International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Emerging Trends and Innovations in Web-Based Applications and Technologies
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Enhancing Deep Fake Detection Generalization through One-Shot Test-Time Adaptation

Harsh. B. Shrivas[1], Ishika. S. Mendhekar[2], Prof. Anupam Chaube[3], Prof. Usha Kosarkar[4]

[1,2,3,4]Department of Science and Technology,
[1,2]G H Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra, India
[3,4]G H Raisoni College of Engineering and Management, Nagpur, Maharashtra, India

ABSTRACT

State-of-the-art deep fake locators perform well in recognizing imitations when they are assessed on a test set comparative to the preparing set, but battle to preserve great execution when the test frauds display distinctive characteristics from the preparing pictures, e.g., frauds are made by concealed deep fake strategies. Such a frail generalization capability prevents the pertinence of current deep fake finders. In this paper, we present a modern learning worldview uncommonly planned for the generalizable deep fake discovery assignment. Our key thought is to develop a test sample-specific assistant errand to overhaul the demonstrate some time recently applying it to the test. Particularly, we synthesize pseudo-training tests from each test picture and make a test-time preparing objective to upgrade the show. In addition, we propose to use meta-learning to guarantee that a quick single-step test-time angle plunge, named one-shot test-time preparing (OST), can be adequate for great deep fake location execution. Broad comes about over a few benchmark datasets illustrate that our approach performs favorably against existing expressions in terms of generalization to concealed information and strength to diverse post-processing steps.

1. INTRODUCTION

Profound neural organize models have brought surprising progresses to picture altering and era procedures. Whereas such advance is revolutionizing the mixed media generation industry, it too makes negative social impacts since it has never been simpler to form profoundly deceivable fraud pictures. Among those unused technologies, deepfake, which employments profound learning models to substitute the character of one individual with another or modify the facial highlights in a representation, is especially hurtful since it can lead to extreme computerized wrongdoing and weaken the social believe framework. To neutralize such a negative effect, deep fake discovery procedure is created to naturally recognize perfect or fraud and is accepting expanding consideration within the investigate community. So far away, existing deep fake location strategies accomplish favorable exhibitions when preparing and test frauds are from the same dataset and created by the same deep fake strategy. In hone, the test frauds are ordinarily created by obscure strategies or connected with distinctive picture post processing approaches. This disparity will unavoidably make a conveyance float between training and test information. Tragically, existing deep fake locators don't generalize well, and their execution tends to diminish essentially when assessed over datasets. This wonder motivates later thinks about on moving forward show generalizations to recognize confront imitations produced from inconspicuous strategies. For illustration, a two heads arrange36th Conference on Neural Information Processing Systems (NeurIPS 2022).test pseudo pretrained updated sample sample detector one-shot detector online training real / fake

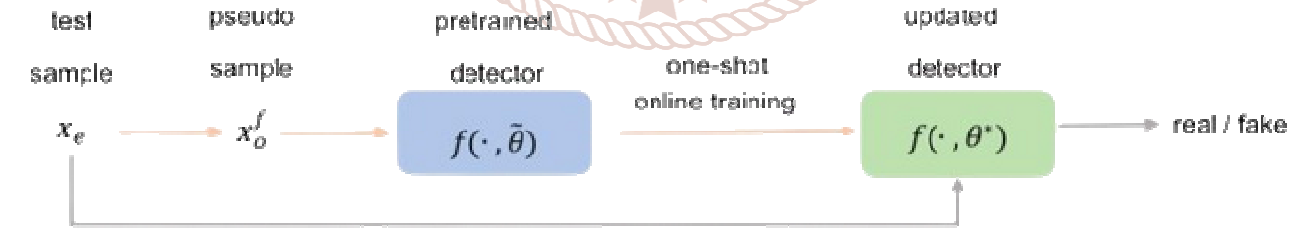


Figure 1: A see of the proposed one-shot test-time preparing system amid online expectations. For every test sample we first synthesis size pseudo training sample based on. Then the pertained locator can be upgraded by means of a directed learning step with of, i.e. of is with a known name as fake. The ultimate result is gotten by applying the upgraded locator to the test in, and facial source highlight irregularities are found in to identify confront frauds, separately. These strategies attempt to investigate common highlights among the preparing frauds for superior classification. But the test information regularly shows distinctive characteristics, and the learned common highlights may not be shared by them. Since the test samples are not seen within the preparing stages, getting great generalization appears inaccessible for current finders. This work presents a unused learning worldview specially-designed for the generalizable deep fake location. Particularly, we permit the detector to "see" the test tests some time recently making the ultimate forecast by conducting an extra "preparing step" at the test time. One challenge of this thought is that the name of the test picture is inaccessible for the preparing objective. We overcome this issue by synthesizing a pseudo-training test based on the test picture and utilizing it to upgrade the deep fake location show online. In a common deep fake location setting, no matter genuine or fraud for the assessed test, we are sure that the synthesized pseudo test could be an imitation. This one of a kind property empowers our locator to prepare on the synthesized test that has comparable substance to the test, hence superior adjusting to the test characteristic. Besides, we propose to utilize as it were one step angle plunge, named one-shot test-time preparing (OST), to overhaul the show online for way better computational

proficiency. A set of OST is appeared in Figure 1. To guarantee that such an OST scheme can continuously lead to a not too bad demonstrate without over fitting the pseudo-training test, we utilize meta-learning to prepare a great beginning demonstrate in a comparable fashion as MAML]. Note that in spite of the fact that test-time preparing (TTT) has been already proposed for the common picture classification assignment, our approach embraces a diverse TTT objective that's more particular for deep fake location. In the mean time, as both recommended by a later work and our exploratory consider, the general-purposed self-supervised TTT calculation in may come up short to bring any enhancements but may indeed fall apart the location exactness. On the opposite, our OST strategy can altogether move forward the generalization execution

2. Related Works

In this section, we conduct a brief survey on the most relevant arts, including existing deepfake detection methods and test-time training (TTT)-based works.

2.1. Deepfake Detection

Since the deep fake frauds have driven to extraordinary dangers to societal security, it is of vital significance to create successful locators against it. By defining the identifying as a vanilla double classification issue (i.e. flawless or imitation), current end-to-end prepared locators with a straightforward Exception pattern can get tall location exactness. Other than, with more capable arrange structures and more enlightening picture highlights implanted within the arrange inputs, existing strategies are able to attain indeed more surprising victory when the preparing and test imitations are synthesized by the same deep fake calculations. Hence, the genuine challenge in this assignment lies in how to generalize a learned finder to frauds made by inconspicuous strategies.

A few works have been given to tending to the generalizing issue as of late. For example, proposes that the mixing operation is omnipresent within the current deep fake synthesizing process. As a result, they propose to identify the mixing boundaries covered up within the frauds and utilize them as classification clues. Besides, appears that the up-sampling step in synthesizing models can bring artifacts to the synthesized frauds, and they utilize the stage spectrums of the imitations to capture these artifacts. As existing fraud synthesizing steps regularly include two pictures from distinctive characters and diverse sources, recommend utilizing high-pass channels from SRM to uncover detail disparities of the frauds. A comparable thought is received in, where they utilize the signal of the source highlight irregularity inside the fashioned pictures for location. In spite of the fact that these strategies are successful in many cases, the low-level artifacts they depend on are touchy to post-processing steps that change in different datasets, hence jeopardizing their generalization. A few other works propose to borrow highlights from other assignments, such as lips perusing, facial picture decay, and point of interest geometric, to suggest the abnormity of frauds. In spite of the fact that these highlights can bring certain changes, there's a incredible chance that future deep fake calculations will be planned based on these finders to synthesize more normal imitations, causing indeed greater threats to societal security.

Compared to existing finders, the preferences of our strategy are as takes after:
(1) we receive a MAML based OST system to empower the quick adjustment of the learned locator to the test information, which moves forward generalization in any case of the changing post-processing steps; (2) OST does not depend on hand-crafted or borrowed highlights, which clears out less follows for the deep fake calculations to assault.

2.2. Test-time Training

The concept of TTT was firstly presented in for generalization to out-of-distribution test information, where a self-supervised turn forecast errand is utilized with the most classification errand amid preparing, and as it were the self-supervised task is embraced to assist progress the visual representation amid deduction, which in a roundabout way progresses semantic classification. This system is hypothetically demonstrated to be compelling and is advance utilized in other related ranges. For example, proposes a recreation errand inside the most posture estimation system, which can be prepared by comparing the reproduced picture with the ground truth borrowed from other outlines. Illustrate that the predictions with lower entropy have lower mistake rates, and they use entropy to supply fine-tuning signals when given a test picture. Rather than as it were minimizing the entropy of the anticipated back, too recommends maximizing the commotion vigor of the include representation amid test. In a few later works, the TTT system has moreover been utilized inside a show freethinker meta-learning (MAML) worldview which permits the prepared demonstrate to be optimized in a way such that it can rapidly adjust to any test pictures. To empower quick adjustment, these works utilize contrastive misfortune or smooth misfortune to fine-tune the models amid the meta-test stage.

Be that as it may, in spite of a few empowering comes about, current TTT strategies point to select observational self supervised errands, which is at tall hazard of falling apart the performance when the assignments are not legitimately chosen. This work presents a modern learning worldview extraordinarily outlined for the deep fake location errand. Review that a fraud can be effortlessly synthesized by mixing two diverse pictures. We hence utilize a recently synthesized forgery. As pseudo-training sample to finetune the pretrained detector during inference. Our method is easy to implement and can avoid the tedious work of selecting an effective self-supervised task.

3. Proposed Method

Our approach includes an offline meta-training phase and an online test-time training phase. During the online test-time training, a pseudo-training sample is generated for each test image, followed by a single gradient descent update to fine-tune the model with sample-specific parameters. The offline meta-training stage is designed to replicate the test-time training process by creating training episodes from available data. Below, we first explain the test-time training procedure, covering both pseudo-training sample generation and one-shot model updates. Subsequently, we provide details on the meta-learning framework.

3.1. OnlineTest-TimeTraining

In the following discussion, we assume that a deepfake detection model θ has already been trained. The OST̃ process updates the model parameters, resulting in θ· which adapts to each specific test image. This adaptation is performed by first generating a pseudo-training sample, which is then used to construct a mini-training set for a one-shot training update to refine the model parameters.

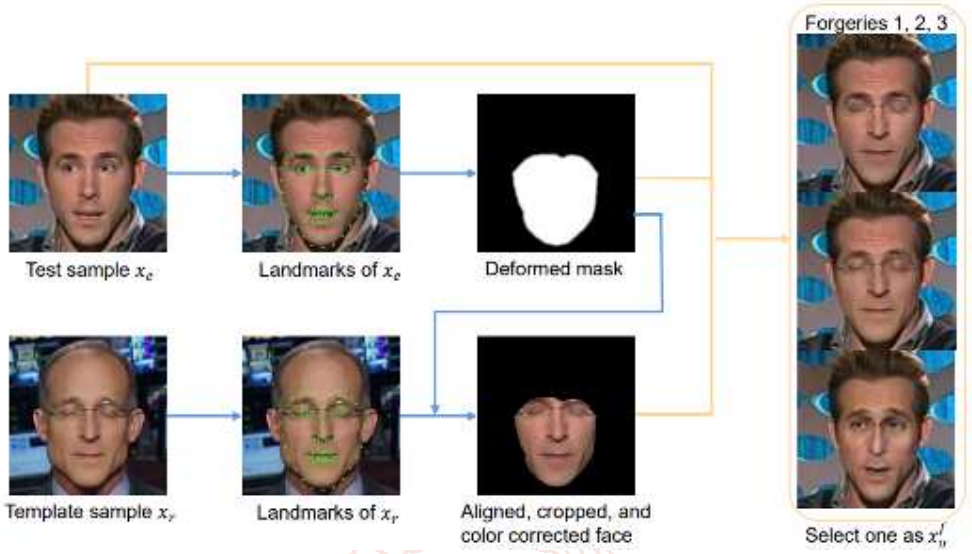


Figure 2: Pipeline for generating pseudo training samples. Forgeries 1, 2, 3

Generatingpseudo-training samples: As showing Figure 2, For every test sample x_e, we first randomly select a template image x_r from the training dataset and align these two images in geometry based on their landmarks.

The proposed test-time training approach can be interpreted as a domain adaptation technique. In this framework, each test image is treated as a unique domain, characterized by its content, which may differ from the training data due to a domain gap. The pseudo-training sample generated through this method is more closely related to the test image than the original training samples, as it is synthesized based on the test image itself. By performing rapid adaptation on this generated sample, the detector can better align with the test image, improving its performance. Additional evidence supporting this analysis is provided.

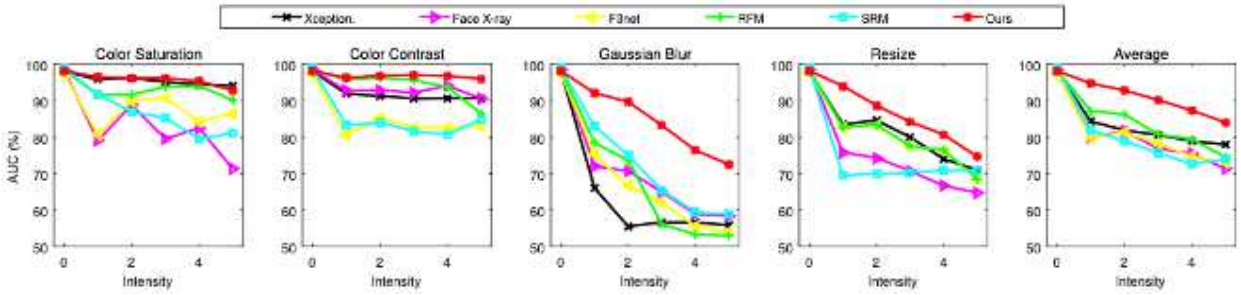
4. Experiments

This section first presents the setups and then shows extensive experimental results to demonstrate the superiority of our approach. Please refer to the supplementary material for more experimental results.

4.1. Settings

Training and test datasets. Following the protocols in existing deepfake detection methods, we use the data in the Faceforencis++ (FF++) dataset for training. This dataset contains

Method	DF			F2F			FS			NT			Avg.
	DFDC	DFD	DF1.0	DFDC	DFD	DF1.0	DFDC	DFD	DF1.0	DFDC	DFD	DF1.0	
Xception [43]	0.654	0.808	0.617	0.708	0.695	0.745	0.708	0.657	0.605	0.646	0.724	0.838	0.700
Face X-ray [28]	0.609	0.811	0.668	0.633	0.679	0.766	0.646	0.625	0.795	0.613	0.645	0.866	0.696
F3Net [42]	0.682	0.812	0.658	0.679	0.729	0.761	0.679	0.641	0.651	0.672	0.826	0.932	0.727
RFM [50]	**0.758**	0.833	0.717	0.736	0.711	0.732	0.714	0.593	0.714	0.726	0.816	0.846	0.741
SRM [35]	0.679	0.855	0.720	0.687	0.801	0.775	0.671	0.705	0.771	0.656	0.791	**0.936**	0.754
Ours	0.757	**0.869**	**0.938**	**0.798**	**0.880**	**0.947**	**0.802**	**0.824**	**0.909**	**0.752**	**0.841**	0.929	**0.854**



The dataset consists of **1,000 videos**, with **720 used for training**, **140 reserved for validation**, and the remaining videos allocated for testing. Each real video undergoes manipulation using four deepfake techniques: **DeepFake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTexture (NT)**, resulting in four corresponding synthetic videos. Additionally, the dataset is available in three different quality levels—**raw, lightly compressed**

(HQ), and heavily compressed (LQ)—with **HQ being the default** unless specified otherwise.

To assess the generalization capability of the proposed method, we evaluate it across four additional benchmark datasets:

- **DeepfakeDetection (DFD)**: Comprising **363 real videos and 3,068 deepfake videos** generated using an enhanced DeepFake approach.
- **Deepfake Detection Challenge (DFDC)**: Includes **over 1,000 real and more than 4,000 fake videos**, manipulated through various deepfake, GAN-based, and traditional non-learned techniques.
- **DeeperForensics-1.0 (DF1.0)**: Consists of **over 11,000 deepfake videos** created using the **DFVAE** method.
- **CelebDF**: Contains **408 real and 795 synthetic videos**, produced with an improved DeepFake method.

Notably, the fake videos in the training and test sets do not share the same content or generation techniques, ensuring a fair evaluation of model performance.

Implementation Details

The **Xception network** is employed for face alignment, with aligned faces resized to **256 × 256** pixels. The model weights are initialized using **pretrained ImageNet parameters**, and **DLIB** is utilized for face extraction.

For training, we optimize the model using the **Adam optimizer** with **β1 = 0.9 and β2 = 0.999**, and a **meta-batch size of 20**. The learning rates for the **inner update (γ)** and **meta update (λ)** are set to **0.0005 and 0.0002**, respectively, for both the offline and online training phases.

Further Analysis

During the forgery generation process, a training sample is randomly chosen and blended with the test sample. To evaluate the effectiveness of this selection method, we conduct ablation studies comparing it against two alternative strategies:

1. **Nearest Neighbor (NN) Sampling** – The training sample with the closest feature distance to the test sample is selected for blending.
2. **Average (Avg) Sampling** – The performance of multiple training samples is averaged for evaluation.

These configurations are tested on four types of data within the **FF++ dataset** and their performance is further validated on the **DFDC, DFD, and DF1.0 datasets** to ensure generalizability.

OST Adaptation

The proposed learning framework is based on **Model-Agnostic Meta-Learning (MAML)**, which facilitates rapid adaptation to new tasks. To assess the role of **One-Shot Test-Time Training (OST)** within this framework, we compare its performance against a conventional training scheme while keeping the dataset and **Xception** backbone unchanged. Results, presented in Table 6, show that OST with standard end-to-end training performs comparably to MAML. However, when the OST process is removed, detection performance declines, indicating that the method enhances conventional training configurations and strengthens widely used network architectures. While both approaches are effective, the MAML framework accelerates adaptation during inference—reducing test sample processing time by a factor of **10×**. In practical applications, we integrate OST within the MAML framework to maximize efficiency.

Impact of Multiple Gradient Descent Steps

To determine whether additional gradient descent steps enhance detection accuracy, we perform ablation studies by varying the number of updates during evaluation. The model, trained on **FF++**, is tested on **DFDC, DFD, and DF1.0** datasets. As shown in Table 7, increasing the number of gradient descent steps does not yield substantial improvements in accuracy, while computational cost scales proportionally. Additionally, using multiple updates significantly increases memory consumption in an MAML-based framework. To balance efficiency and accuracy, we opt for a **single gradient descent step** in our final method.

Conclusion and Discussions

In this research, we present a novel learning paradigm tailored specifically for the generalizable deepfake detection challenge. To summarize, we recommend finetuning the pretrained detector with a pseudo-training sample, which is created by blending the test samples with a randomly picked template picture, prior to the classification phase. We empirically demonstrate that the proposed online training strategy allows the pretrained model to adjust to sample-specific statistics, hence improving generalizability. We implement our method in a MAML-based framework to allow for rapid adaption to varied test samples, and it outperforms state-of-the-art methods for generalization to previously encountered forgeries and other postprocessing processes.

Limitations and future work. Because the pseudo-training examples are synthesized using existing deepfake pipelines, our method cannot be applied to scenarios where the fake images are made using different protocols, such as when fake images are completely synthesized using GAN-based methods. Our next work will focus on developing approaches for deepfakes as well as GAN-synthesized fake images. Meanwhile, DLIB is employed in our forgery synthesis workflow to identify and extract facial landmarks. Given that there are instances in which DLIB may fail simultaneously with OST. Thus, a more effective facial detection system promotes OST. Ethical statement. This initiative aims to assist people in combating the exploitation of deepfake technology. It does not.involve any human or animal subjects, and there is no infringement of personal privacy during the experiment. We do not expect any possible negative implications for our efforts. We believe that our research and the release of our code will increase scientific and society awareness of the subject of generalizable deepfake detection. Acknowledgements. Liang Chen is financed by the China Scholarship Council (CSC Student ID: 202008440331).

References

[1] Isao Echizen, Junichi Yamagishi, Vincent Nozick, and Darius Afchar. Mesonet: a small network for detecting face video forgeries. In WIFS, 2018.

[2] Koki Nagano, Yuming Gu, Mingming He, Hany Farid, Hao Li, and Shruti Agarwal. defending global leaders from deepfakes. 2019's CVPR Workshops.

[3] Alberto Del Bimbo, Leonardo Galteri, Roberto Caldelli, and Irene Amerini. Deepfake video detection using a CNN based on optical flow. In Workshops on ICCV, 2019.

[4] Matthias Nießner and Shivangi Aneja. Few-shot and generalized zero-shot transfer for detecting facial forgeries.2020; arXiv preprint arXiv:2006.11863.

[5] Mario Döbler, Bin Yang, Felix Wiewel, Andre Bühler, and Alexander Bartler. Mt3: Self-supervised test-time adaptation using meta test-time training. Preprint arXiv:2103.16201, 2021 and arXiv.

[6] Ngoc-Trung Tran, Ngai-Man Cheung, and Keshigeyan Chandrasegaran. A more thorough examination of fourier spectrum inconsistencies for the detection of CNN-generated images.

[7] Xiaoguang Han, Xiaoqing Liu, Jiongcheng Li, Yizhou Yu, and Chaoqi Chen. compound domain generalization by the encoding of meta-knowledge. CVPR, 2022.

[8] Yizhou Yu, Yue Huang, Gangming Zhao, Feng Liu, Luyao Tang, and Chaoqi Chen. Mix and reason: Using data mixing to provide domain generalization while reasoning over semantic topology. 2022, in NeurIPS. Jibing Song, Liang Chen, Yong Zhang, Lingqiao Liu, and Jue Wang [3]. Self-supervised learning of an adversarial example: Moving toward sound generalizations for the detection of deepfakes. CVPR, 2022.

[9] Bingbing Ni, Yanhao Ge, Xuanhong Chen, and Renwang Chen. Simswap: An effective framework for face switching in high fidelity. In 2020, ACM MM. Chollet, François [5]. Deep learning using depthwise separable convolutions is called Xception. 2017's CVPR. Deepfaked detection. https://ai.googleblog.com/2019/09/contributing-data-to-deepfaked etection.html

[10] www.github.com/deepfakes/faceswap DeepFakesAccessed April 24, 2021.

[11] Li-Jia Li, Kai Li, Li Fei-Fei, Richard Socher, Wei Dong, and Jia Deng. An extensive hierarchical image database is called Imagenet. 2009; in CVPR.

[12] The issue of deepfake detection. This link will take you to the deepfake-detection-challenge on Kaggle. Accessed April 24, 2021.

[13] Baining Guo, Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, DongChen, and FangWen [5]. utilizing an identity consistency transformer to safeguard celebrities. 2022, in CVPR.

[14] Freddie Witherden, Karan Shah, and Tarik Dzanic. Fourier spectrum differences in images produced by deep networks. 2020's NeurIPS.

[15] FaceSwap. Accessed April 24, 2021. www.github.com/MarekKowalski/FaceSwap.

[16] Sergey Levine, Pieter Abbeel, and Chelsea Finn. Model-independent meta-learning for rapid deep network adaptation. ICML, 2017.

[17] Jan Kodovsky and Jessica Fridrich. Rich models for digital image steganalysis. IEEE TIFS, 7(3), 2012, 868–882.

[18] Praveer Singh, Nikos Komodakis, and Pyros Gidaris. Predicting picture rotations allows for unsupervised representation learning. In 2018, the arXiv preprint arXiv:1803.07728.