Face Forensics in the Wild

Tianfei Zhou\textsuperscript{1}, Wenguan Wang\textsuperscript{1*}, Zhiyuan Liang\textsuperscript{2}, Jianbing Shen\textsuperscript{3,2}
\textsuperscript{1}ETH Zurich  \textsuperscript{2}Beijing Institute of Technology  \textsuperscript{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\textsubscript{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\textsubscript{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\textsuperscript{1} would have an immediate and far-reaching impact in alleviating the malicious intents of deepfakes, such as, face recognition attacks\textsuperscript{2} and fake news\textsuperscript{[36, 4]}. To help with this, several benchmarks have been established. The pioneering large-scale dataset, i.e., FaceForensics++\textsuperscript{[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\textsuperscript{[62]}. More recent datasets (e.g., DeeperForensics-1.0\textsuperscript{[37]}, DFDC\textsuperscript{[22]}, Celeb-DF\textsuperscript{[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\textsuperscript{[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.

To take the research of face forensics into a new level, we introduce a new large-scale dataset, called FFIW\textsubscript{10K}, promoting empirical study of face forgery detection in multi-person scenarios. In FFIW\textsubscript{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\textsubscript{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\textsubscript{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

\textsuperscript{*}Corresponding author: Wenguan Wang.
\textsuperscript{1}In this work, “forgery” refers to altering imagery by swapping faces.

\textsuperscript{2}Answer: The middle person in the left image and the rightmost of the back row in the right image are fake.
forged faces. ii) 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-regulated 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 as 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 DeepForensics-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.

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).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Pub.</th>
<th>#Real</th>
<th>#Fake</th>
<th>#Synthetic Methods</th>
<th>#Face Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>UADFV [79]</td>
<td>2018</td>
<td>ICASSP</td>
<td>49</td>
<td>49</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FaceForensics++ [63]</td>
<td>2019</td>
<td>ICCV</td>
<td>1,000</td>
<td>4,000</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>DFDC Preview [22]</td>
<td>2019</td>
<td>arXiv</td>
<td>1,131</td>
<td>4,113</td>
<td>~1</td>
<td></td>
</tr>
<tr>
<td>Celeb-DF [50]</td>
<td>2020</td>
<td>CVPR</td>
<td>590</td>
<td>5,639</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DeepForensics-1.0 [37]</td>
<td>2020</td>
<td>CVPR</td>
<td>50,000</td>
<td>10,000</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>FFIW$^{10K}$ (Ours)</td>
<td>2020</td>
<td>-</td>
<td>10,000</td>
<td>10,000</td>
<td>3</td>
<td>3.15</td>
</tr>
</tbody>
</table>

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 DeepForensics-1.0 [37], each fake video is randomly perturbed for augmentation, and the perturbed videos are counted as new fake videos. Ignoring perturbation, DeepForensics-1.0 only contains 1,000 unique fake videos, much fewer than the 10,000 in FFIW$^{10K}$.
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-level [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 FFIW10K enables us to make a more in-depth exploration of this new direction.

### 3. FFIW10K Dataset

Challenging datasets are catalysts for progress in the computer vision community. We therefore introduce FFIW10K to provide a better benchmark and help identify conditions under which current algorithms fail, with the hope of promoting further research efforts. Exemplars of FFIW10K are shown in Figs. 1 and 2. In the following, we present some important aspects of FFIW10K.

**Figure 2: Exemplar frames of deepfake datasets.** From left to right: 1st row shows tampered faces in FaceForensics++ [63]. Celeb-DF [50] and DeeperForensics-1.0 [37]; 2nd row shows examples in FFIW10K 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
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 FFIW10K. The network allows us to build FFIW10K 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 FFIW10K.

3.3. Dataset Features and Statistics

To offer deeper insights into FFIW10K, 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, FFIW10K 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 FFIW10K 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 FFIW10K. Specifically, a total of 50 computer science students are invited to assess the realness of synthetic videos in FFIW10K 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 FFIW10K 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, FFIW10K 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 FFIW10K 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.

Dataset Annotation. For completeness, FFIW10K 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 FFIW10K 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 accompanied 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 $G^Q$ (VGG16-based) that automatically evaluates the quality of each swapped face and hence allows us to effortlessly 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], FS-GAN[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 ground-truth 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 FFIW10K due to the domain shift between the training and testing distributions. To address this issue, we introduce domain-adversarial regularization [29] to encour-
image 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 \( F^Q \) predict both the quality score and domain label. As shown in Fig. 5(a), with all the training samples \( \{ I_i, s_i, d_i \} \), \( F^Q \) is learned by minimizing:

\[
L_Q = \sum_i L_i(s_i, s_i) - \alpha L_{CE}(d_i, d_i),
\]

where \( L_i \) and \( L_{CE} \) are \( l_1 \) and cross-entropy losses, respectively, \( s \) and \( d \) indicate the estimated quality score and domain label, respectively, and \( \alpha > 0 \). Through the domain-adversarial regularization (i.e., \( -\alpha L_{CE}(\hat{d}_i, d_i) \)), 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 \( F^Q \), 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 \( \text{FFIW}_{10K} \). For each swapped face \( \tau \), created by DeepFaceLab, FSGAN or FaceSwap, we compute its quality score through Q-Net: \( \hat{s} = F^Q(\tau) \). 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 \( \mathcal{y} \in \{ \text{fake, real} \} \). 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:

\[
\max_{\alpha \in [0,1]^K} \log p(\mathcal{y}_k | \{ \mathcal{y}_k \}_{k=1}^K, \alpha) + \log p(\alpha),
\]

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

**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 \( \tau_t \) is represented by a feature vector \( x_t \in \mathbb{R}^D \). Here we aim to learn a compact representation \( \mathbf{Y} \) for \( \Gamma \):

\[
\mathbf{Y} = F(x_1, \ldots, x_T) \in \mathbb{R}^D,
\]

where the aggregation function \( F \) can be naturally implemented as a global pooling operation (e.g., max-pooling, average pooling or log-sum-exponential pooling) 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) in manipulated face sequences. We thus propose a multi-temporal-scale feature aggregation module for more discriminative instance representation learning. Formally, suppose \( \mathbf{X} = [x_1, \ldots, x_T] \in \mathbb{R}^{D \times T} \) be the raw tracklet representation in matrix form. \( F \) is achieved by a sequence of short-term \( F^s \), long-term \( F^l \) and global \( F^g \) aggregation operations (see Fig. 6):

- **short-term aggregation**: \( \mathbf{S} = F^s(\mathbf{X}) \in \mathbb{R}^{D \times T} \),
- **long-term aggregation**: \( \mathbf{L} = F^l(\mathbf{S}) \in \mathbb{R}^{D \times T} \),
- **global aggregation**: \( \mathbf{Y} = F^g(\mathbf{L}) \in \mathbb{R}^D \).

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

**Short-Term Feature Aggregation.** We propose a densely connected dilated temporal convolution module to achieve \( F^s \) 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, \( F^s \) is a stack of \( L \) atrous convolutional layers, i.e., \( \{ f_{\text{atrous conv}}^{(i)} \}_{i=1}^{L} \), where the dilation rate \( r_i \)
is increased layer by layer. Each \( F^{\text{at}, \text{conv}}_l \) takes the concatenated features of all proceeding layers, \([S_0, S_1, \ldots, S_{l-1}] \in \mathbb{R}^{(l \times D) \times T}\) as inputs, and outputs:

\[
S_l = F^{\text{at}, \text{conv}}_l ([S_0, S_1, \ldots, S_{l-1}]) \in \mathbb{R}^{D \times T},
\]

where \( S_0 = X \). Here, \( F^{\text{at}, \text{conv}} \) 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, \( F^{\text{at}} \) finally produces a powerful, short-term descriptor \( S \) for the tracklet \( \Gamma \), by comprehensively modeling and fusing context over different local temporal scales.

In practice, each \( F^{\text{at}, \text{conv}}_l \) is implemented by: \( \text{bn-relu-conv}(1 \times 1) \)-bn-relu-conv(3 \times 3, \( r_l \))-bn-conv(1 \times 1). Here, the first \( 1 \times 1 \text{ conv} \) reduces the feature dimension to \((1 \times D)/4\) for computational efficiency, the \( 3 \times 3 \text{ conv} \) with dilation rate \( r_l \) facilitates multi-scale feature learning, and the second \( 1 \times 1 \text{ conv} \) outputs the feature \( S_l \in \mathbb{R}^{D \times D \times T} \) at \( l \)-th layer. We use \( L = 3 \) layers of dilated convolution with rates \( r \in \{1, 2, 4\} \), 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 \( F^{\text{L}} \) (Eq. (5)) over short-term feature \( S \) to learn a non-local informative representation for the tracklet \( \Gamma \). Specifically, we employ self-attention \cite{vaswani2017attention, chen2018long} to model the long-range, multi-level dependencies among temporal features in \( S \) (see Fig. 6). We first compute the normalized correlation between each pair of temporal feature vectors in \( S = \{s_1, \ldots, s_T\} \in \mathbb{R}^{D \times T} \) through pairwise dot product:

\[
A = \text{softmax}(S^\top S)
= \text{softmax}(\{s_1, \ldots, s_T\}^\top \{s_1, \ldots, s_T\}) \in [0, 1]^{T \times T}.
\]

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

\[
L = F^{\text{L}}(S) = SA + S \in \mathbb{R}^{D \times T}.
\]

Thus \( L \) encodes both the long-term \( SA \) and short-term information \( S \), with enhanced representability.

### 5.3. Attention-Based Bag Feature Aggregation

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

\[
O_V = \sum_{k=1}^K a_k Y_k \in \mathbb{R}^D,
\]

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

\[
a_k = \frac{\exp(\mathbf{w}^\top \tanh(W^\top Y_k))}{\sum_{k'=1}^K \exp(\mathbf{w}^\top \tanh(W^\top Y_{k'}))} \in [0, 1].
\]

Here \( \mathbf{w} \in \mathbb{R}^C \) and \( W \in \mathbb{R}^{D \times C} \) 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 \( a_Y \). For face forgery localization, we regard a tracklet \( \Gamma_k \) as fake if \( a_k > 0.75 \). The threshold 0.75 is determined by grid search over \( \text{FFIW}_{10K} \) val.

### 5.4. Loss Function

Given the video-level feature representation \( O_V \in \mathbb{R}^D \), 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 $\alpha_V$ to select a few most possibly tampered faces. Thus the overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\hat{I}_V, I_V) + \beta \mathcal{L}_{\text{Sparsity}}(\alpha_V). \quad (12)$$

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 $\alpha_V$, i.e., $\mathcal{L}_{\text{Sparsity}}(\alpha_V) = \| \alpha_V \|_1$. 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 $224 \times 224$ 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.

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).

We next analyze the 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.

6.1. Face Forgery Classification

We first investigate the classification performance of the approaches on $FFIW_{10K}$ test. As shown in Table 2, the video-based methods 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.

6.2. 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 consistently outperform image-based methods [63, 3, 13, 49] in terms of mAP.

<table>
<thead>
<tr>
<th>Methods</th>
<th>classification ACC (%)</th>
<th>AUC (%)</th>
<th>localization mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xception [63]</td>
<td>54.1</td>
<td>56.1</td>
<td>17.9</td>
</tr>
<tr>
<td>MesoNet [3]</td>
<td>53.8</td>
<td>55.4</td>
<td>17.7</td>
</tr>
<tr>
<td>PatchForensics [13]</td>
<td>58.9</td>
<td>61.6</td>
<td>18.9</td>
</tr>
<tr>
<td>FWA [49]</td>
<td>60.2</td>
<td>63.1</td>
<td>19.2</td>
</tr>
<tr>
<td>TSN [74]</td>
<td>61.1</td>
<td>62.8</td>
<td>21.7</td>
</tr>
<tr>
<td>C3D [70]</td>
<td>64.3</td>
<td>65.5</td>
<td>23.9</td>
</tr>
<tr>
<td>I3D [12]</td>
<td>68.8</td>
<td>69.5</td>
<td>29.7</td>
</tr>
<tr>
<td>S-MIL [47]</td>
<td>59.8</td>
<td>61.2</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>69.4</td>
<td>70.9</td>
<td>30.8</td>
</tr>
</tbody>
</table>

Table 2: Quantitative results for face forgery classification and localization on test set of $FFIW_{10K}$ (§6.1 and §6.2).
Figure 7: Visual results for face forgery localization on test set of FFIW\textsubscript{10K} (§6.2).

<table>
<thead>
<tr>
<th>Methods</th>
<th>FF++ [63]</th>
<th>DFDC Preview [22]</th>
<th>Celeb-DF [50]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xception-FF++ [63]</td>
<td>99.2</td>
<td>49.9</td>
<td>48.2</td>
</tr>
<tr>
<td>Capsule-FF++ [57]</td>
<td>96.6</td>
<td>53.3</td>
<td>57.5</td>
</tr>
<tr>
<td>Xception-c40-FF++ [63]</td>
<td>95.5</td>
<td>69.7</td>
<td>65.5</td>
</tr>
<tr>
<td>F\textsuperscript{2}Net-FF++ [62]</td>
<td>99.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Face X-ray-FF++ [46]</td>
<td>99.2</td>
<td>73.5</td>
<td>74.8</td>
</tr>
<tr>
<td>TRN-FF++ [53]</td>
<td>99.1</td>
<td>-</td>
<td>76.7</td>
</tr>
<tr>
<td>Xception-c40-FFIW\textsubscript{10K}</td>
<td>95.7</td>
<td>71.3</td>
<td>66.9</td>
</tr>
<tr>
<td>Ours-FF++</td>
<td>99.5</td>
<td>72.8</td>
<td>75.3</td>
</tr>
<tr>
<td>Ours-FFIW\textsubscript{10K}</td>
<td>99.3</td>
<td>74.1</td>
<td>78.3</td>
</tr>
</tbody>
</table>

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

At§6.2. 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\textsubscript{10K}. We train our model as well as Xception-c40 [63] on FFIW\textsubscript{10K} train and report their performance on other datasets. We see that both “Ours-FFIW\textsubscript{10K}” and “Xception-c40-FFIW\textsubscript{10K}” outperform their alternatives, i.e., “Ours-FF++” and “Xception-c40-FF++”, which are trained on FF++ train. Hence, “Ours-FFIW\textsubscript{10K}” shows superior performance on DFDC Preview and Celeb-DF. These experiment results reveal that FFIW\textsubscript{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\textsubscript{10K} test. The results are summarized in Table 4.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Variants</th>
<th>classification ACC (%)</th>
<th>AUC (%)</th>
<th>localization mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Model</td>
<td></td>
<td>69.4</td>
<td>70.9</td>
<td>30.8</td>
</tr>
<tr>
<td>Instance Feature Aggregation</td>
<td>max-pooling</td>
<td>64.5</td>
<td>66.2</td>
<td>24.6</td>
</tr>
<tr>
<td>w/o short-term (Eq. (4))</td>
<td>avg-pooling</td>
<td>63.9</td>
<td>65.7</td>
<td>29.6</td>
</tr>
<tr>
<td>w/o long-term (Eq. (5))</td>
<td></td>
<td>69.0</td>
<td>70.4</td>
<td>30.2</td>
</tr>
<tr>
<td>Bag Feature Aggregation</td>
<td>max-pooling</td>
<td>67.3</td>
<td>69.5</td>
<td>-</td>
</tr>
<tr>
<td>w/o avg-pooling (Eq. (5))</td>
<td></td>
<td>64.7</td>
<td>66.8</td>
<td>-</td>
</tr>
<tr>
<td>Loss Function (§5.3) w/o (L_{\text{sparsity}}) (Eq. (12))</td>
<td></td>
<td>68.6</td>
<td>70.3</td>
<td>28.5</td>
</tr>
</tbody>
</table>

Table 4: Ablation study on test set of FFIW\textsubscript{10K} (see §6.4).

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 \(L_{\text{Sparsity}}\). At last, we study the necessity of the sparsity constraint \(L_{\text{Sparsity}}\) 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


