampered examples [50, 37]. Thus building large-scale face forensics datasets is costly and time-consuming. We instead devise a quality assessment network to post-control the face manipulation procedure. The network is trained with domain-adversarial learning on a set of automatically collected training samples and is independent from face synthesis techniques. Thus it yields high flexibility and generalization, and allows large-scale dataset construction in a labor-efficient manner.

Face Forgery Detection. Recently, active research has been devoted to synthetic content detection in portrait videos [68, 73], in order to fight against the emerging threat of face swapping techniques. Early models [8, 49, 42, 54, 79, 4] make use of hand-crafted features (*e.g.*, blinking patterns, temporal flickering, face warping artifacts), which are manually designed to capture visual artifacts and inconsistencies generated in the fake face synthesis process. However, due to the limited representation ability of hand-designed features, they do not fit well towards more sophisticated facial manipulation techniques. With the advance of deep neural networks, recent approaches are built upon modern network architectures (*e.g.*, Xception [16], I3D [12]), and address image- [3, 80, 63, 75, 57, 68] or video-level [31, 6, 71, 37, 53] forgery classification. Some methods [7, 56, 46, 35, 23, 43] further focus on fine-grained localization of manipulated regions to provide better interpretability. Additionally, frequency domain analysis [62, 28, 15, 25, 24, 53], texture statistics [51], audio features [17, 55], and biological signals [18, 33, 61, 26] are also becoming popular for recognizing fake content.

Despite their success, current deep learning methods typically conduct forgery classification on trimmed videos and are prone to failing in real-world, multi-person scenarios. Though [47] made an initial attempt to address this, it is still confined to constrained scenarios due to the lack of large-scale datasets. In contrast, our *FFIW*$_{10K}$ enables us to make a more in-depth exploration of this new direction.

3. *FFIW*$_{10K}$ Dataset

Challenging datasets are catalysts for progress in the computer vision community. We therefore introduce *FFIW*$_{10K}$ to provide a better benchmark and help identify conditions under which current algorithms fail, with the hope of promoting further research efforts. Exemplars of *FFIW*$_{10K}$ are shown in Figs. 1 and 2. In the following, we present some important aspects of *FFIW*$_{10K}$.

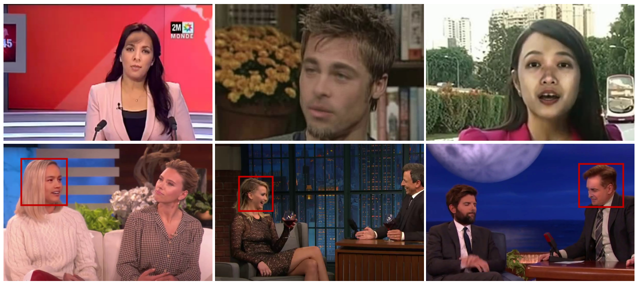


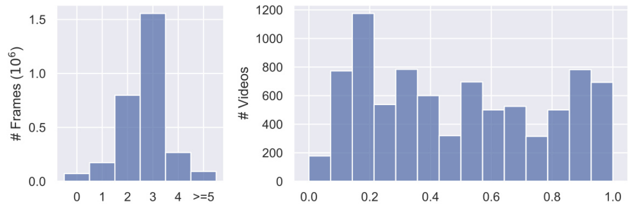
Figure 2: **Exemplar frames of deepfake datasets.** From left to right: 1_{st} row shows tampered faces in FaceForensics++ [63], Celeb-DF [50] and DeeperForensics-1.0 [37]; 2_{nd} row shows examples in *FFIW*$_{10K}$ created by FSGAN [59], DeepFaceLab [60] and FaceSwap [2]. Manipulated faces are denoted by red boxes.

3.1. Pristine Video Collection

To align with our target of multi-person face forgery detection, we collect pristine videos in the wild, ensuring that a large number of videos contain more than one individual. We start by searching a collection of videos from *YouTube* based on diverse keyword queries. To alleviate selection bias, the search is conducted by 10 people with self-chosen queries in different languages. For video quality, we only download high-resolution videos (480p or higher), yielding a total of 4,000 raw videos. Then, we split each video into four uniform clips, and randomly select one 12-second sequence from each. We filter out static or crowded sequences, as well as sequences containing few human faces. This results around 12,000 sequences, which we used as pristine videos for facial manipulation.

3.2. Facial Manipulation Procedure

For face swapping, we randomly select two videos from the pristine collection, *i.e.*, a *target* video in which a target face will be replaced, and a *source* video providing the identity of a source face that will be swapped onto the target face. Since both the source and target videos contain multiple identities, we pre-process them with off-the-shelf face detection and tracking algorithms [44, 77] to obtain a set of face tracklets. We then select the tracklets with the longest duration and highest resolution for swapping. To enrich the diversity of manipulated videos, we create each video with one of three face swapping methods, including two learning-based methods (DeepFaceLab [60] and FSGAN [59]), and one graphic-based method (FaceSwap [2]). Though these methods can produce compelling results, they still show weaknesses under varying conditions. For example, DeepFaceLab performs poorly in the presence of glasses and extreme poses, and FSGAN is weak in maintaining an even skin tone in dark scenes. Instead of previous works [50, 37] involving dense human interventions in dataset construction, we design a fully automatic procedure so that our dataset can be easily scaled. We develop a quality assessment network to quantitatively score each



(a) Number of faces per frame (b) Histogram of fake faces per video

Figure 3: **Statistics of *FFIW*$_{10K}$** (§3.3). (a) Distribution of face number per frame. (b) Ratio distribution of tampered faces.

manipulated face, and discard the synthetic faces with low scores (see Fig. 5 (b)), *i.e*., only $10,000$ fake videos with high quality scores and from different pristine videos are selected to build *FFIW*$_{10K}$. The network allows us to build *FFIW*$_{10K}$ with low human cost. For conciseness, we defer the discussion of the quality control procedure to §4, after we have provided all the necessary details of *FFIW*$_{10K}$.

3.3. Dataset Features and Statistics

To offer deeper insights into *FFIW*$_{10K}$, we next discuss its various attractive properties and descriptive statistics.

Real-World Complexity. Existing datasets fall in short of containing just one or two identities in each video (see Table 1), which does not accurately reflect the distribution in the real world. However, *FFIW*$_{10K}$ is designed to involve more human faces (1–15 per frame, 3.15 on average). The distribution of face number in each frame is shown in Fig. 3 (a). Another challenge *FFIW*$_{10K}$ provides is that each video contains both real and fake faces. We analyze the ratio of the number of tampered faces against the number of all faces in each video in Fig. 3 (b). As seen, in many videos, only a small percentage of faces are manipulated. These statistics are more representative of real world-applications and allow for in-depth benchmark analysis.

High Fidelity with Low Human Cost. We organize a user study to verify the quality of *FFIW*$_{10K}$. Specifically, a total of 50 computer science students are invited to assess the realness of synthetic videos in *FFIW*$_{10K}$ as well as two previous high-quality datasets (*i.e*., Celeb-DF [50] and DeeperForensics-1.0 [37]). Following [37], we randomly select 30 videos from each dataset and prepare a web-based platform to play each video once to the participants. Each participant is asked to score each video at five levels (0.2 – clearly fake, 0.4 – fake, 0.6 – borderline, 0.8 – real, 1.0 – clearly real). For each video, we average the scores of all users as the final score. Fig. 4 shows the results. As seen, more videos in *FFIW*$_{10K}$ are rated as 'real' and 'clearly real' than in the other two datasets. This can be attributed to: i) the intrinsic difficulties of multi-person face forgery detection; and ii) the effectiveness of the Q-Net in quality control.

Large Scale. As shown in Table 1, *FFIW*$_{10K}$ consists of 10K synthetic as well as 10K real videos, with about 33 hours and more than 7.2M frames in total. Note that the number of *unique* fake videos in *FFIW*$_{10K}$ is ten orders of magnitude larger than DeeperForensics-1.0, which treats a manipulated video and its distorted versions (perturbed by Gaussian blur, JPEG compression, *etc*.) as different videos.

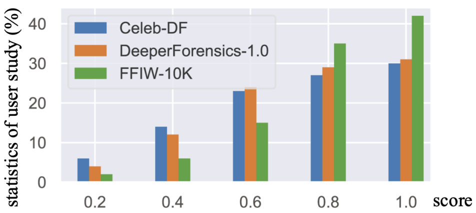


Figure 4: **Statistics of user study on data fidelity** (§3.3).

Dataset Annotation. For completeness, *FFIW*$_{10K}$ provides both face-level and video-level labels; a total of 3.2M real faces and 1.1M fake faces of 3.6K persons are annotated. Note that our method explores only the usage of video-level labels, addressing high utility in practical applications.

Dataset Split. We split *FFIW*$_{10K}$ into separate `train`, `val` and `test` sets. Following random selection of pristine video clips, we arrive at a unique split consisting of 16,000 training, 500 validation, and 3,500 test videos. In each split, each fake video is companied with its real video.

4. Domain-Adversarial Quality Control

Domain-Adversarial Quality Assessment Network (Q-Net). As mentioned in §3.2, for facilitating dataset construction, we design a Q-Net $\mathcal{F}^Q$ (VGG16-based) that automatically evaluates the quality of each swapped face and hence allows us to effortless filter out low-fidelity faces. The main challenge here is that it is hard to collect precise quality annotations directly from the face swapping algorithms (*i.e*., DeepFaceLab[60], FSGAN[59], FaceSwap[2]). Inspired by [30], we collect data in a semi-supervised way. Our algorithm is built on the observation that, for most generative models, the quality of their synthesized images progressively improves as the training continues. This enables us to collect face images generated by various unconditional generative models (*i.e*., StyleGAN [39], StyleGAN2 [40], PGGAN [38]) in different iterations and use the corresponding iteration number as the pseudo groundtruth quality score. Specifically, for each generated face I_i, the pseudo score is defined as: $s_i = 0.9 \times n/N$, where n and N indicate the iteration number and the maximum iteration, respectively. We train each model [39, 40, 38] on FFHQ[39] for $N = 5{,}000$ iterations and select 20 images per iteration, leading to a total of 300,000 training samples $\{I_i, s_i\}_i$.

Directly learning on the collected data is insufficient, since the Q-Net may overfit to specific artifacts of different generative models [80, 24], rather than learning discriminative and informative representations for realism perception. This will lead to poor generalization when tested on *FFIW*$_{10K}$ due to the *domain shift* between the training and testing distributions. To address this issue, we introduce domain-adversarial regularization [29] to encour-

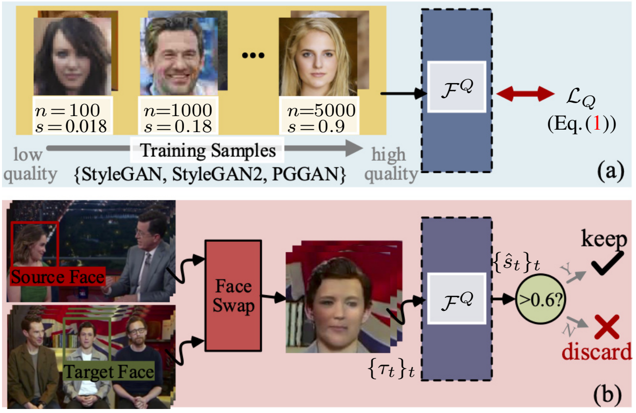


Figure 5: (a) Training stage of Q-Net. (b) Facial manipulation and Q-Net based quality control during the construction of *FFIW*$_{10K}$.

age domain-invariant feature learning in the course of the optimization. Concretely, we assign each training sample I_i an extra domain label $d_i \in \{\text{StyleGAN}, \text{StyleGAN2}, \text{PGGAN}\}$. For each I_i, we let $\mathcal{F}^Q$ predict both the quality score and domain label. As shown in Fig. 5 (a), with all the training samples $\{I_i, s_i, d_i\}_i$, $\mathcal{F}^Q$ is learned by minimizing:

$$\mathcal{L}_Q = \sum\nolimits_i \mathcal{L}_{l_1}(\hat{s}_i, s_i) - \alpha\mathcal{L}_{\text{CE}}(\hat{d}_i, d_i), \tag{1}$$

where $\mathcal{L}_{l_1}$ and $\mathcal{L}_{\text{CE}}$ are l_1 and cross-entropy losses, respectively, $\hat{s}$ and $\hat{d}$ indicate the estimated quality score and domain label, respectively, and $\alpha > 0$. Through the domain-adversarial regularization (*i.e.*, $-\alpha\mathcal{L}_{\text{CE}}(\hat{d}_i, d_i)$), our Q-Net is enforced to learn discriminative feature representations that are invariant to the change of data distributions, hence gaining improved generalization ability.

To evaluate $\mathcal{F}^Q$, we carry out a user study over a set of images to measure the consistency between model predictions and human assessments (see *supplementary*). The results from the user study confirm our Q-Net is effective.

Q-Net based Quality Control. After training, Q-Net is employed to automate the construction of *FFIW*$_{10K}$. For each swapped face τ, created by DeepFaceLab, FSGAN or FaceSwap, we compute its quality score through Q-Net: $\hat{s} = \mathcal{F}^Q(\tau)$. Then, for each manipulated face tracklet, it will be preserved only if its quality score, averaged over the swapped faces it contains, is larger than 0.6 (see Fig. 5 (b)).

5. Face Forgery Detection Framework

5.1. Discriminative Attention Model

In this section, we elaborate on our discriminative attention model for multi-person face forgery detection, which falls into the multiple instance learning (MIL) regime [20, 52]. In MIL, labels are associated with groups of instances (or *bags*), while instance labels are unobserved. The learning procedure aims to combine instance knowledge and predict labels on the bag level. In our problem, each video $\mathcal{V}$ corresponds to a bag, with its class label $l_{\mathcal{V}} \in \{\text{fake}, \text{real}\}$. A bag consists of K instances with unknown labels, each of which is a tracklet of faces, obtained by [44, 77]. We then formulate MIL-based face forgery detection as:

$$\max_{\boldsymbol{a}_{\mathcal{V}} \in [0,1]^K} \log p(l_{\mathcal{V}} | \{\boldsymbol{Y}_k\}_{k=1}^K, \boldsymbol{a}_{\mathcal{V}}) + \log p(\boldsymbol{a}_{\mathcal{V}}), \tag{2}$$

where $\boldsymbol{Y}_k$ denotes the representation of the k-th tracklet instance, and $\boldsymbol{a}_{\mathcal{V}}$ is a tracklet-aware attention vector, in which each value measures the likelihood of the corresponding tracklet being fake. The first term $\log p(l_{\mathcal{V}} | \{\boldsymbol{Y}_i\}_{i=1}^K, \boldsymbol{a}_{\mathcal{V}})$ prefers $\boldsymbol{a}_{\mathcal{V}}$ with high discriminative capacity for classification, while the second term $\log p(\boldsymbol{a}_{\mathcal{V}})$ models the prior distribution of $\boldsymbol{a}_{\mathcal{V}}$. With Eq. (2), we design a *multi-temporal-scale instance feature aggregation* module (§5.2) to learn instance representations $\{\boldsymbol{Y}_k\}_{k=1}^K$, an *attention-based bag feature aggregation* module (§5.3) to fuse $\{\boldsymbol{Y}_k\}_{k=1}^K$ into a video descriptor according to $\boldsymbol{a}_{\mathcal{V}}$, and a *sparse attention regularization loss* (§5.4) to model the distribution of $\boldsymbol{a}_{\mathcal{V}}$.

5.2. Multi-Temporal-Scale Instance Feature Aggregation

Let us denote $\Gamma = \{\tau_1, \ldots, \tau_T\}$ as a tracklet instance with T face regions detected from $\mathcal{V}$, each face region τ_t is represented by a feature vector $\boldsymbol{x}_t \in \mathbb{R}^D$. Here we aim to learn a compact representation $\boldsymbol{Y}$ for Γ:

$$\boldsymbol{Y} = \mathcal{F}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) \in \mathbb{R}^D, \tag{3}$$

where the aggregation function $\mathcal{F}$ can be naturally implemented as a global pooling operation (*e.g.*, max-pooling, average pooling or log-sum-exponential pooling [10]) over all the input features. However, global statistics cannot describe rich relations among different face regions, especially the *temporal order* within the tracklet, which are informative for recognizing temporal inconsistency (*e.g.*, eye blinking patterns [48], temporal artifacts [31]) in manipulated face sequences. We thus propose a multi-temporal-scale feature aggregation module for more discriminative instance representation learning. Formally, suppose $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{D \times T}$ be the raw tracklet representation in matrix form. $\mathcal{F}$ is achieved by a sequence of short-term $\mathcal{F}^s$, long-term $\mathcal{F}^l$ and global $\mathcal{F}^g$ aggregation operations (see Fig. 6):

$$\textit{short-term aggregation:} \quad \boldsymbol{S} = \mathcal{F}^s(\boldsymbol{X}) \in \mathbb{R}^{D \times T}, \tag{4}$$

$$\textit{long-term aggregation:} \quad \boldsymbol{L} = \mathcal{F}^l(\boldsymbol{S}) \in \mathbb{R}^{D \times T}, \tag{5}$$

$$\textit{global aggregation:} \quad \boldsymbol{Y} = \mathcal{F}^g(\boldsymbol{L}) \in \mathbb{R}^{D}. \tag{6}$$

Here, $\mathcal{F}^g$ denotes max-pooling. $\boldsymbol{S}$ and $\boldsymbol{L}$ are intermediate features after short-term $\mathcal{F}^s$ and long-term $\mathcal{F}^l$ aggregation operations, respectively, which will be detailed later.

Short-Term Feature Aggregation. We propose a densely connected dilated temporal convolution module to achieve $\mathcal{F}^s$ in Eq. (4). The module combines the advantages of atrous convolution [14] and dense connectivity [34] to effectively enlarge the field of view of filters to capture large temporal context. Specifically, $\mathcal{F}^s$ is a stack of L atrous convolutional layers, *i.e.*, $\{\mathcal{F}_l^{\text{atr\_conv}}\}_{l=1}^L$, where the dilation rate r_l

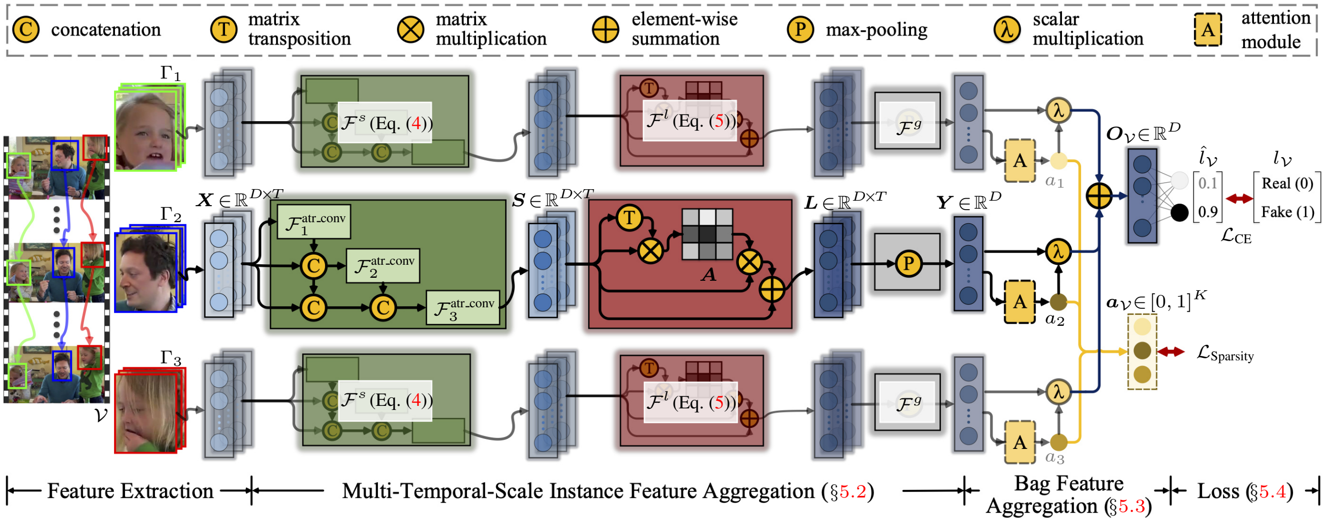

C concatenation
T matrix transposition
⊗ matrix multiplication
⊕ element-wise summation
P max-pooling
λ scalar multiplication
A attention module
Γ_1
Γ_2
Γ_3
$\mathcal{V}$
$\boldsymbol{X} \in \mathbb{R}^{D\times T}$
$\mathcal{F}^s$ (Eq. (4))
$\mathcal{F}_1^{\text{atr\_conv}}$
$\mathcal{F}_2^{\text{atr\_conv}}$
$\mathcal{F}_3^{\text{atr\_conv}}$
$\boldsymbol{S} \in \mathbb{R}^{D\times T}$
$\mathcal{F}^l$ (Eq. (5))
$\boldsymbol{A}$
$\boldsymbol{L} \in \mathbb{R}^{D\times T}$
$\mathcal{F}^g$
$\boldsymbol{Y} \in \mathbb{R}^{D}$
a_1
a_2
a_3
$\boldsymbol{O}_{\mathcal{V}} \in \mathbb{R}^{D}$
$\hat{l}_{\mathcal{V}}$
$l_{\mathcal{V}}$
0.1
0.9
Real (0)
Fake (1)
$\mathcal{L}_{\text{CE}}$
$\boldsymbol{a}_{\mathcal{V}} \in [0,1]^K$
$\mathcal{L}_{\text{Sparsity}}$
Feature Extraction
Multi-Temporal-Scale Instance Feature Aggregation (§5.2)
Bag Feature Aggregation (§5.3)
Loss (§5.4)

Figure 6: **Framework of the proposed discriminative attention model** (§5) for face forgery detection in multi-person scenarios.

is increased layer by layer. Each $\mathcal{F}_l^{\text{atr\_conv}}$ takes the concatenated features of all proceeding layers, $[\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{l-1}] \in \mathbb{R}^{(l\times D)\times T}$ as inputs, and outputs:

$$\boldsymbol{S}_l = \mathcal{F}_l^{\text{atr\_conv}}([\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{l-1}]) \in \mathbb{R}^{D\times T}, \tag{7}$$

where $\boldsymbol{S}_0 = \boldsymbol{X}$. Here, $\mathcal{F}^{\text{atr\_conv}}$ is able to efficiently capture temporal patterns over a relatively wide range without drastically increasing the number of parameters. The dense connection structure enables gradually assembling more temporal cues from different layers. Therefore, with a large receptive field, $\mathcal{F}^s$ finally produces a powerful, short-term descriptor $\boldsymbol{S}$ for the tracklet Γ, by comprehensively modeling and fusing context over different local temporal scales.

In practice, each $\mathcal{F}_l^{\text{atr\_conv}}$ is implemented by: bn-relu-conv(1×1)-bn-relu-conv$(3\times 3, r_l)$-bn-conv(1×1). Here, the first 1×1 conv reduces the feature dimension to $(l\times D)/4$ for computational efficiency, the 3×3 conv with dilation rate r_l facilitates multi-scale feature learning, and the second 1×1 conv outputs the feature $\boldsymbol{S}_l \in \mathbb{R}^{D\times T}$ at l-th layer. We use $L=3$ layers of dilated convolution with rates $r=\{1,2,4\}$, respectively, as shown in Fig. 6.

Long-Term Feature Aggregation. In addition to the short-term temporal context learning, we conduct long-term context aggregation $\mathcal{F}^l$ (Eq. (5)) over short-term feature $\boldsymbol{S}$ to learn a non-local informative representation for the tracklet Γ. Specifically, we employ self-attention [72, 76] to model the long-range, multi-level dependencies among temporal features in $\boldsymbol{S}$ (see Fig. 6). We first compute the normalized correlation between each pair of temporal feature vectors in $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T] \in \mathbb{R}^{D\times T}$ through pairwise dot product:

$$\begin{aligned}\boldsymbol{A} &= \texttt{softmax}(\boldsymbol{S}^\top \boldsymbol{S}) \\ &= \texttt{softmax}([\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T]^\top [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T]) \in [0,1]^{T\times T}.\end{aligned} \tag{8}$$

The affinity matrix $\boldsymbol{A}$ stores similarity scores corresponding to all pairs of features in $\boldsymbol{S}$, *i.e.*, the (i,j)-th element of $\boldsymbol{A}$ gives the similarity between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. $\texttt{softmax}(\cdot)$ normalizes each column of the input. Next, attention summaries are computed as $\boldsymbol{S}\boldsymbol{A} \in \mathbb{R}^{D\times T}$, and used to generate the long-term descriptor $\boldsymbol{L}$ for Γ in a residual form:

$$\boldsymbol{L} = \mathcal{F}^l(\boldsymbol{S}) = \boldsymbol{S}\boldsymbol{A} + \boldsymbol{S} \in \mathbb{R}^{D\times T}. \tag{9}$$

Thus $\boldsymbol{L}$ encodes both the long-term $\boldsymbol{S}\boldsymbol{A}$ and short-term information $\boldsymbol{S}$, with enhanced representability.

5.3. Attention-Based Bag Feature Aggregation

After applying max-pooling based global aggregation $\mathcal{F}^g$ over $\boldsymbol{L}$ (Eq. (6)), we get a compact and discriminative representation $\boldsymbol{Y} \in \mathbb{R}^D$ for each tracklet Γ. For video $\mathcal{V}$, all the K detected tracklets form a bag. We further adaptively aggregate all instance features $\{\boldsymbol{Y}_k\}_{k=1}^K$ into a global bag-level representation, using learnable attention:

$$\boldsymbol{O}_{\mathcal{V}} = \sum\nolimits_{k=1}^{K} a_k \boldsymbol{Y}_k \in \mathbb{R}^D, \tag{10}$$

where $\boldsymbol{a}_{\mathcal{V}} = (a_1, \ldots, a_K) \in [0,1]^K$ is a vector of scalar attention weights, and each weight a_k is computed by:

$$a_k = \frac{\exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{W}^\top \boldsymbol{Y}_k)\}}{\sum_{k'=1}^{K} \exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{W}^\top \boldsymbol{Y}_{k'})\}} \in [0,1]. \tag{11}$$

Here $\boldsymbol{w} \in \mathbb{R}^C$ and $\boldsymbol{W} \in \mathbb{R}^{D\times C}$ are learnable parameters. Through the attention-aware pooling, our method enjoys *high flexibility* to absorb faithful knowledge from representative instances for more accurate video-level classification, and *better interpretability* to locate the manipulated faces according to the attention $\boldsymbol{a}_{\mathcal{V}}$. For face forgery localization, we regard a tracklet Γ_k as fake if $a_k > 0.75$. The threshold 0.75 is determined by grid search over *FFIW*$_{10K}$ val.

5.4. Loss Function

Given the video-level feature representation $\boldsymbol{O}_{\mathcal{V}} \in \mathbb{R}^D$, a fully-connected layer is added for forgery classification. In addition, since only a sparse subset of faces are manipulated in most videos, we introduce sparse regularization over the attention vector $\boldsymbol{a}_{\mathcal{V}}$ to select a few most possibly tampered faces. Thus the overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\hat{l}_{\mathcal{V}}, l_{\mathcal{V}}) + \beta \mathcal{L}_{\text{Sparsity}}(\boldsymbol{a}_{\mathcal{V}}). \qquad (12)$$

Here, $\mathcal{L}_{\text{CE}}$ indicates the binary cross-entropy loss, and $\mathcal{L}_{\text{Sparsity}}$ is a sparse regularization term that is formulated as the l_1 norm of $\boldsymbol{a}_{\mathcal{V}}$, *i.e.*, $\mathcal{L}_{\text{Sparsity}}(\boldsymbol{a}_{\mathcal{V}}) = \|\boldsymbol{a}_{\mathcal{V}}\|_1$. The coefficient $\beta > 0$ controls the trade-off between the two terms.

5.5. Implementation Details

Preprocessing. For each video, we detect human faces in each frame [44] and associate them across frames to obtain a set of tracks [77]. To incorporate more spatial context, we extend each face region by a factor of 1.2 along the width and height, uniformly resized into 224×224 resolution.

Training Details. We employ ResNet-50 [32] as the backbone network and extract features after the average pooling layer as the representation of each face ($D = 2048$). The whole network is trained end-to-end using the Adam optimizer with learning rate 1e-4 and batch size 32. During training, we apply random perturbations (*e.g.*, horizontal flipping, color jitter) on each track for data augmentation. The coefficient β in Eq. (12) is empirically set to 0.001.

Reproducibility. Our model is implemented on PyTorch and trained on four NVIDIA Tesla V100 GPUs. To provide full details of our method, our codes are released.

6. Experiment

On top of *FFIW*$_{10K}$, we examine the proposed as well as representative face forgery detection methods on two tasks: face forgery classification (§6.1) and localization (§6.2). Then, we conduct experiments for assessing cross-dataset generalization abilities of various approaches in §6.3. Finally, in §6.4, a set of ablation studies are performed.

Competitors. Most previous approaches are designed for single-person scenarios, and, in practice, suffer from training divergence on *FFIW*$_{10K}$ (due to the influence of large amounts of real faces in manipulated videos). Therefore, we train these models on *FFIW*$_{10K}$ `train` with face-level labels. In particular, we select four frame-based (*i.e.*, Xception [63], MesoNet [3], FWA [49], PatchForensics [13]) and three video-based (*i.e.*, C3D [70], TSN [74], I3D [12]) models for comparison. Note that these models show compelling performance on existing datasets [63, 50, 37]. In addition, S-MIL [47] is employed as another baseline which can be trained using only video-level labels. All training protocols follow the original papers unless stated otherwise.

Evaluation Protocol. To fairly benchmark *FFIW*$_{10K}$, we devise a unified evaluation protocol that is applicable to all the methods. In particular, each test video is first parsed into a set of face tracklets, and each approach determines the possibility of each tracklet to be fake. This is natural for video-based methods since they work on tracklets, while for frame-level methods we use the average score of all faces in each tracklet as its score. Based on the tracklet-level predictions, we compute area under the receiver operating characteristic curve (AUC) as the metric of the classification task, as well as mean average precision (mAP) for the localization task. Following conventions [63, 50, 37], we also report video-level accuracy score (ACC) for classification.

Methods	classification		localization
	ACC (%)	AUC (%)	mAP (%)
frame-based methods: using face-level labels as supervision			
Xception [63]	54.1	56.1	17.9
MesoNet [3]	53.8	55.4	17.7
PatchForensics [13]	58.9	61.6	18.9
FWA [49]	60.2	63.1	19.2
video-based methods: using face-level labels as supervision			
TSN [74]	61.1	62.8	21.7
C3D [70]	64.3	65.5	23.9
I3D [12]	68.8	69.5	29.7
video-based methods: using video-level labels as supervision			
S-MIL [47]	59.8	61.2	-
Ours	**69.4**	**70.9**	**30.8**

Table 2: **Quantitative results for face forgery classification and localization** on `test` set of *FFIW*$_{10K}$ (§6.1 and §6.2).

6.1. Face Forgery Classification

We first investigate the classification performance of the approaches on *FFIW*$_{10K}$ `test`. Although this task has been well studied in single-person scenarios, we observe from Table 2 that previous methods produce poor classification results on *FFIW*$_{10K}$, even though they are trained with face-level labels. Our model outperforms all the compared methods by a large margin. This is encouraging given that our model only accesses to video-level labels. We note that our model significantly outperforms S-MIL [47], which is also trained using video-level labels. Additionally, we can observe that the top performance in *FFIW*$_{10K}$ is still far from being satisfactory, thus we hope that our new dataset could encourage continuous efforts in this challenging task.

6.2. Face Forgery Localization

We next analyze the performance of the approaches on face forgery localization. This task is more practical and challenging, yet is rarely explored in the literature. As shown in Table 2, the video-based methods [74, 70, 12] consistently outperform image-based methods [63, 3, 13, 49] in terms of mAP. Benefiting from our multi-temporal-scale feature aggregation (§5.2) and attention-based selection (§5.3) mechanisms, our approach achieves the best performance, even without precise, face-level supervision. Some visual results are depicted in Fig. 7, showing the strong capability of our model in isolating high-fidelity tampered faces from complex, multi-person scenes.

6.3. Cross-Dataset Evaluation and Generalization

Furthermore, we examine the cross-dataset performance of various approaches and the generalization ability of our

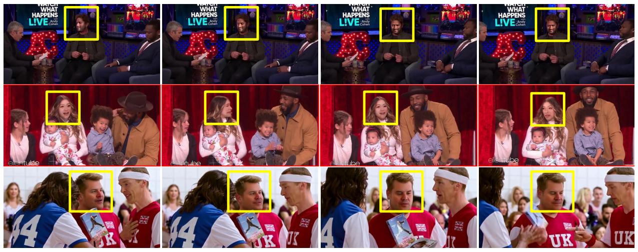


Figure 7: **Visual results for face forgery localization** on `test` set of *FFIW*$_{10K}$ (§6.2).

Methods	FF++ [63]	DFDC Preview [22]	Celeb-DF [50]
Xception-FF++ [63]	99.2	49.9	48.2
Capsule-FF++ [57]	96.6	53.3	57.5
Xception-c40-FF++ [63]	95.5	69.7	65.5
F^{3}Net-FF++ [62]	**99.9**	-	-
Face X-ray-FF++ [46]	99.2	73.5	74.8
TRN-FF++ [53]	99.1	-	76.7
Xception-c40-FFIW$_{10K}$	95.7	71.3	66.9
Ours-FF++	99.5	72.8	75.3
Ours-FFIW$_{10K}$	99.3	**74.1**	**78.3**

Table 3: **Cross-dataset generalization evaluation for face forgery classification**, in terms of AUC (%). See §6.3 for details.

FFIW$_{10K}$. As shown in Table 3, all comparative methods are trained on FF++ `train` [63] and evaluated on test sets of FF++ [63], DFDC Preview [22], and Celeb-DF [50], respectively. "Ours-FF++", also trained on FF++, produces comparable performance against other competitors, verifying the efficacy of our model. In addition, we evaluate the generalization ability of our *FFIW*$_{10K}$. We train our model as well as Xception-c40 [63] on *FFIW*$_{10K}$ `train` and report their performance on other datasets. We see that both "Ours-*FFIW*$_{10K}$" and "Xception-c40-*FFIW*$_{10K}$" outperform their alternatives, *i.e.*, "Ours-FF++" and "Xception-c40-FF++", which are trained on FF++ `train`. Hence, "Ours-*FFIW*$_{10K}$" shows superior performance on DFDC Preview and Celeb-DF. These experiment results reveal that *FFIW*$_{10K}$ has a low data bias and is well qualified to be used for training and evaluating face forgery detection models.

6.4. Model Ablations

We next conduct ablative studies of our model on *FFIW*$_{10K}$ `test`. The results are summarized in Table 4.

Aspect	Variants	classification ACC(%)	 AUC(%)	localization mAP(%)
Full Model	-	**69.4**	**70.9**	**30.8**
Instance Feature Aggregation (§5.2)	max-pooling	64.5	66.2	24.6
	avg-pooling	63.9	65.7	24.1
	w/o short-term (Eq. (4))	68.3	69.7	29.6
	w/o long-term (Eq. (5))	69.0	70.4	30.2
Bag Feature Aggregation (§5.3)	max-pooling	67.3	69.5	-
	avg-pooling	64.7	66.8	-
Loss Function (§5.4)	*w/o* $\mathcal{L}_{\text{sparsity}}$ (Eq. (12))	68.6	70.3	28.5

Table 4: **Ablation study** on `test` set of *FFIW*$_{10K}$ (see §6.4).

Instance Feature Aggregation. To study the impact of our multi-temporal-scale context aggregation (Eqs. (4-6)), we first develop two baselines by directly applying max- or avg-pooling over raw tracklet feature $\boldsymbol{X}$ to obtain a global compact descriptor $\boldsymbol{Y}$, without multi-temporal-scale feature learning. We can easily find that both models perform significantly worse than our full model across all metrics. This confirms that the global statistics are not eligible for encoding high-order relationships in $\boldsymbol{X}$, resulting in poor performance. We further separately analyze the short-term (Eq. (4)) and long-term (Eq. (5)) aggregation modules. As seen, by dropping the short-term module, the model encounters a performance drop (70.9% → 69.7% over AUC, and 30.8%→29.6% over mAP). A similar trend is also observed after discarding the long-term aggregation module.

Bag Feature Aggregation. To investigate the efficacy of the learnable attention mechanism for bag feature aggregation, we compare it with two baseline models which carry out the video-level feature summarization in Eq. (10) by max- and avg-pooling, respectively. We see that our attention-based aggregation mechanism brings favorable performance improvements over the baselines for classification, which can be attributed to its ability to automatically highlight the most possible instances for discrimination.

Efficacy of $\mathcal{L}_{\text{Sparsity}}$. At last, we study the necessity of the sparsity constraint $\mathcal{L}_{\text{Sparsity}}$ in Eq. (12). Since the term provides an appropriate modeling of the data distribution (*i.e.*, tampered faces are sparse in manipulated videos), it contributes to great performance improvements, especially in the localization task (28.5%→30.8% in terms of mAP).

7. Limitation and Discussion

For our dataset, its difficulty is limited to the adopted face swapping algorithms. This limitation is also shared by existing datasets. Considering the rapid advance of face swapping and forgery detection techniques, it is hard to maintain a long life-span for face forensics datasets. Our domain-adversarial quality control strategy may provide a feasible solution – one can automatically update the dataset by using more advanced deepfake techniques. For our model, it faces difficulties in the scenes with slow illumination change and stable motions. In such cases, the manipulated faces within a same tracklet usually show strong consistency and less artifacts. Thus the long-term features are less informative, easily leading to inferior performance. Hence, current study for "forgery" is mainly around face swapping. However, given the broader concerns about how imagery is being altered in order to influence political sphere, "forgery" should be explored in a larger extent, such as manipulating body movements, changing facial expressions, synthesizing realistic talking head videos, or swapping faces under controllable camera characteristics.

Acknowledgment This work was supported by Zhejiang Lab's Open Fund (No. 2019KD0AB04) and CCF-Baidu Open Fund.

References

[1] Deepfakes. https://github.com/deepfakes/faceswap. accessed November 10, 2020. 2

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap/. accessed November 10, 2020. 2, 3, 4

[3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, 2018. 3, 7

[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshop*, 2019. 1, 3

[5] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A sift-based forensic method for copy–move attack detection and transformation recovery. *IEEE TIFS*, 6(3):1099–1110, 2011. 2

[6] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *ICCV Workshop*, 2019. 3

[7] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE TIP*, 28(7):3286–3300, 2019. 3

[8] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. *IEEE TIFS*, 7(3):1003–1017, 2012. 3

[9] Google AI blog. Contributing data to deepfake detection research. 2019. 2

[10] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 5

[11] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *SIGGRAPH*, 1997. 2

[12] João Carreira, Andrew Zisserman, and Quo Vadis. Action recognition? a new model and the kinetics dataset. In *CVPR*, 2018. 3, 7

[13] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020. 7

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 5

[15] Zehao Chen and Hua Yang. Manipulated face detector: Joint spatial and frequency domain attention network. *arXiv preprint arXiv:2005.02958*, 2020. 3

[16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 3

[17] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. *arXiv preprint arXiv:2005.14405*, 2020. 3

[18] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE TPAMI*, 2020. 3

[19] Tiago José De Carvalho, Christian Riess, Elli Angelopoulou, Helio Pedrini, and Anderson de Rezende Rocha. Exposing digital image forgeries by illumination color classification. *IEEE TIFS*, 8(7):1182–1194, 2013. 2

[20] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 2, 5

[21] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2

[22] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 1, 2, 8

[23] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable forgery detection with locality-aware autoencoder. *arXiv preprint arXiv:1909.05999*, 2019. 3

[24] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 3, 4

[25] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. 3

[26] Steven Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Jha. Predicting heart rate variations of deepfake videos using neural ode. In *ICCV Workshop*, 2019. 3

[27] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Videoforensicshq: Detecting high-quality manipulated face videos. *arXiv preprint arXiv:2005.10360*, 2020. 2

[28] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. *arXiv preprint arXiv:2003.08685*, 2020. 3

[29] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 4

[30] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *ECCV*, 2020. 4

[31] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *AVSS*, 2018. 3, 5

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[33] Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*, 2020. 3

[34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 5

[35] Yihao Huang, Felix Juefei-Xu, Run Wang, Xiaofei Xie, Lei Ma, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. Fakelocator: Robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. *arXiv preprint arXiv:2001.09598*, 2020. 3

[36] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via

learned self-consistency. In *ECCV*, 2018. 1

[37] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020. 1, 2, 3, 4, 7

[38] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2017. 4

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4

[40] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 4

[41] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37(4):1–14, 2018. 2

[42] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 1, 2, 3

[43] Jia Li, Tong Shen, Wei Zhang, Hui Ren, Dan Zeng, and Tao Mei. Zooming into face forensics: A pixel-level analysis. *arXiv preprint arXiv:1912.05790*, 2019. 3

[44] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *CVPR*, 2019. 3, 5, 7

[45] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *CVPR*, 2020. 2

[46] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 3, 8

[47] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *ACM MM*, 2020. 3, 7

[48] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018. 5

[49] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 3, 7

[50] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *CVPR*, 2020. 1, 2, 3, 4, 7, 8

[51] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, 2020. 3

[52] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, 1998. 2, 5

[53] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, 2020. 3, 8

[54] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACV Workshop*, 2019. 3

[55] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: A deepfake detection method using audio-visual affective cues. *arXiv preprint arXiv:2003.06711*, 2020. 3

[56] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 3

[57] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019. 3, 8

[58] Thanh Thi Nguyen, Cuong M Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 1, 2019. 2

[59] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019. 2, 3, 4

[60] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Jian Jiang, Luis RP, Sheng Zhang, Pingyu Wu, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 3, 4

[61] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. *arXiv preprint arXiv:2006.07634*, 2020. 3

[62] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 1, 3, 8

[63] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 1, 2, 3, 7, 8

[64] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2019. 2

[65] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 38(4):1–12, 2019. 2

[66] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM TOG*, 34(6):183–1, 2015. 2

[67] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016. 2

[68] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179*, 2020. 3

[69] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1

[70] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 7

[71] Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. Interpretable deepfake detection via dynamic prototypes. *arXiv preprint arXiv:2006.15473*, 2020. 3

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-

reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 6

[73] Luisa Verdoliva. Media forensics and deepfakes: an overview. *arXiv preprint arXiv:2001.06564*, 2020. 3

[74] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 7

[75] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *IJCAI*, 2020. 3

[76] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 6

[77] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 3, 5, 7

[78] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 2018. 2

[79] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, 2019. 2, 3

[80] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019. 3, 4