Towards Benchmarking and Evaluating Deepfake Detection

Jingyi Deng, Chenhao Lin, Member, IEEE, Pengbin Hu, Chao Shen, Senior Member, IEEE, Qian Wang, Fellow, IEEE, Qi Li, Senior Member, IEEE, and Qiming Li

Abstract—Deepfake detection automatically recognizes the manipulated media by analyzing whether it contains forgeries generated through deep learning. It is natural to ask which among the existing deepfake detection approaches stand out as top performers. This question is pivotal for identifying promising research directions and offering practical guidance. Unfortunately, conducting a sound benchmark comparison of popular detection approaches based on literature results is challenging due to inconsistent evaluation conditions across studies. In this paper, our objective is to achieve a sound comparison between detection approaches by establishing a comprehensive and consistent benchmark, developing a repeatable evaluation procedure, and performing extensive performance evaluation. Accordingly, a challenging dataset consisting of the manipulated samples generated by more than 12 different methods is collected. Subsequently, we implement and evaluate 13 prominent detection approaches (comprising 11 algorithms) from existing literature, utilizing five fair-minded and practical evaluation metrics. Finally, we provide up to 882 comprehensive evaluations by training 117 detection models. The results, along with the shared data and evaluation methodology, constitute a benchmark for comparing deepfake detection approaches and measuring progress.

Index Terms—Benchmark, deepfake detection, face swapping, forensic datasets.

I. INTRODUCTION

DEEPAKE technology is a recent deep learning-based development designed to manipulate faces in both images and videos. This technology presents significant challenges to personal privacy protection and holds the potential to mislead individuals with fabricated words and deeds. Among various deepfake manipulations, face swapping stands out as the most popular manipulation for creating deepfake content. This form of manipulation is recognized as a new threat to the violation of privacy, identity, financial, legal, and even various security [1]. Consequently, deepfake detection techniques have gained significant attention to mitigate these emerging threats.

With the advent of deepfake detection approaches, several large-scale deepfake forensic datasets and benchmarks [2], [3], [4] have emerged. However, these benchmarks often fail to provide a fair and comprehensive comparison between state-of-the-art detection approaches. The most obvious issue is the inconsistent choices of training data, resulting in unfair comparisons. Although benchmarked methods are evaluated on consistent test sets, the models trained on larger, higher-quality, and more diverse deepfake datasets are likely to achieve superior evaluation results in such benchmarks. In addition, existing benchmarks often compare data scale, manipulation diversity, and perturbation diversity of proposed datasets, while lacking a quantitative analysis indicating which dataset is more challenging or suitable for training a robust deepfake detection model. Consequently, there is an urgent need to construct a fair, consistent, and systematic benchmark to assess current achievements and identify future needs in deepfake detection.

Besides, although several advanced deepfake detection methods have been proposed and proved effective, soundly quantifying the contribution of existing works poses significant challenges for the following reasons.

- First, unlike popular image classification and object detection tasks that usually use the same datasets for training to ensure a fair comparison, many deepfake detection methods undergo training on distinct datasets while being evaluated on the consistent test dataset. For example, several existing works [5], [6] employ publicly available pre-trained models rather than re-implementing these methods on the same training data for evaluation. Such inconsistency in training data introduces challenges in making fair comparisons, making it difficult to measure whether the performance contributions stem from the methods themselves or the specific training data they utilize [7], [8].
- Second, most deepfake detection methods with outstanding performance can be impractical due to the overfitting problem and poor transferability. This is particularly evident when considering the limited manipulation and

5112 IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 21, NO. 6, NOVEMBER/DECEMBER 2024
perturbation on which these methods are trained and evaluated. Predictably, these methods may experience a significant performance decrease when deployed in realistic scenarios with diverse and complicated manipulation data [9], [10].

- Third, widely used evaluation metrics, including AUC (area under the ROC curve) and accuracy, are insufficient for a comprehensive evaluation of detection methods. Essential and practical evaluation metrics, including time and space complexity, have drawn nil attention in previous research, potentially leading to the low efficiency of top detection methods for large-scale forged videos or images in realistic scenarios.

To address these limitations, this paper proposes a benchmark to conduct a fair, comprehensive, and rigorous comparison of deepfake detection approaches. The primary focus of this comparison lies in the detection of face swapping manipulation, a prevalent deepfake manipulation in existing datasets and a target of numerous popular detection methods. Specifically, we establish this benchmark on our collected standard dataset and our proposed Imperceptible and Diverse test (ID test) set, both containing data in image or video modalities. The standard dataset integrates several representative existing deepfake datasets and is used to train and evaluate detection methods. On the other hand, the Imperceptible and Diverse test (ID test) set is designed to simulate real-world media conditions. It comprises hard and diverse samples selected from public datasets and our hosted private dataset, exclusively reserved for evaluation purposes. The forged videos within the ID test set are synthesized using various manipulation approaches and are highly indistinguishable, proving challenging for both human eyes and detection algorithms.

To facilitate a comprehensive comparison in the benchmark, we begin by categorizing existing deepfake datasets and detection methods. Subsequently, we integrate several representative datasets and state-of-the-art methods to conduct evaluations. Furthermore, to provide a thorough illustration of the performance of these methods, we employ not only the commonly used detection ability measurement (e.g., Area under the ROC curve) but also four complementary evaluation metrics. These metrics are applied to measure these methods from distinct perspectives, including generalization ability, robustness, efficiency, and practicability. Through a comprehensive and quantitative analysis of the results from 882 evaluation experiments, this work presents several important findings.

- First, the forgery detection ability of all 13 popular deepfake detection approaches drops significantly on a realistic and challenging dataset, indicating a failure to meet the requirements of real-world applications.
- Second, we find that the overall performance of popular detection methods shows no significant difference under strictly uniform evaluation conditions, contrary to previous studies that asserted the superiority of one detection method over another in specific evaluation configurations.
- Third, considering detection ability, generalization, robustness, and practicability simultaneously, no one method shows comprehensive superiority over others. While certain methods exhibit specific strengths in individual evaluation metrics, more detailed discussions are provided in Section IV-G.

The remainder of this paper is organized as follows: Section II provides a review of related work in deepfake creation and detection, encompassing existing datasets, detection methods, evaluation metrics, and forensic benchmarks; Section III presents the evaluation methodology that our benchmark established on; Section IV presents evaluation results and qualitative observations; Section V discusses the limitations of our benchmark and proposes solutions; and Section VI concludes with a discussion and further work.

II. DEEPFAKE CREATION AND DETECTION

To establish a comprehensive benchmark, we first summarize the advances in deepfake detection across four dimensions, including existing popular datasets, detection methods, evaluation metrics, and forensic benchmarks. The creation methods, associated datasets, and detection approaches reviewed in this section are categorized to facilitate the subsequent establishment of a systematic benchmark. Moreover, according to our focus, this overview is primarily on deepfake creation approaches and datasets related to face swapping manipulation.

A. Deepfake Creation and Forensic Datasets

1) Deepfake Creation: Existing popular face manipulation approaches are mainly based on three types of manipulation strategies, namely autoencoder-based manipulation, GAN-based manipulation, and graphic-based manipulation.

   a) Autoencoder-based Manipulation: This manipulation employs an autoencoder to implement deepfake face swapping. The model is trained to encode the facial features of one individual in the latent space and subsequently decode these features onto another face. This process effectively results in a face swap between the two identities. As one of the most widely used manipulation strategies, a majority of forensic datasets contain this type of manipulated data. For example, UADFV [11] uses an autoencoder-based software, FakeApp [17], to generate 49 fake videos. FaceForensics++ [13] applies FaceSwap [18] to manipulate 1,000 videos. DeeperForensics-1.0 [3] designs a novel autoencoder-based framework, DF-VAE, to manipulate 10,000 videos.

   b) GAN-based Manipulation: This manipulation leverages generators trained through contests with discriminators to generate entire swapping face or face attributes in manipulated face images. Benefiting from its powerful learning ability and manipulation performance, DeepFake-TIMIT [12] employs a GAN-based face-swapping algorithm [19] to synthesize 620 fake videos. FaceForensics++, on the other hand, manipulates 1,000 videos using a two-stage face swapping method named FaceShifter [20]. Additionally, ForgeryNet utilizes DeepFakes [21], FSGAN [22], and FaceShifter [20] to generate tens of thousands of videos and images.

   c) Graphic-based Manipulation: It forges face images by modeling and deforming source face landmarks to align with
TABLE I
OVERVIEW OF POPULAR FORENSIC DATASETS

| Dataset       | Total Frames/Images | Video-level | Manipulation Method | Extra Test Set | Perturb* |
|---------------|---------------------|-------------|---------------------|----------------|----------|
|              | Real     | Fake      | Split                | AE     | GAN     | Graphic |                |                |
| UADFV [11]    | 17,329   | 16,991    | 3:1:1                | 1      | –       | –       | –              | –              |
| DF-TIMIT [12] | 34,003   | 34,023    | 3:1:1                | –      | 1       | –       | –              | –              |
| FaceForensics++ [13] | 509,914 | 1,321,408 | 5:1:1                | 1      | 1       | 3       | –              | 1              |
| DFD [14]      | 315,381  | 2,242,692 | unknown              | 1      | –       | –       | –              | –              |
| Celeb-DF-v2 [2] | 358,790 | 2,116,768 | 13:1:1               | 1      | –       | –       | –              | –              |
| DFDC-preview [15] | 488,376 | 1,783,305 | 6:1 (train:test)     | unknown | unknown | unknown | –              | 3              |
| DeeperForensics-1.0 [3] | 509,128 | 508,944   | 7:1:2                | 1      | –       | –       | ✓              | 7              |
| DFFD [6]      | 509,914  | 2,304,108 | 10:1:9               | 1      | 2       | 3       | –              | –              |
| DFDC [4]      | 5,655,501 | 29,075,744 | 5:1:1                | 2      | 3       | 1       | ✓              | 19             |
| ForgeryNet [16] | 2,848,548 | 1,054,671 | 48:3:7               | –      | 6       | –       | ✓              | 36             |

*Number of unique perturbations, regardless of perturbed parameters.

We list the most crucial information reflecting the pros and cons of datasets for training and evaluating deepfake detection models. Dataset scale and data distribution are quantified by the measurement of total frames. Dataset quality is measured by the diversity of manipulation methods, perturbation types, and the inclusion of hard examples.

the target face landmarks, often followed by a blending operation [23]. In existing forensic datasets, FaceForensics++ adopts 3D-Faceswap [24] to manipulate 1,000 videos, while DFDC [4] uses a custom frame-based morphable-mask model proposed in [23] to forge parts of videos.

2) Forensic Datasets: According to the above analysis and categorization of deepfake creation approaches, existing forensic datasets commonly consist of deepfake videos and images generated by limited manipulation strategies due to the high consumption of resources and time. For example, datasets such as UADFV, DeepFake-TIMIT, Celeb-DF [2], DFD [14], DeeperForensics-1.0, and ForgeryNet are constructed by one or two types of forgery approaches. In contrast, a few well-considered datasets, including FaceForensics++, DFFD [6], and DFDC, incorporate complete deepfake creation strategies to offer complex and practical data. A comprehensive overview of these popular forensic datasets is summarized and listed in Table I. For datasets containing multiple facial deepfake types, such as face swapping, face reenactment, attribute manipulation, and entire face synthesis, only information related to face swapping is listed here.

Additionally, to better simulate real-world scenarios and prevent detection bias, more recent datasets like DeeperForensics-1.0, DFDC, and ForgeryNet have introduced hidden or private test sets. These additional test sets ensure both manipulation diversity and the visual authenticity of fake videos. However, certain limitations of these datasets remain undisclosed. For example, while DeeperForensics-1.0 carries out a user study to ensure the visual authenticity of the fake videos, the data distribution and diversity are unknown. Conversely, DFDC and ForgeryNet provide details of their hidden test sets, whereas the authenticity of the manipulated videos is not verified for their capacity to deceive the human eye.

B. Forgery Detection

To defend the abused synthetic media, considerable efforts have been undertaken by researchers and communities these years, promoting the rise of multiple detection methods. In this section, we roughly categorize the existing forgery detection methods into intra-frame detection with several subclasses and inter-frame detection.

1) Intra-Frame Detection: Intra-frame detection involves image-level binary classification conducted on frame images. Based on the principle difference, intra-frame detection can be further subdivided into knowledge-driven detection, data-driven detection, and multi-stream-driven detection.

a) Knowledge-Driven Detection: Knowledge-driven detection incorporates domain-specific knowledge to emphasize predefined manipulation artifact clues. These clues then serve as discriminative features upon which a classification model is trained. These detection methods typically require method-specific data pre-processing to either extract artifact-related features or generate artifact-emphasized supervision information. For instance, Headpose [25] and Face-Cutout [26] leverage facial landmarks to estimate 3D head pose and implement data augmentation to explore manipulation clues, respectively. F3Net [27] adopts frequency-aware decomposed image components and local frequency statistics as discriminative features. FWA [28] uses resolution inconsistency between the face and background region as a clue. Face X-ray [9] considers the existence of blending boundary as a forgery clue, and audiovisual methods [29], [30] adopt matching correspondence between the visual and auditory modalities as a clue.

b) Data-Driven Detection: These approaches aim to design specific deep neural networks or modules tailored to deepfake characteristics. This facilitates learning multi-level discriminative features from extensive datasets comprising both real and fake images [10], [31], [32], [33], [34]. Typically, these approaches rely on image-level binary labels as supervision, and discussions on intuition, as well as decision mechanisms, commonly follow the training phase.

c) Multi-Stream-Driven Detection: These approaches leverage the strengths of both aforementioned strategies by often developing multiple streams. These streams simultaneously learn interpretable predefined features and latent features, which are subsequently integrated to enhance forgery detection capabilities [35], [36], [37], [38].

2) Inter-Frame Detection: Inter-frame detection, also referred to as video-level detection, aims to learn temporal artifacts
from consecutive manipulated video frames. The majority of existing works are based on architectures such as CNN-RNN [39], [40], [41], [42], CNN-CNN [43], or 3D CNN [3], [44]. These architectures successively or concurrently extract intra-frame spatial and inter-frame temporal artifacts [39], [40], [41]. Furthermore, some incorporate inter-frame domain knowledge, such as optical flow [43], [45], lipreading [46], and biological signal [11], [45], as auxiliary information to enhance deepfake detection.

C. Evaluation Metrics and Forensic Benchmarks

Various metrics, including AUC [9], [25], [28], precision [10], and accuracy [13], have been widely employed in the evaluation of deepfake detection. Apart from these conventional metrics, DFDC [4] has introduced the Weighted PR, addressing the class imbalance between fake and real videos in organic traffic by assigning a weight to false positives of precision. Despite the widespread use of these metrics, they often fall short of providing a comprehensive understanding of the efficiency of detection methods in real-world scenarios. This limitation stems from their primary focus on predictive performance metrics, neglecting other computational evaluation aspects.

Regarding existing benchmarks, Celeb-DF [2] has evaluated ten methods using publicly accessible pre-trained models, while DeeperForensics-1.0 [3] has introduced an evaluation of five methods not specifically designed for deepfake detection. Initiatives such as the FaceForensics Benchmark [47], DeepFake Detection Challenge (DFDC) [48], and ForgeryNet [16] establish online benchmarks or forensic challenges, requiring participants to submit prediction results or well-trained models. Despite these efforts, most of these benchmarks fail to conduct a fair comparison due to the absence of models trained with consistent data and in a consistent environment.

III. EVALUATION METHODOLOGY

The increased accessibility of face manipulation through mature algorithms and widely available software has heightened the demand for practical forensic approaches. However, evaluating the applicability of existing popular deepfake detection methods in real-world scenarios poses a challenge due to the absence of fair and adequately evaluated benchmarks. To address this concern and promote further research in this field, we propose a fair, comprehensive, and strict benchmark by integrating 7 popular forensic datasets and 13 representative forgery detection methods. According to our categorization in Section II, this benchmark investigates the detection capabilities of each type of state-of-the-art forensic method against different types of manipulation strategies. Moreover, we construct an Imperceptible and Diverse Test (ID Test) Set and apply 4 complementary graphs to thoroughly evaluate the generalization ability, robustness, and practicability of the benchmarked detection methods.

A. Evaluation Datasets and Algorithms

1) Standard Dataset: The standard dataset consists of 7 sub-datasets, which respectively comprise real and manipulated data originating from autoencoder-based UADFV, Celeb-DF, DF-1.0, GAN-based DF-TIMIT (higher quality), and mixed-manipulation-based FF++ (Raw), DFDC, and ForgeryNet. Each sub-dataset is split into training, validation, and test sets to support model training and evaluation. In our benchmark, to better represent the differences of our integrated existing datasets and provide a reasonable data foundation for fair benchmark evaluation, the extracted data used for model training and evaluation satisfy three factors: 1) the total scale of extracted data of different sub-datasets is proportional to their original base dataset scale; 2) the proportion of extracted fake data and real data of each sub-dataset is identical to its original base dataset; 3) the total scale of extracted frame data and the total scale of extracted consecutive frame data are approximately consistent, which is guaranteed to support a fair evaluation between frame-level methods and video-level methods.

Accordingly, following FaceForensics++-v1 [13], taking the extracted data scale of 2,527,384 frames as a scale baseline, the extracted data size of other datasets proportionally doubles or decreases. For dataset splitting, the videos of each sub-dataset are first split into training, validation, and test set according to their default base dataset setting if it is released. Otherwise, we carry out a reasonable split as illustrated in Table I. Afterward, frames and consecutive frames are randomly extracted with an amount ratio of 14:1:1 from the training, validation, and test videos.

2) Imperceptible and Diverse Test (ID Test) Set: To explore the detection ability of forensic approaches when confronting the threats of fake videos with high visual authenticity and diverse manipulations, we construct a high-quality Imperceptible and Diverse test (ID test) set by integrating the hard (high imperceptible) examples from our benchmarked 7 public datasets and our hosted private dataset. The samples from our private dataset consist of self-generated fake examples manipulated by GAN-based FSGAN [22] and autoencoder-based MegaFS [49] approaches, of which the original data are images from CelebA [50] and raw videos from FaceForensics++ respectively. To generate these samples, the officially released pretrained FSGAN model [51] and our self-trained MegaFS model are utilized. The training of the MegaFS model follows the parameters and settings as described in the original paper.

To guarantee the high visual authenticity of hard samples in the ID test set, the fake samples from public datasets are selected by a two-phase selection pipeline, namely detection model selection and user perception selection. Since the dataset size of the private dataset are relatively small, fake samples from this dataset are merely selected by user perception selection. The detection model selection aims to retain falsely accepted fake examples with high confidence. Specifically, we use the data-driven detection method, Xception, and the knowledge-driven method, Face-X-ray, trained on each sub-dataset of the standard dataset to filter corresponding in-domain fake samples. These fake samples are in a video format and are evaluated frame by frame, and their average scores are set as prediction fake scores. The first two-thirds of fake samples with the lowest fake scores predicted by the aforementioned methods are retained and form the intersection set for the next phase of selection.
In the next user perception selection phase, we engage 30 participants, who are experienced in discriminating deepfake forensics, to carry out a blind experiment. Through this process, we aim to preserve high-quality fake videos with imperceptible manipulation clues observed by human eyes. Concretely, equivalent amounts of real samples are first mixed with the filtered fake samples to form a user perception set. Then, referring to [52], the participants are asked to observe these samples and respond to the statement “This video clip looks real to me.” and give a five-level score (1-disagree, 2-disagree, 3-don’t know, 4-agree, 5-strongly agree). Based on the scores provided by each participant, a comprehensive user perception score is assigned to each fake sample. According to this score, we keep a certain quantity of fake samples with the highest scores. Finally, through this selection pipeline, 976 fake videos and 2348 real videos are filtered to construct the ID test set.

To construct a balanced ID test set for fair evaluation, two factors are guaranteed when implementing frame and consecutive frame extraction. On the one hand, the data size of frame-level data and video-level data extracted from different types of manipulation videos is approximately identical, enabling the fair evaluation between frame-level and video-level detection methods. On the other hand, the data size of different manipulation is approximately equivalent, allowing fair evaluation between detection methods trained by different manipulation data. As such, overall 25,697 fake images and 25,697 real images are extracted from videos of the ID test set. The extracted data distribution is shown in Table II and Fig. 1. As is shown, the ID test set consists of diverse forensic samples, manipulated by more than 13 approaches, achieving full coverage of manipulation strategies proposed in Section II-A1.

Moreover, to explore the robustness of detection methods, we construct a perturbed ID test set by applying 6 types of common perturbations to the samples. These perturbations include image-level perturbations such as color contrast change, color saturation change, Gaussian blur, JPEG compression, and white Gaussian noise, as well as video-level perturbation of H.264 compression. The strengths of perturbations are randomly determined within a rational range to ensure the visual perception of the media data. This setting of random strength is compatible with the data post-processing operations brought by the online social platforms, on which the forgery data is usually uploaded and transmitted. Given that most online social platforms manipulate the uploaded images in a lossy fashion, it is challenging to precisely determine the specific type and strength of the imposed lossy operations due to their adaptive nature and complex enhancement filtering [53], [54]. Thus, imposing random strength of perturbations on the ID test set is more suitable to simulate the adaptive data noise in realistic scenarios and able to observe the model robustness under such conditions. Specifically, the color contrast and color saturation are adjusted by the Pillow library [55] with enhancement factors in the range of [0.5, 2.5] and [0.0, 2.5], respectively, Gaussian blur is applied with a kernel size ranging from [10, 20], JPEG compression is implemented with the image quality in the range of [30, 95], white Gaussian noise is applied with the mean and variance in the range of [−0.3, 0.3] and [0, 1], and video compression is employed using a constant rate quantization parameter equal to 23 and 40. Fig. 2 presents representative examples of each perturbation.

3) Forensic Detection Algorithms: Similar to the existing benchmark studies [7], [56] in other fields, our benchmark evaluates several representative deepfake detection methods extensively used for comparison in previous works. Specifically, we evaluate 13 declared state-of-the-art forensic detection approaches (11 algorithms), covering the overall categories described in Section II-B, as shown in Table III. For intra-frame detection algorithms, we choose 3 knowledge-driven algorithms, including Headpose [25], Face X-ray [9], and FWA [28], 5 data-driven algorithms, such as Xception [13], Mesonet [31], Patch-forensics [10], FFD [6] and Multiple-attention [5], and 1 multi-stream-driven algorithm of M2TR [34]. For inter-frame detection algorithms, we evaluate Convolutional LSTM [39] and LRNet [45]. Generally, for the methods proposed with multiple

| Table II: Overview of the Data Distribution According to Manipulation Strategy in the ID Test Set |
|-------------------------------|
| AE | GAN | Graphic | Unknown |
|---|---|---|---|
| Video | 522 | 202 | 22 | 230 |
| Image | 10,514 | 10,778 | 2171 | 2234 |

“Unknown” refers to data lacking relevant information of manipulation type from the source dataset.
TABLE III
OVERVIEW OF EVALUATED FORENSIC DETECTION ALGORITHMS

| Methods      | Category      | OS          | Data Pre     | Backbone                  | Param(M) | GFLOPs | Infer T(ms) |
|--------------|---------------|-------------|--------------|---------------------------|----------|--------|------------|
| HeadPose     | Frame-Know    | yes         | no           | SVM                       |          |        | 159.70     |
| FWA-resnet50 | Frame-Know    | part        | detect + add | Resnet50                  | 25.56    | 8.24   | 101.21     |
| Face X-ray   | Frame-Know    | no          | detect + add | HRNet-W48-C               | 77.47    | 42.58  | 35.62      |
| Xception     | Frame-Data    | part        | detect       | XceptionNet               | 20.81    | 16.84  | 6.05       |
| MSesonet-4   | Frame-Data    | part        | detect + align | 4-layer Conv          | 0.28     | 0.12   | 5.11       |
| Meosolception-4 | Frame-Data | part        | detect + align | 2-Inception+2-Conv | 0.28     | 0.11   | 7.31       |
| Patch ResNet Layer1 | Frame-Data | yes        | detect       | Resnet18                  | 0.13     | 2.10   | 0.73       |
| Patch Xception Block2 | Frame-Data | yes        | detect       | XceptionNet              | 0.19     | 3.34   | 1.17       |
| FFD          | Frame-Know    | yes         | detect + add | XceptionNet + Reg. Map    | 20.82    | 16.84  | 6.04       |
| Multiple-attention | Frame-Know | part    | detect + align | EfficientNet-b4         | 18.83    | 6.80   | 25.48      |
| M2TR         | Frame-Multi   | yes         | detect + add | Eff-b4+Transformer+Frequency Filter | 18.61   | 2.92   | 34.18      |
| Conv LSTM    | Video         | no          | detect       | InceptionV3 + LSTM        | 30.36    | 229.48 | 221.64     |
| LKNet        | Video         | yes         | detect + add | RNN                       | 0.17     | 0.004  | 1.87       |

We list their category, open-source status (OS), required data pre-processing procedure (Data Pre), adopted backbone in our experiment, number of parameters (Param), GFLOPs, and inference time (Infer T). Categories include frame-level knowledge-driven method (Frame-Know), frame-level data-driven method (Frame-Data), frame-level multi-stream-driven method (Frame-Multi), and video-level method (Video). The data pre-processing procedure includes face detection (detect), face alignment (align), and method-specific additional operation (add).

backbones, we evaluate the model with the best-performing backbone.

a) Pre-Processing: Deepfake data pre-processing includes common data pre-processing and method-specific data pre-processing. Common data pre-processing refers to the necessary operations used in the majority of detection methods, while method-specific data pre-processing involves additional algorithm-specific operations. These algorithm-specific operations commonly generate additional supervision or extract discriminative features to guide model learning. In our experiments, we use the image processing toolkit Dlib [57] and OpenCV-Python [58] for implementing common data pre-processing operations, including frame extraction, face cropping, and face alignment. Method-specific pre-processing operations for HeadPose, FWA, Face X-ray, and FFD algorithms are performed following the original presentations and implementations. The following section elaborates on the method-specific data pre-processing pipelines for algorithms without open-sourced codes.

Face X-ray requires additional mask supervision, i.e., Face X-ray, to guide the model in learning the blending boundary of fake images. Specifically, Face X-ray is defined as an image $I$ with

$$I_{i,j} = 4 \cdot M_{i,j} \cdot (1 - M_{i,j})$$  (1)

where $M$ is the soft non-binary mask delimiting the manipulated region, and the subscript $(i, j)$ indicates the pixel location. Following the original paper, we generate the ground-truth Face X-ray for fake images. First, the binary difference mask can be generated by computing the absolute element-wise difference between the manipulated image and its paired target face image. Subsequently, a Gaussian blur followed by normalization is applied to generate the soft non-binary mask. Finally, we use the above formulation to generate Face X-ray.

FWA adopts a self-supervised learning strategy to generate negative examples during the model training phase. In line with its methodology, we initially align the face image to multiple scales and randomly select an aligned image. Subsequently, Gaussian blur is applied to the aligned image, followed by sequential affine warping back to the original size. Finally, we randomly change the image color information and paste the entire face region or landmark region to the original frame image to generate the negative example.

FFD requires the ground-truth binary modification mask for computing the attention map loss. Following the paper, we compute the absolute pixel-wise difference of RGB image pairs between manipulated images and their corresponding target face images. These are then converted to grayscale images, and binary masks are derived by thresholding the normalized grayscale images.

b) Implementation Details: To ensure fair comparisons, all the selected algorithms are retrained on the consistent training set, i.e., the standard dataset in our benchmark. For the majority of these algorithms, reimplementation is necessary as open-source training codes were not available. Conversely, for the four algorithms with accessible source codes, as outlined in Table III, we directly adopt the provided implementations. In the subsequent part, we summarize the benchmarked detection methods and the detailed parameter settings applied in our experiment. In general, hyper-parameters in each method are assigned initial values. Certain hyper-parameters, defined by iteration numbers, such as the iteration number for learning rate decay [9], [28], [31] and network warm start [9], are adjusted considering the total iteration number is related to the training data size when given a fixed batch size. We set the initial values of such hyper-parameters as adaptive values proportional to the correlation between our training data size and the original training data size. For other hyper-parameters that remained independent of iteration numbers, such as batch size and initial learning rate, we retain their values as specified in the original papers. Subsequently, we tune the network parameters and hyper-parameters for all the benchmarked methods until convergence was achieved on the validation datasets. Afterward, we adopt the models with the best validation performance to conduct benchmark evaluations.
Face X-ray [9] is an intra-frame level knowledge-driven detection method. This method applies Face X-ray, the blending boundary produced by the blending procedure in face manipulation, as an interpretable artifact clue to guide the model in learning additional supervised information for detection. The original paper uses the self-supervised BI dataset along with public datasets for model training. In our experiment, we exclusively adopt the public dataset to train models, ensuring a fair evaluation of benchmarked datasets. We adopt HRNet-W48-C [59] as the backbone network in the experiment and set the batch size to 32. The initial learning rate is set as 0.0002 using the Adam optimizer, identical to the original paper. The number of warming start iterations, fine-tuning iterations, and learning rate decay iterations are set as adaptive values in proportion to the correlation between our training data size and the original training data size. Then parameters are tuned until convergence is achieved on validation datasets.

FWA [28] is classified as an intra-frame level knowledge-driven method. By exploring the artifacts introduced by the affine transform procedure in the deepfake generation pipeline, this method leverages the resolution inconsistency between the manipulated face region and its surrounding region as an interpretable artifact clue. To impose the detection model to focus on this artifact, FWA adopts self-supervised learning to generate negative examples for training. In our experiment, we utilize half of the training positive samples from each dataset to generate these negative examples. We use ResNet50 as the detection model and set the batch size to 64. The number of learning rate decay steps, fine-tuning steps, and hard mining steps are set to adaptive values proportional to the correlation between our training data size and the original training data size. Additionally, following the experimental setup in the original paper, the learning rate starts from 0.001 in the fine-tuning stage and begins at 0.0001 in the hard mining stage, with a decay rate of 0.95.

HeadPose [25] is considered as an intra-frame level knowledge-driven method. It adopts facial landmarks to predefined an interpretable forgery clue. Specifically, the difference in 3D head poses estimated by facial landmarks between the entire face region and the central face region are extracted as discriminative features. Subsequently, a Support Vector Machine (SVM) classifier is leveraged to learn this difference and distinguish the original images from the DeepFake images.

MesoNet-4/MesoInception-4 [31] is an intra-frame level data-driven method. By analyzing images at a mesoscopic level, this method introduces two networks, MesoNet-4 and MesoInception-4. MesoNet-4 is a shallow network with a sequence of four layers of successive convolutions and pooling, followed by a dense network with one hidden layer. MesoInception-4 maintains a similar network structure which replaces the first two convolutional layers with an inception module with dilated convolutions. In our experiments, we set the batch size to 75 and use the Adam optimizer during the training process. The learning rate starts from $10^{-3}$ and is divided by 10 every adaptive iterations down to $10^{-6}$.

Patch ResNet-Layer1/Patch Xception Block2 [10] belongs to the intra-frame level data-driven method. This method analyzes images based on patch-level predictions by truncating ResNet and Xception after intermediate blocks. In our experiments, we adopt patch-level labels and predictions to calculate losses, forcing the model to learn local features in the training phase. In the testing phase, we aggregate patch-level predictions in an average manner to obtain the image-level predictions and then calculate the image-level losses and metrics. Following the original paper, we set the batch size to 32, consisting of 16 real and 16 fake images, and use the Adam optimizer with default parameters and learning rate.

Xception [13] is categorized as an intra-frame level data-driven method. This method adopts XceptionNet [60] as the backbone for binary classification. Similar to the original paper, we set the batch size as 32 and train the network with a learning rate of 0.0002 using the Adam optimizer.

FFD [6] can be considered as an intra-frame level knowledge-driven method. This method applies an attention mechanism to detect and localize manipulation regions. The authors propose two types of attention-based layers, namely the manipulation appearance model and direct regression, to guide the network to focus on discriminative regions. Meanwhile, three types of loss functions are proposed to supervise the learning progress. In our implementation, we adopt the XceptionNet [60] as the backbone, direct regression as the attention-based layer, and supervised learning as the attention map learning strategy to train models.

Multiple-attention [5] is an intra-frame level knowledge-driven method. This method considers deepfake detection as a fine-grained classification problem and proposes a multi-attentional deepfake detection network. This network includes three key components: an attention module generating multiple attention maps to assist in exploring local discrimination, densely connected convolutional layers enhancing subtle texture artifacts in shallow feature maps, and bilinear attention pooling guided by the attention maps to aggregate the low-level textural features and high-level semantic features. Moreover, this method designs a regional independent loss to learn multiple attention maps and applies the AGDA mechanism to force the attention to mine more useful information. Following the original paper, we use EfficientNet-b4 [61] as the backbone network and set the $SL_t$ and $SL_a$ as L2 and L5, respectively.

Conv LSTM [39] is an inter-frame level detection method. It takes consecutive frame images as input, adopting InceptionV3 and LSTM to extract frame-level features and identify temporal anomalies. To optimize performance, diverging from the original paper of using consecutive frame images as input, we incorporate face detection in the data pre-processing pipeline. Consequently, we use 20 consecutive face images as input. Moreover, we train our model using the Adam optimizer with an initialized learning rate of 1e-5, and we set the batch size to 4.

LRNet [45] is an inter-frame level detection method. It models temporal characteristics on precise biological signal-based geometric features, e.g., facial landmarks, by adopting a two-stream Recurrent Neural Network (RNN). To capture precise landmarks, they devise a calibration module adopting optical flow calculation algorithms to improve the discrimination and stability of landmark features. This method adopts
geometric features of 60 consecutive frames as input in official implementation. Different from such official implementation, we adopt 20 consecutive frames as input in our experiment. This setting aims to keep comparison fairness between inter-frame level detection methods. Moreover, different from the dataset split in the original paper, the default dataset split in our paper is used in our experiments. Other experiment settings are identical to the original paper.

M2TR [34] is an intra-frame level multi-stream-driven detection method. It adopts a two-stream architecture, where one stream utilizes a multi-scale transformer to capture forgery patterns in the spatial domain, and the other stream utilizes frequency filters to filter out forgery clues in the frequency domain. Finally, through a cross modality fusion block, the complementary forgery features are fused for classification. In our experiment, consistent with the default setting of official implementation, we merely adopt classification loss to train the model, and other settings are identical to the statements of the paper.

To ensure the implementation correctness of these detection algorithms, we have conducted verification experiments for methods lacking open-sourced codes or with only partial availability. This is done to guarantee consistency between our results and those reported in the original papers. Comparative results are presented in Table IV. For Conv LSTM, where the test data from the original paper was unavailable, we used FF++/DF for testing and observed better results. For other methods, we performed validation on test data with a similar domain distribution to that of the original studies. Due to the unavailability of identical frame images, we randomly extract frame images from each test dataset to ensure a fair comparison. Our re-implementation achieved better or comparable results for detection approaches such as Face X-ray, FW-resnet50, Xception, MesoInception-4, Multiple-attention, and Conv LSTM. However, discrepancies were noted for FW-resnet50 (UADFV) and Xception. Specifically, FW-resnet50 achieved superior performance on high-quality negative examples of DeepfakeTIMIT but performed poorly on low-quality UADFV, suggesting a strong reliance on the image quality of the generated training samples may exist. Regarding Xception, a similar performance gap observed in our experiment was also mentioned in [31], where the authors reported achieving only 96.1% accuracy after fine-tuning Xception.

### B. Proposed Evaluation Metrics

In response to the limitation of evaluation metrics, we apply the commonly adopted AUC and four complementary graphs to comprehensively analyze the detection ability, robustness, practicability, and efficiency of forensic algorithms.

1) **AUC (Area Under the ROC Curve):** AUC is the most commonly adopted deepfake detection metric since it is insensitive to data imbalance. It represents the area under the ROC curve and the ROC curve plots the relationship of False Positive Rate (FPR) vs. True Positive Rate (TPR) of a deepfake detection model at all classification thresholds. Typically,

\[
TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN},
\]

where TP, FN, FP, and TN are true positive, false negative, false positive, and true negative. Deepfake detection methods generally utilize two types of AUC, i.e., frame-level AUC and video-level AUC, to report model performance. Frame-level AUC adopts image predictions to compute AUC and video-level AUC uses the video-level predictions derived from the average or mode of frame predictions within a video to compute AUC. In this paper, we compute frame-level AUC for image-level methods, and for video-level methods, we compute segment-level AUC by averaging image predictions within a video segment.

2) **Perturbation Versus AUC:** The perturbation vs. AUC graph provides an assessment of the robustness of forensic classifiers when facing different types and degrees of perturbed data in realistic scenes. This graph also enables an observation of the impact of various perturbations on model performance. To generate such a graph, we compute the AUC of models on each type of perturbed data in ID test set. Therefore, models with smooth performance fluctuations on perturbed data compared with their original performance would be more robust.

3) **FLOPs Versus AUC:** The FLOPs (Floating Point Operations) vs. AUC graph provides insights into the practicability of forensic classifiers, where FLOPs is a commonly used metric to calculate the computational complexity of deep learning models. From this graph, the model with a high AUC score and low FLOPs possesses both excellent forgery detection ability and an optimal network architecture with minimal computational requirements. This efficiency in computing power makes it applicable in real-world application environments.

4) **Number of Parameters Versus AUC:** The correlation between the number of parameters and AUC provides a deeper

---

**Table IV**

| Method                  | Test Dataset | Metric | Reported Result | Our Result |
|-------------------------|--------------|--------|----------------|------------|
| Face X-ray              | FF++/DF      | AUC    | 99.17 [9]      | 99.4       |
|                         | FF++/FS      | AUC    | 99.20 [9]      | 99.8       |
| FW-resnet50             | DeepfakeTIMIT| AUC    | 87.4           | 99.1       |
|                         | UADFV        | AUC    | 79.0           | 57.3       |
| Mesolnception-4         | paper released data | ACC | 91.7 [31]      | 91.2      |
|                         |              |        |                |            |
| Xception                | FF++/DF      | AUC    | 99.59 [13]     | 96.7       |
| Multiple-attention      | FF++         | AUC    | 99.29 [5]      | 99.4       |
| Conv LSTM               | Unavailable data | ACC | 96.7 [39]       | 97.9       |

FF++/DF and FF++/FS refers to FF++/Deepfakes and FF++/FaceSwap.
The reported evaluation results are shown in Table V, in which the results of frame-level methods and video-level methods are measured by the frame-level AUC and segment-level AUC. The segment-level AUC is computed by averaging the frame-level predictions over the frames in a video segment. From the results, we observe that most deep learning-based approaches achieve superior performance under the in-domain evaluation setting. Despite video-level methods adopting more samples for predictions compared to image-level methods, they do not show overall superiority over frame-level methods when their test data scale is aligned. Among deep learning-based methods, M2TR gains the best average AUC score of 99.1% (99.4%), while FWA-resnet50 exhibits the poorest performance of 63.6% (67.0%). The performance gain of M2TR mainly benefits from its multi-task training framework and its emphasis on multi-modal and multi-scale discriminative feature exploration. This design assists the model in capturing comprehensive discrepancies. In contrast, FWA-resnet50 relies exclusively on self-supervised data, emphasizing simple predefined artifacts, leading to the exploration of limited discrepancy features. In addition to method comparisons, we can see that models trained on DFDC and ForgeryNet generally exhibit lower performance. This can be attributed to the introduction of unseen perturbations in their test dataset.

Moreover, our evaluation results reveal that the significant performance gap declared in previous non-strictly evaluations has been overstated. Given this observation, model comparison at a specific training configuration cannot fully demonstrate the superiority of one model over another. For example, the performance gap on UADFV of 14.1% (84.3% (Meso4) / 98.4% (FFD)) in [6] decreases to 1.4% (97.7% (Meso4) / 99.1% (FFD)) in our experiments. This indicates that a rigorous benchmarking approach is crucial for fair evaluation.

A. Evaluation Results of Forensic Detection Ability

In this section, we evaluate the intra-domain forensic detection ability of our benchmarked methods by training and testing models on each sub-dataset of the standard dataset. The evaluation results are shown in Table V, in which the results of frame-level methods and video-level methods are measured by the frame-level AUC and segment-level AUC. The segment-level AUC is computed by averaging the frame-level predictions over the frames in a video segment. From the results, we observe that most deep learning-based approaches achieve superior performance under the in-domain evaluation setting. Despite video-level methods adopting more samples for predictions compared to image-level methods, they do not show overall superiority over frame-level methods when their test data scale is aligned. Among deep learning-based methods, M2TR gains the best average AUC score of 99.1% (99.4%), while FWA-resnet50 exhibits the poorest performance of 63.6% (67.0%). The performance gain of M2TR mainly benefits from its multi-task training framework and its emphasis on multi-modal and multi-scale discriminative feature exploration. This design assists the model in capturing comprehensive discrepancies. In contrast, FWA-resnet50 relies exclusively on self-supervised data, emphasizing simple predefined artifacts, leading to the exploration of limited discrepancy features. In addition to method comparisons, we can see that models trained on DFDC and ForgeryNet generally exhibit lower performance. This can be attributed to the introduction of unseen perturbations in their test dataset.

Moreover, our evaluation results reveal that the significant performance gap declared in previous non-strictly evaluations has been overstated. Given this observation, model comparison at a specific training configuration cannot fully demonstrate the superiority of one model over another. For example, the performance gap on UADFV of 14.1% (84.3% (Meso4) / 98.4% (FFD)) in [6] decreases to 1.4% (97.7% (Meso4) / 99.1% (FFD)) in our experiments. This indicates that a rigorous benchmarking approach is crucial for fair evaluation.

Understanding of practicability, in which the number of parameters quantifies learnable parameters in a deep learning model. From this graph, classifiers with outperformed AUC and fewer parameters require less memory consumption, which is a crucial strength in practical scenes.

5) Inference Time Versus AUC: The inference time vs. AUC graph reveals the efficiency of forensic classifiers, where the inference time is estimated by testing a single face image or a single image segment consisting of sequential face images. Considering the large scale of media data in real-world applications, deepfake detection classifiers should be efficient and provide smooth experiences. Therefore, models with shorter inference times and higher AUC scores are more practical for deployment.

C. Environment

For fair evaluation, we conduct all the experiments and evaluation in a uniform environment, with NVIDIA 2080Ti GPU and 128 GB of RAM.

IV. Evaluation Results and Discussions

We present comprehensive evaluations and discussions of the forgery detection ability, generalization ability, robustness, practicality, and efficiency/effectiveness trade-off of our benchmarked forgery detection methods in Sections IV-A, IV-B, IV-C, IV-D, and IV-E respectively. Moreover, we provide the class activation maps of evaluated models to explore their classification decision mechanisms in Section IV-F. Finally, we report some conclusive key findings in Section IV-G. Among these evaluations, the forgery detection ability evaluation is performed on the test sets of the standard dataset, and others are performed on our ID test set. Due to the space restriction, when plotting the graphs of complementary metrics illustrated in Section III-B for a specific detection method, the average AUC score of the models trained on different training sets is adopted.

The results for HeadPose on Celeb-DF, DFDC, and ForgeryNet are not reported due to challenges in training SVM models on such large-scale datasets. The last column presents the average AUC for each row, where results without parentheses are the average AUC across all datasets, and those with parentheses are averages across commonly trained datasets that exclude DFDC and ForgeryNet.
including training and evaluating models on datasets with a consistent configuration, is necessary to establish an objective and fair measurement.

Furthermore, considering that certain detection methods are also applicable to face reenactment manipulation, we conduct in-domain evaluations on FF++/NeuralTextures and FF++/Face2Face [62] as well. From the results of Table VI, we can conclude that most methods exhibit comparable and superior in-domain detection abilities for face reenactment manipulation as they do for face swap manipulation.

B. Evaluation Results of Generalization Ability

In this section, algorithms are trained on the standard dataset and evaluated on our ID test set to perform cross-domain evaluations. As the results shown in Table VII, since the data in the ID test set is diverse and high-quality enough to simulate real-world scenarios, significant drops in detection performance of all 13 deepfake detection approaches can be observed. This suggests that the existing methods are still far from the expectations for real-world deployment.

Specifically, M2TR achieves the best average AUC score of 67.9% (64.1%) among frame-level detection methods, demonstrating that incorporating complementary cross-modality features and comprehensive multi-scale local features improves generalization ability. Moreover, Multiple-attention, Patch Resnet Layer1, and Patch Xception Block2 gain relatively top performance variations, as indicated by a population standard deviation of 0.61 (0.53), suggesting that replacing face images with low-dimension geometric features as model inputs assists in training a more robust model. This observation suggests that the data-driven frame-level methods are more inclined to learn high frequency, low-level clues. Notably, LRNet exhibits minimal performance drop is due to the cross-domain data distribution and 10.6% performance drop is due to the high-quality sample mining. Therefore, we can conclude that the performance drop on the ID test set can be primarily attributed to the multi-domain distribution, with a partial contribution from imperceptible data artifacts. This validation experiment indicates that evaluating algorithms on the ID test set can effectively reflect their generalization ability. Furthermore, based on this perspective, the results of average AUC among different detection methods for datasets shown in Table VII lead to the conclusion that the superior generalization ability of DFDC and ForgeryNet on the ID test set is due to their inclusion of more diverse manipulation data.

C. Evaluation Results of Detection Robustness

Detection robustness is evaluated on our perturbed ID test set and measured by the AUC vs. perturbation graph, as shown in Fig. 3. The average performance is calculated across two sets of models: one trained on each sub-dataset of the standard dataset and the other trained on each commonly trained sub-dataset (sub-datasets of standard dataset except for DFDC and ForgeryNet). The results reveal that all methods experience varying levels of performance fluctuations when exposed to perturbations. Notably, white Gaussian noise significantly degrades the performance of most models. Perturbations influencing the high frequency content of images (i.e., Gaussian blur and JPEG compression) exhibit minimal impact on the performance of most knowledge-driven frame-level methods (e.g., Face X-ary, FFD, and WFA-resnet50) compared to data-driven frame-level methods. This observation suggests that the data-driven frame-level methods are more inclined to learn high frequency, low-level clues. Notably, LRNet exhibits minimal performance variations, as indicated by a population standard deviation of 0.61 (0.53), suggesting that replacing face images with low-dimension geometric features as model inputs assists in training a more robust model. This observation is intuitive, given that geometric features, especially facial landmarks, exclusively capture facial structure information and remain resilient to the degradation of image quality. In addition, the curve of M2TR consistently outperforms other methods, underlining its superior generalization ability.
TABLE VII
RESULTS OF THE GENERALIZATION ABILITY OF DETECTION METHODS MEASURED BY FRAME-LEVEL AUC AND SEGMENT-LEVEL AUC

| Train Test | UADFV | DF-TMIT | Celeb-DF | DF-1.0 | FF++/DF | FF++/FaceShifter | DFDC | ForgeryNet | Average AUC% |
|------------|-------|---------|----------|--------|---------|-----------------|------|-------------|--------------|
| Face X-ray | 53.9  | 52.1    | 64.3     | 66.4   | 54.8    | 59.3            | 54.9 | -           | 57.9(57.9)   |
| FWA-ResNet50 | 54.7  | 55.1    | 49.9     | 54.3   | 49.8    | 54.3            | 50.0 | 53.2        | 50.0(52.5)   |
| HeadPose   | 52.1  | 48.3    | -        | 50.4   | 51.8    | 54.0            | 50.7 | -           | 51.2(51.2)   |
| MesoNet-4  | 57.2  | 59.4    | 59.9     | 47.4   | 53.7    | 57.6            | 48.6 | 61.6        | 56.8(54.8)   |
| MesoInception-4 | 60.0  | 55.8    | 59.6     | 48.3   | 54.5    | 63.0            | 49.8 | 60.5        | 54.7(55.8)   |
| Patch ResNet Layer1 | 57.1  | 55.9    | 54.1     | 54.5   | 60.0    | 64.9            | 57.7 | 59.8        | 56.7(57.7)   |
| Patch Xception Block2 | 53.3  | 53.5    | 56.4     | 53.2   | 60.3    | 63.4            | 64.0 | 52.8        | 58.3(57.7)   |
| Xception   | 55.9  | 61.9    | 52.9     | 59.8   | 61.6    | 52.1            | 48.0 | 56.3        | 57.0(56.0)   |
| FFD        | 53.9  | 63.4    | 62.1     | 64.3   | 57.8    | 57.4            | 43.5 | -           | 57.4(57.4)   |
| Multi-attention | 52.0  | 54.0    | 64.1     | 59.3   | 54.8    | 54.5            | 57.4 | 74.7        | 69.3(68.3)   |
| M2TR       | 61.1  | 66.6    | 63.9     | 60.1   | 73.7    | 68.9            | 54.4 | 79.5        | 63.4(67.9)   |
| Conv LSTM  | 56.3  | 57.7    | 59.9     | 50.0   | 62.9    | 55.5            | 39.9 | 52.2        | 50.1(53.6)   |
| LRNet      | 52.3  | 42.0    | 55.4     | 51.8   | 61.3    | 56.4            | 53.8 | 54.0        | 61.8(54.3)   |
| Average AUC% | 55.3  | 55.8    | 58.5     | 55.3   | 58.2    | 58.5            | 51.7 | 60.4        | 61.3         |

Models trained on various sub-datasets within the standard dataset are evaluated on the ID test set. The last row calculates the average AUC of detection methods trained on each sub-dataset.

Fig. 3. Perturbation vs. AUC graphs evaluated on ID test set. Note that points belonging to a specific method are interconnected to enhance the visibility of performance variation trends.

TABLE VIII
COMPARISON RESULTS OF DETECTION PERFORMANCE OF ALGORITHMS ON THE IN-DOMAIN STANDARD TEST DATASET, ID TEST SET, AND IN-DOMAIN DATA WITHIN ID TEST SET

| Train Test | DFDC | DFDC | DFDC | DFDC |
|------------|------|------|------|------|
| FWA-ResNet50 | 93.8  | 61.6  | 90.2  | 61.6  |
| MesoNet-4   | 94.8  | 60.5  | 88.3  | 60.5  |
| MesoInception-4 | 90.1  | 59.8  | 77.4  | 59.8  |
| Patch ResNet Layer1 | 92.7  | 32.8  | 72.8  | 32.8  |
| Xception    | 79.7  | 56.3  | 70.5  | 56.3  |
| Multi-attention | 99.0  | 74.7  | 88.7  | 74.7  |
| Conv LSTM   | 86.7  | 67.9  | 74.8  | 67.9  |
| Average AUC% | 90.9  | 61.9  | 80.3  | 61.9  |

D. Evaluation Results of Practicability

We evaluate the practicability of forgery detection methods by analyzing AUC vs. FLOPs graph and AUC vs. number of parameters graph in this section, as shown in Figs. 4 and 5. From these graphical representations, we can observe that M2TR possesses an optimal balance between FLOPs and AUC. This can be attributed to its method-specific model design, which incorporates the EfficientNet-b4 [61] backbone, the multi-scale transformer module, and the global filter [64] module. These design choices collectively maintain low computation complexity. Besides, the lightweight Patch ResNet Layer1 model possesses acceptable performance while preserving lower FLOPs and few parameters, thanks to the design of utilizing the truncated model to capture high frequency discriminative clues learned in shallow layers. Consequently, this model proves more suitable for practical applications constrained by computational resources.

E. Evaluation Results of Efficiency/Effectiveness Trade-Off

The evaluation results depicting the trade-off between efficiency and effectiveness are shown in Fig. 6. From the results, we can infer that Patch ResNet Layer1 achieves acceptable performance with remarkably short inference time, while M2TR achieves the best AUC score but with a relatively longer inference time. This can be primarily attributed to the differences in their model designs. Moreover, video-level methods using image inputs exhibit longer inference times compared to their image-level counterparts, given that video-level methods require testing...
multiple images during inference. However, this video-level inference time can be significantly reduced by replacing image inputs with low-dimensional features (e.g., LRNet). Overall, as illustrated in Table III, the inference time for the majority of methods exceeds 5 milliseconds, implying that the existing deepfake detection algorithms can be time-consuming and less practical for detecting large-scale forgery data in realistic scenarios.

F. Evaluation Results of Classification Decision

To understand the decision-making mechanisms of the evaluated models, we apply Grad-cam [63] to visualize their saliency maps. As shown in Fig. 7, we conduct in-domain visualizations on real and fake samples from FF++ and Celeb-DF-v2 datasets. Moreover, cross-domain visualizations of models trained by FF++/DF and FF++/FS are performed on samples from FF++/FS and FF++/DF datasets. We also provide the manipulated masks to indicate the ground truth manipulation regions. These masks are generated by computing the absolute pixel-wise difference between the samples and their paired source images, followed by Gaussian blur with a kernel size of $7 \times 7$. As the in-domain results show, models trained with weak supervision of image labels, such as Xception and M2TR in our experiments, mainly focus on local regions to make decisions, lacking interpretability. However, models trained with additional supervision or attention
maps demonstrate the ability to concentrate on interpretable forgery clues related to different manipulations. This is because methods merely supervised by image-level labels usually learn discriminative features that can accelerate the decrease of cross-entropy loss and therefore are prone to learn specific local forgeries [65]. In contrast, methods supervised by additional predefined supervisions are capable of learning corresponding forgery clues, such as blending boundary [9] and regional independent discrepancies [5]. Therefore, the highlighted activation regions might not strictly match the manipulated regions but indicate meaningful detection clues. Regarding cross-domain evaluation, the activation regions of most models deviate from the regions associated with designed forgery clues or even concentrate on the forgery-irrelevant background regions, making models challenging to generalize effectively.

G. Analysis of Evaluation Results

Based on the above results and observations, we highlight some key findings.

First, most deep learning methods can achieve promising performance in intra-domain scenarios but suffer from significant performance degradation when confronting cross-domain and manipulation-imperceptible forgery data. This indicates that the detection performance of existing methods is far from satisfactory in practical real-world scenarios.

Second, in-domain evaluation results reveal that the performance disparities between methods may have been excessively emphasized in previous works. This is because the evaluation configurations are inconsistent between compared models, including the inconsistency of training data or test data.

Third, the detection performance evaluation results of our benchmarked methods suggest that adopting state-of-the-art backbones to extract features, along with exploring multi-modal and multi-scale features, contributes to capturing comprehensive discriminative representations. This is beneficial for both in-domain and cross-domain detection. Moreover, enhancing decisive features with low-level features and integrating diverse manipulation data into the training set may improve cross-domain detection performance.

Fourth, concerning the evaluation of detection robustness for our benchmarked methods, replacing image inputs with low-dimensional geometric features proves effective in handling perturbations.

Fifth, considering AUC vs. FLOPs, AUC vs. number of parameters, and AUC vs. inference time, the primary differentiation among methods is derived from their model design. Therefore, equipping the model with efficient backbones or specific modules tailored to the characteristics of deepfake would improve the method’s practicability. For instance, truncating networks to explore low-level, high frequency features could serve as a promising strategy.

Sixth, considering detection ability, generalization, robustness, and practicability simultaneously, no single method demonstrates overall superiority. While considering different perspectives, M2TR achieves the best AUC in terms of detection and generalization performance while maintaining a high time complexity. On the other hand, Patch-Xception-Block2 and Patch-Resnet-Layer1 exhibit comparatively superior detection abilities, characterized by minimal inference time and memory consumption.

V. DISCUSSION

Although we have established a comprehensive and consistent benchmark and attempted to investigate the pros and cons of
existing popular and top-performing deepfake detection algorithms through large-scale experiments, several algorithms in recent studies have not been evaluated [44, 66, 67, 68, 69]. This is due to the unavailability of open-sourced codes of most detection methods and it is difficult to evaluate all up-to-date state-of-the-art detection algorithms in one work. Therefore, we are developing an online deepfake detection platform [70] with our strict benchmark suit. We hope to make our benchmark serve the deepfake detection community as a standardized benchmark and encourage the researchers to incorporate and evaluate their approaches through our platform. This platform is an ongoing work for us to include more challenging manipulation data, state-of-the-art detection methods, and evaluation metrics. It is also a continuous work for facilitating deepfake detection development.

VI. CONCLUSION
This paper has established a comprehensive and consistent benchmark for a holistic and fair evaluation of existing deepfake detection approaches. By performing large-scale experiments with several fair-minded and practical evaluation metrics, we have concluded that dataset inconsistencies can lead to unfair comparisons among popular approaches. A challenging ID test set, including manipulated samples indistinguishable to both humans and detection algorithms, is constructed to better evaluate and understand state-of-the-art deepfake detection methods. The evaluation results reveal that the existing popular deepfake detection algorithms remain far from the expectations for real-world deployment. The evaluation from multiple perspectives indicates that different algorithms have their advantages, and no one method shows superiority comprehensively over others.

To continuously benefit the deepfake detection community, we are developing an online deepfake detection platform with our strict benchmark suit, allowing the researchers to fairly and comprehensively evaluate their approaches through our platform.

REFERENCES
[1] B. Chesney and D. Citron, “Deep fakes: A looming challenge for privacy, democracy, and national security,” California Law Rev., vol. 107, 2019, Art. no. 1753.
[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DFF: A large-scale challenging dataset for deepfake forensics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3207–3216.
[3] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2886–2895.
[4] B. Dollhansky et al., “The deepfake detection challenge dataset,” 2020, arXiv: 2006.07397.
[5] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2185–2194.
[6] H. Dang, F. Liu, J. Stiehourney, X. Liu, and A. K. Jain, “On the detection of digital face manipulation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5781–5790.
[7] J. Baek et al., “What is wrong with scene text recognition model comparisons? Dataset and model analysis,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 4715–4723.
[8] L. Brigata, B. Barz, L. Iocchi, and J. Denzler, “Tune it or don’t use it: Benchmarking data-efficient image classification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 1071–1080.
[9] L. Li et al., “Face X-ray for more general face forgery detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5001–5010.
[10] L. Chai, D. Bau, S.-N. Lim, and P. Isola, “What makes fake images detectable? Understanding properties that generalize,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 103–120.
[11] Y. Li, M.-C. Chang, and S. Lyu. “In iuct oculi: Exposing AI generated fake face videos by detecting eye blinking,” 2018, arXiv: 1806.02877.
[12] P. Korshunov and S. Marcel, “DeepFakes: A new threat to face recognition? Assessment and detection,” 2018, arXiv: 1812.08685.
[13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Forensics: Learning to detect manipulated facial images,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 1–11.
[14] N. Dufour et al., “Deepfakes detection dataset by Google & Jigsaw,” Sep. 2019. [Online]. Available: https://blog.google/research/2019/09/contributing-data-to-deepfake-detection.html
[15] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The deepfake detection challenge (DFDC) preview dataset,” 2019, arXiv: 1910.08554.
[16] Y. He et al., “ForgeryNet: A versatile benchmark for comprehensive forgery analysis,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 4360–4369.
[17] “FakeApp,” 2018. [Online]. Available: https://www.malavida.com/en/face/fakeapp
[18] “Faceswap,” 2017. [Online]. Available: https://github.com/deepfakes/faceswap
[19] “Faceswap-GAN,” 2018. [Online]. Available: https://github.com/shoaibul/faceswap-GAN
[20] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, “FaceShifter: Towards high fidelity and occlusion aware face swapping,” 2019, arXiv: 1912.13457.
[21] I. Perov et al., “DeepFaceLab: A simple, flexible and extensible face swapping framework,” 2020, arXiv: 2005.05535.
[22] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject agnostic face swapping and reenactment,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 7184–7193.
[23] D. Huang and F. De La Torre, “Facial action transfer with personalized bilinear regression,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 144–158.
[24] “3D-FaceSwap,” 2017. [Online]. Available: https://github.com/MarekKowalski/FaceSwap/
[25] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2019, pp. 8261–8265.
[26] S. Das et al., “Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3776–3785.
[27] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 86–103.
[28] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” 2018, arXiv: 1811.00856.
[29] H. Cheng, Y. Guo, T. Wang, Q. Li, T. Ye, and L. Nie, “Voice-face homogeneity tells deepfake,” 2022, arXiv:2203.02195.
[30] Y. Zhou and S.-N. Lim, “Joint audio-visual deepfake detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 14800–14809.
[31] A. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur., 2018, pp. 1–7.
[32] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, “Forensics: Weakly-supervised domain adaptation for forgery detection,” 2018, arXiv: 1812.02510.
[33] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, “Video face manipulation detection through ensemble of CNNs,” in Proc. 25th Int. Conf. Comput. Vis., 2021, pp. 5012–5019.
[34] J. Wang et al., “M2TR: Multi-modal multi-scale transformers for deepfake detection,” in Proc. Int. Conf. Multimedia Retrieval, 2022, pp. 615–623.
[35] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Two-stream neural networks for tampered face detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2017, pp. 1831–1839.
[36] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, “Attention-based two-stream convolutional networks for face spoofing detection,” IEEE Trans. Inf. Forensics Security, vol. 15, pp. 578–593, 2019.
[37] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in Proc. IEEE 10th Int. Conf. Biometrics Theory Appl. Syst., 2019, pp. 1–8.
Qian Wang (Fellow, IEEE) received PhD degree in electrical and computer engineering from the Illinois Institute of Technology. He is currently a professor with the School of Cyber Science and Engineering, and an adjunct professor with the School of Computer Science, Wuhan University. His research interests include AI security, data storage, search and computation outsourcing security and privacy, wireless systems security, Big Data security and privacy, and applied cryptography etc.

Qi Li (Senior Member, IEEE) received the PhD degree from Tsinghua University, Beijing, China. He is currently an associate professor with the Institute for Network Sciences and Cyberspace, Tsinghua University. His research interests include network and system security, particularly Internet security, mobile security, and Big Data security. He is currently an editorial board member of *IEEE Transactions on Dependable and Secure Computing* and ACM Digital Threats: Research and Practice, and has served on the organization or program committees of various premier conferences.

Qiming Li received the master’s degree in engineering from the Harbin Institute of Technology, Harbin, China. He is currently a senior engineer with the Guangdong OPPO Mobile Telecommunications Corporation, Ltd. His current research interests include adversarial attack and defense, object detection, Markov jump system and 3D reconstruction.