International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02208-2

Test-time Forgery Detection with Spatial-Frequency Prompt Learning

Junxian Duan[1,2,3] · Yuang Ai[1,2,3] · Jipeng Liu[4] · Shenyuan Huang[1] · Huaibo Huang[1,2,3] · Jie Cao[1,2] · Ran He[1,2,3]

Received: 15 September 2023 / Accepted: 27 July 2024

Abstract
The significance of face forgery detection has grown substantially due to the emergence of facial manipulation technologies. Recent methods have turned to face detection forgery in the spatial-frequency domain, resulting in improved overall performance. Nonetheless, these methods are still not guaranteed to cover various forgery technologies, and the networks trained on public datasets struggle to accurately quantify their uncertainty levels. In this work, we design a Dynamic Dual-spectrum Interaction Network that allows test-time training with uncertainty guidance and spatial-frequency prompt learning. RGB and frequency features are first interacted in multi-level by using a Frequency-guided Attention Module. Then these multi-modal features are merged with a Dynamic Fusion Module. As a bias in the fusion weight of uncertain data during dynamic fusion, we further exploit uncertain perturbation as guidance during the test-time training phase. Furthermore, we propose a spatial-frequency prompt learning method to effectively enhance the generalization of the forgery detection model. Finally, we curate a novel, extensive dataset containing images synthesized by various diffusion and non-diffusion methods. Comprehensive evaluations of experiments show that our method achieves more appealing results for face forgery detection than recent state-of-the-art methods.

Keywords Face forgery detection · Spatial-frequency prompt learning · Test-time training · Generalization · Diffusion model

Communicated by Segio Escalera.

✉ Ran He
rhe@nlpr.ia.ac.cn

Junxian Duan
junxian.duan@ia.ac.cn

Yuang Ai
aiyuang2023@ia.ac.cn

Jipeng Liu
liujipeng@iie.ac.cn

Shenyuan Huang
huaibo.huang@ia.ac.cn

Huaibo Huang
hsywatchingu@foxmail.com

Jie Cao
jie.cao@cripac.ia.ac.cn

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA, Beijing, China

[3] University of Chinese Academy of Sciences, Beijing, China

[4] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

1 Introduction

In recent years, the research in face forgery has experienced remarkable advancements. However, deep learning technology has notably elevated the quality of digital forgery (Sushko et al., 2022). Forgery images may deceive the facial recognition systems or give a chance to malicious exploitation, giving rise to profound trust-related and security apprehensions within the society (Liu et al., 2023). Consequently, there is a compelling need to explore more efficient and robust methods for face forgery detection.

The majority of existing approaches concentrate on within-database detection, as seen in literature like Guo et al. (2023) and Li et al. (2023), where both training and testing datasets comprise manipulated images from the same forgery technique. This is highlighted in Fig. 1, where significant stylistic distinctions exist among the synthesized images from different forgery techniques. Consequently, an ongoing issue in face forgery detection is how to detect forgeries that are unseen in training data (Guo et al., 2022). As shown in Fig. 2a, notable frequency distribution variations between real and fake images occur in certain datasets, posing difficulties in differentiation within the RGB domain. Recent

Published online: 13 August 2024

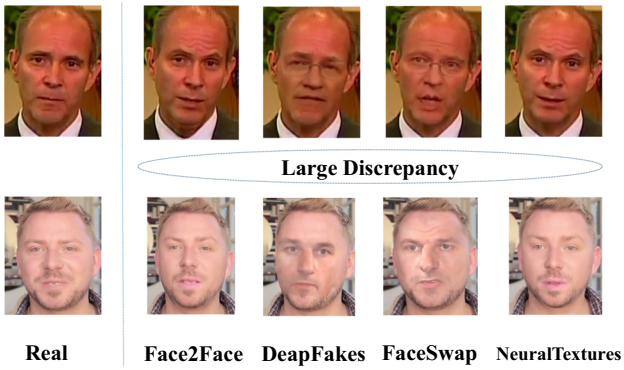

Fig. 1 Real images alongside their fake images through various forgery techniques (The images are collected from the FaceForensics++ dataset)

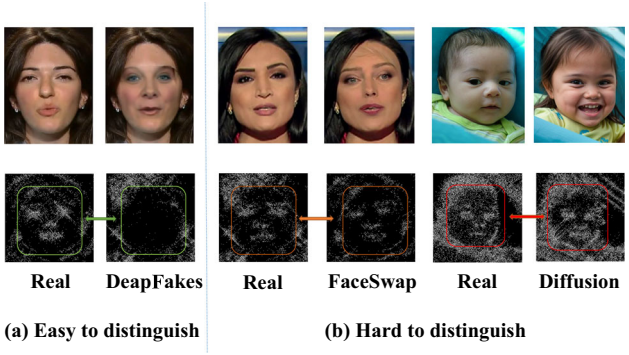

Fig. 2 Quality differences between the RGB and frequency domains. Notably, the difference in (**a**) highlighted with green boxes is more evident, while the red box in (**b**) makes it challenging to distinguish forgery traces. (The DeapFakes and FaceSwap images are collected from the FaceForensics++ dataset.)

advancements have introduced a face forgery frequency network to uncover forgery patterns within the frequency domain. Chen et al. (2021) propose a similarity model utilizing frequency features to enhance performance in previously unexplored domains. Similarly, Luo et al. (2021) assume that the presence of high-frequency noise in images could lead to the removal of color texture, thereby unveiling traces of forgery. Nonetheless, as forgery trace is also embedded within RGB features, the efficacy of the frequency domain is not consistently optimal, as shown in Fig. 2b. Additionally, diffusion-based methods (Zhang & Chen, 2023) can generate high-quality images, which present new challenges for forgery detection. We find it hard to distinguish the difference between the real and fake images in Fig. 2b using only the frequency information, especially in the case of diffusion-based techniques. Therefore, in Sect. 4.2, we will introduce our innovative dataset, which is based on newly proposed diffusion models, thereby enhancing the diversity of the forgery detection dataset.

Hence, there are still challenging issues: a) the reliability of frequency-based approaches can not be consistently assured across various forgery techniques, and b) networks trained on public datasets struggle to accurately quantify their levels of uncertainty when meeting the unseen data.

Addressing the discrepancies in feature quality and model uncertainty, we introduce a novel Dynamic Dual-spectrum Interaction Network (DDIN) to assess quality differences between RGB and frequency domains. During test-time training, we use uncertain perturbations to enhance forgery detection on unseen data. Additionally, inspired by previous work in visual prompt learning (Sanh et al., 2022; Wang et al., 2022b), we employ a spatial-frequency prompt learning method to boost the generalization of the pre-trained DDIN model.

First, we employ Discrete Cosine Transform (DCT) to convert an RGB image into its frequency domain representation. These images are then fed into a Transformer-based network. During the multi-level interaction phase, the Frequency-guided Attention Module (FAM) improves forgery detection by considering frequency-related aspects. The Cross-modal Attention Module (CAM) combines features from the dual-stream network, providing enriched information about the forged areas. The Dynamic Fusion Module (DFM) model dynamically enhances multi-modal information for improved generalization. Moreover, we introduce test-time training with uncertainty and incorporate spatial-frequency prompt learning. This approach fine-tunes the dynamic fusion module using uncertainty estimation from unseen test data. Uncertainty-guided perturbations encourage the network to probabilistically predict quality weights. We leverage the spatial-frequency prompt learning method during test-time training to efficiently fine-tune the parameters. This fine-tuning process helps the model adapt to distribution biases caused by uncertainty, narrowing the prediction distributions for forgery features in both training and testing datasets.

In brief, the main contributions are summarized as follows:

- We propose a dynamic dual-spectrum interaction network, which incorporates the frequency-guided attention module and dynamic fusion module. This framework enables the dynamic fusion of features by considering quality differences, effectively mitigating their impact.
- We propose a test-time training scheme with uncertainty guidance. Through applying uncertain perturbations, we improve the generalization of forgery detection.
- We introduce spatial-frequency prompt learning in test-time training, which contributes to the efficiency of parameter fine-tuning and improves the detection accuracy.
- We construct a new large-scale face forgery detection dataset for diffusion-generated images, which contains

more than 100k real images and 140k synthetic images under the newly proposed techniques based on diffusion models and a non-diffusion model.
– Extensive experiments demonstrate that our proposed method outperforms existing methods in terms of both effectiveness and generalization.

This paper serves as an expansion of our prior conference version Huang et al. (2023c), incorporating three noteworthy enhancements over the initial iteration. **First,** to augment the accuracy and generalization capabilities of the training model, we introduced the uncertain-guided test-time training method in the preliminary version. In the current version, we introduce spatial-frequency prompt learning, which contributes to the efficiency of parameter fine-tuning. Notably, the model integrated with spatial-frequency prompt learning demonstrates robust performance across multiple databases, resulting in heightened accuracy for forgery detection. **Second,** diffusion-based methods bring new challenges for forgery detection but are less explored in the existing literature. There are high-resolution images with fine details and the diffusion models update rapidly. Thus, we have introduced a new expansive dataset dedicated to face forgery detection with the newly proposed techniques based on diffusion models. To highlight the significance of detecting diffusion forgeries, the dataset also includes non-diffusion images. This dataset comprises 100,000 authentic images and 140,000 synthetic images, substantially enhancing the diversity of face forgery datasets available. **Finally,** we have conducted additional experiments and visualization to verify the generalization of our approach and the competing methods. Then we also delved into more insightful analyses of detection accuracy. Significantly, the present version outperforms the preliminary iteration across all datasets including the new dataset we proposed, reflecting substantial improvements in the performance of generalization and effectiveness.

The rest of the paper is organized as follows. We review related work and give a brief introduction of the preliminary background in Sect. 2. In Sect. 3, we describe the proposed method combined with the spectrum transformation, multi-level interaction, multi-modal fusion, and uncertainty-guided test-time training with spatial-frequency prompt learning. Then, we introduce the new large-scale dataset and evaluate the performance of the proposed method in Sect. 4. Finally, we conclude this paper in Sect. 5.

2 Related Work

In this section, we review related studies and summarize the main attributes involved in the forgery detection scenarios. Their differences to our method are also discussed.

2.1 Face Forgery Detection

The field of face forgery detection has witnessed remarkable progress (Huang et al., 2023a; Sun et al., 2022), marked by the successive development of forgery detection models tailored to practical applications. In the initial stages, approaches (Das et al., 2021; Sun et al., 2021; Yang et al., 2021) placed substantial emphasis on identifying semantic visual anomalies through intricate model architectures. Dang et al. (2020) introduce a segmentation task, optimizing it alongside a classification backbone to concurrently predict regions afflicted by forgery. Another strategy, Liu et al. (2021a) highlight how up-sampling in synthetic models can introduce anomalies in fabricated forgeries. Then they utilize the phase spectrums of forgeries to capture anomalies. In a similar vein, Zhu et al. (2023) consider a face image as the production of the interaction between 3D geometry and lighting environment. Somepalli et al. (2023) study image retrieval frameworks to compare training samples with generated images by diffusion models. Guillaro et al. (2023) combine the RGB image and a learned noise-sensitive fingerprint to extract both high-level and low-level traces. While these methods yield satisfactory outcomes within their respective datasets, their performance during cross-dataset evaluations exposes shortcomings in terms of generalizability.

Recently, numerous research efforts have been dedicated to tackling the issue of generalization. For example, both Li et al. (2021a) and Chen et al. (2021) have recognized content-independent, low-level attributes capable of uniquely identifying the sources of manipulation. To prompt the generalization of the detection model in cross-database evaluation, Miao et al. (2022) designed an HFI-Net Network to analyze a wide range of frequency-related forgery cues for the detection of face manipulation. Furthermore, in addressing the fine-grained forgery detection challenge, Miao et al. (2023) introduce the High-Frequency Fine-Grained Transformer (F2Trans) network, leveraging spatial and frequency domain information. While Li et al. (2021a) suggest identifying these subtle features along facial boundaries, Chen et al. (2021) try to uncover spatial-local inconsistencies. Wang et al. (2023) propose to capture both spatial and temporal artifacts in one model for face forgery detection. Wang et al. (2022a) introduce LiSiam, a Siamese Network, to enhance localization invariance for improved deepfake detection across various image degradations. Bai et al. (2023) propose the Action-Units Relation Learning framework to improve the generality of forgery detection. In a different approach, Gu et al. (2022); Cao et al. (2022); Kwon et al. (2022) combine low-level frequency pattern learning with CNN architectures to enhance generalizability. Nonetheless, these techniques often rely on low-level artifacts that are sensitive to post-processing techniques, which can reduce their effectiveness in different scenarios as the dataset changes.

Dong et al. (2023b) propose a deep neural network to learn features that are applicable to manipulations in unseen data. Despite the incremental progress these methods may offer, it is conceivable that future iterations of deepfake algorithms will generate even more realistic. In contrast to existing approaches, our innovative proposition introduces a dynamic network with dual-spectrum interaction that permits test-time training with guidance from uncertainties.

2.2 Test-Time Training Strategy

The concept of test-time training was originally introduced in Sun et al. (2020) as a strategy to improve generalization to unseen test data. This approach combines the primary classification task with a self-supervised rotation prediction task during the training phase. However, during inference, only the self-supervised task is utilized to enhance visual representation, consequently indirectly boosting semantic classification accuracy. The effectiveness of this framework is theoretically substantiated and has found application in other pertinent domains, as exemplified by Bartler et al. (2022); Zhang et al. (2021); Liu et al. (2022a). For instance, Li et al. (2021b) propose incorporating a reconstruction task within the core pose estimation framework. This task is trained by contrasting reconstructed images with ground truth extracted from alternate frames. Wang et al. (2021) unveil that predictions with lower entropy correspond to decreased error rates. They leverage entropy as a mechanism to provide fine-tuning cues when processing a test image. Furthermore, in the context of dehazing, Liu et al. (2022a) introduce a meta-learning approach to address the challenge of supervision signals from helper networks not consistently contributing to enhanced performance for the dehazing network.

To facilitate swift adaptation, these studies utilize contrastive loss (Bartler et al., 2022) or smooth loss (Zhang et al., 2021) for refining models during the meta-test phase. In a specialized application, Chen et al. (2022b) introduce a one-shot test-time training approach tailored specifically for the task of generalizable forgery detection. However, despite their favorable outcomes, current Test-Time Training (TTT) methods often select empirical self-supervised tasks. This selection process introduces a notable risk of performance degradation when the tasks are not suitably selected, as highlighted in Liu et al. (2021b). In contrast, our scheme avoids the need to select an effective self-supervised task, which can significantly elevate the generalization performance of deepfake detectors. Moreover, our method outperforms existing solutions in a variety of benchmark datasets.

2.3 Visual Prompt Learning

Rather than relying on hand-engineered prompts, recent works learn prompts using the training data from downstream tasks, like advanced visual prompting (VP) for vision tasks. VP can reprogram a fixed, pre-trained source model to optimize downstream tasks by incorporating universal prompts into downstream data (Shu et al., 2022; Chen et al., 2022a).

The diversity of data in image datasets presents challenges for visual prompt learning. Based on the efficiently tuning large language models, Jia et al. (2022) propose an alternative method based on the visual prompt for updating all the backbone parameters in pre-trained large-scale Transformer models. To handle the transfer of complex distributions to the original pre-trained data distribution, Huang et al. (2023b) propose a dataset diversity-aware prompting strategy by meta-prompt initialization. Dong et al. (2023a) leverage several trainable prompts into a frozen pre-trained model to adapt it to long-tailed data. They put shared prompts to learn general features and to adapt a pre-trained model, and use other prompts to gather group-specific features for similar features. Wang et al. (2022b) deal with catastrophic forgetting under continual learning and the prompts are defined as small learnable parameters to instruct the model prediction. Chen et al. (2022a) introduce a visual prompting framework that automatically re-maps the source labels to the target labels.

Inspired by recent advances in prompt learning research in natural language processing (NLP), Zhang et al. (2022) leverage prompt learning to enhance CLIP's (Contrastive Vision-Language Pre-training) adaption capability in fine-tuning additional learnable modules. Zhou et al. (2022b) propose a CLIP-like vision-language model for downstream image recognition. To address the overfits base classes observed during training, Zhou et al. (2022a) also propose dynamic prompts adapt to each instance and are less sensitive to class shift than static prompts. Yao et al. (2021) present cross-modal prompts for pre-trained vision-language models, they redefined the visual grounding as the fill-in-the-blank problem of common reference markers in images and text. Shu et al. (2022) propose a test-time prompt to learn adaptive prompts for image classification, and optimize the prompt to improve the zero-shot top-1 accuracy of CLIP. In our work, we also take advantage of test time and prompt learning for face forgery detection tasks.

3 Proposed Method

In this section, we introduce our scheme for face forgery detection. We leverage two-stage training including early training and test-time training. In the first stage, as illustrated in Fig. 3, there are three phases in the Dynamic

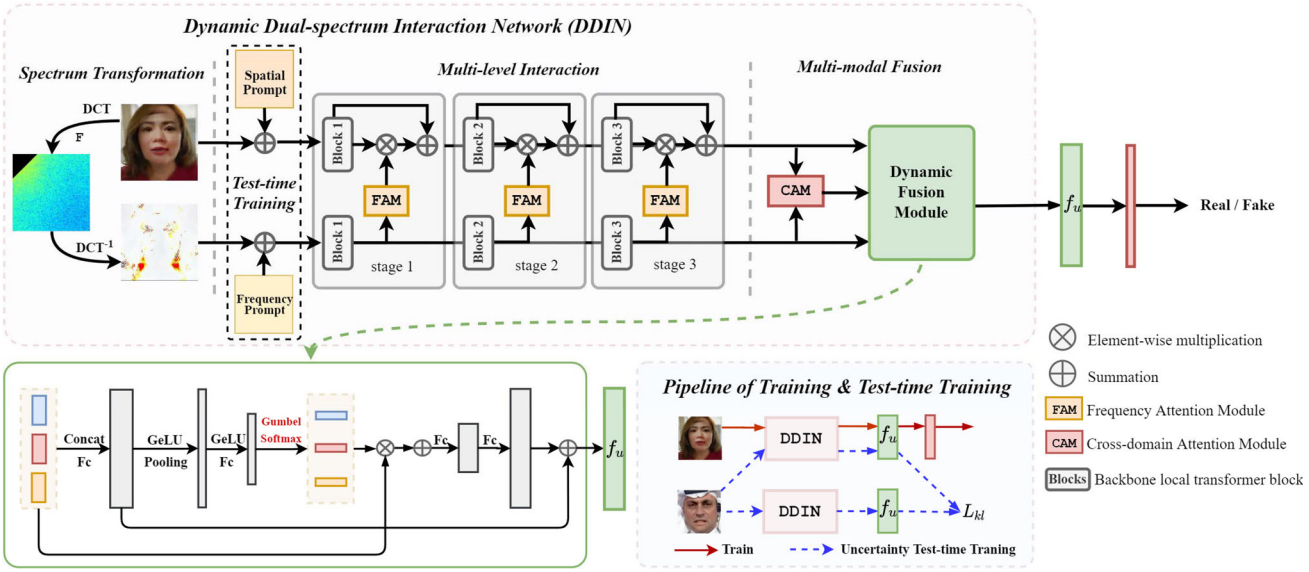

Fig. 3 Framework of our proposed DDIN and the pipeline of test-time training

Dual-spectrum Interaction Network, which will be introduced in Sect. 3.1. Followed by a test-time training of the fine-tuned network using unlabeled test data, as shown in the pipeline of our training scheme, we will demonstrate the uncertainty-guided test-time training method with spatial-frequency prompt learning in Sect. 3.2.

3.1 Dynamic Dual-spectrum Interaction Network

There can be divided into three stages in the DDIN. Firstly in Sect. 3.1.1, the input RGB image is decomposed into frequency components by spectral transformation, which can take to the spatial and frequency prompt in the test-time training phase. Secondly, feature extraction is realized based on three stages in MOA-Transformer, and a frequency-guided attention model is introduced in Sect. 3.1.2. Finally in Sect. 3.1.3, after combining the resulting RGB features with frequency features by a cross-domain attention module, we leverage a dynamic fusion module in the multi-modal fusion phase.

3.1.1 Spectrum Transformation

In the first stage of our DDIN network, we apply spectrum transformation that decomposes an input RGB image into frequency components, assisting the network in mining the distinction between real and forged regions.

Without loss of generality, let $X_{rgb} \in \mathbb{R}^{H\times W\times 3}$ represent the RGB input, with H and W denoting the height and width, respectively. Initially, we employ the Discrete Cosine Transform (DCT) to convert X_{rgb} from RGB to the frequency domain. This transformation places low-frequency responses in the top-left corner and high-frequency responses in the bottom-right corner. According to Qian et al. (2020), the low-frequency band encompasses the initial 1/16 of the spectrum, the middle-frequency band spans from 1/16 to 1/8 of the spectrum, and the high-frequency band encompasses the remaining 7/8 of the spectrum. To accentuate subtle artifacts at high frequencies, we eliminate low and middle-frequency information by setting their frequency bands to zero. Subsequently, in order to retain shift-invariance and local consistency inherent in natural images, we revert the high-frequency spectrum back into RGB format using DCT^{-1}. This process yields the desired representation in the frequency domain, which can be defined as:

$$X_{freq} = \mathcal{D}^{-1}(\mathcal{F}(\mathcal{D}(X_{rgb}))), \tag{1}$$

where $X_{freq} \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB image represented at frequency domain, $\mathcal{D}$ denotes the DCT, $\mathcal{F}$ denotes the filter to obtain high frequency information, and $\mathcal{D}^{-1}$ denotes the DCT^{-1}, which means reverse operation. Then, the initial RGB input is disassembled and subsequently reconstituted as frequency-aware data, all the while preserving spatial relationships. Ultimately, we feed both the RGB and frequency images into the multi-level interaction phase to amplify the forged features.

3.1.2 Multi-Level Interaction

RGB information is valuable in pinpointing abnormal textures within manipulated images, whereas frequency information addresses subtly altered artifacts. Based on the MOA-Transformer (Patel et al., 2022), we take the three

stages as the backbone to realize feature extraction, and each stage features a patch embedding/merging layer and a local transformer block. To further investigate forgery traces, we introduce a Frequency-guided Attention Module inspired by the Convolutional Block Attention Module (CBAM) (Woo et al., 2018). While CBAM derives attention weights from RGB images, we utilize frequency features to generate attention maps and detect the RGB modality from a frequency-oriented viewpoint. After feature extraction by local transformer blocks separately between three stages, we also use FAM to derive the frequency attention map separately as follows,

$$\begin{aligned} \hat{f} &= Conv_{3\times3}(f_{freq}), \\ f_{att} &= \sigma(Conv_{7\times7}(Cat(P_a(\hat{f}), P_m(\hat{f})))), \end{aligned} \tag{2}$$

where f_{freq} denotes the frequency feature after feature extraction, σ denotes the Sigmoid function, P_a and P_m represent global average pooling and global max pooling, respectively. Cat concatenates the features along with the depth. Finally, we opt for a 7×7 convolutional kernel to extract forged traces in the frequency domain. As it is suitable for detecting edge information and covers a more extensive region compared to using three 3×3 convolutional kernels. The resulting attention map, denoted as f_{att}, encompasses subtle forgery traces within the frequency domain. It is challenging to extract forged traces only through the RGB features. Therefore, we implement f_{att} on the RGB feature f_{rgb}, directing f_{rgb} to further mine forgery traces. We can obtain,

$$f_{rgb} = f_{rgb} \oplus (f_{rgb} \otimes f_{att}), \tag{3}$$

where $\oplus$ represents summation and $\otimes$ represents element-wise multiplication.

In addition, our feature extraction process encompasses three distinct levels: low-level, mid-level, and high-level. Low-level features capture texture forgery information, while high-level features provide a more comprehensive view of the overall forgery traces. Consequently, we engage with both RGB and frequency features at multiple levels to obtain a holistic representation of forged features.

Concretely, the output in the frequency domain at the i-th stage, denoted as f^i_{freq}, is employed as the input $\hat{f}^i_{freq}$ for the i+1-th stage. Simultaneously, the input $\hat{f}^{i+1}_{rgb}$ for the i+1-th stage is derived from the RGB feature f^i_{rgb}, which was previously guided in the frequency domain. The input can be computed by,

$$\hat{f}^{i+1}_{rgb} = f^i_{rgb} \oplus (f^i_{rgb} \otimes f^i_{att}), \quad \hat{f}^{i+1}_{freq} = f^i_{freq}. \tag{4}$$

Then we input the high-level output features $f_{rgb} \in \mathbb{R}^{h\times w\times c}$ and $f_{freq} \in \mathbb{R}^{h\times w\times c}$ into multi-modal fusion to mine more discriminative information, where h, w and c are the dimensions of the output features.

3.1.3 Multi-Modal Fusion

In recent years the attention mechanism has been broadly applied in natural language processing (Vaswani et al., 2017) and computer vision (Dosovitskiy et al., 2021). Inspired by these works, the resulting RGB features are combined with frequency features with a Cross-domain Attention Module. Some datasets encompass diverse forgery techniques, resulting in images with varying discriminability in the frequency and RGB domains. Simple methods fail to effectively assign weights based on this discriminability discrepancy. Hence, we designed a dynamic fusion module to allocate quality weights accordingly.

As outlined in Sect. 3.1.2, the frequency modality should serve as a supporting component. With f_{rgb} and f_{freq}, we employ CAM to initially merge these features into a unified representation. This merging process is achieved through the query-key-value mechanism. To be specific, we use 1×1 convolutions to transform f_{rgb} into Q, while f_{freq} is embedded into both K and V.

$$\begin{aligned} Q &= Conv_q(f_{rgb}), \\ K &= Conv_k(f_{freq}), \\ V &= Conv_v(f_{freq}), \end{aligned} \tag{5}$$

where the $Conv_q$, $Conv_k$, $Conv_v$ denote 1×1 convolutions. We execute the attention mechanism by flattening these embeddings into 2D representations ($\hat{Q}$, $\hat{K}$ and $\hat{V} \in \mathbb{R}^{\frac{h\times w}{16}\times c}$) along the channel dimension, resulting in:

$$f_{cam} = softmax\left(\frac{\hat{Q}\hat{K}^T}{\sqrt{h/4 \times w/4 \times c}}\right)\hat{V}, \tag{6}$$

where f_{cam} denotes the preliminary fusion feature which aggregates the RGB and frequency information. To maximize the effective utilization of forged information contained within f_{rgb}, f_{cam}, and f_{freq}, we introduce a Dynamic Fusion Module within the multi-modal fusion phase. In a more detailed breakdown, we feed $S = \{f_{rgb}, f_{cam}, f_{freq}\}$ into the DFM. The DFM computes weights for each branch based on the quality information derived from all branches. These weights are subsequently employed to combine the information from the various branches. To determine the respective weights for each branch, we employ a process that integrates the features from the three branches. This integration involves two fully connected layers denoted as FC_1 and FC_2, global average pooling P_a, and an activation function GELU

denoted as (δ). This process can be written as follows:

$$\begin{aligned} f' &= FC_1(Cat(f_{rgb}, f_{cam}, f_{freq})), \\ f'' &= FC_2(\delta_1(P_a(\delta_2(f')))), \end{aligned} \tag{7}$$

where $f' \in \mathbb{R}^{h\times w\times 3c}$ and $f'' \in \mathbb{R}^{1\times 1\times c}$. Then we set three fully connected layers (F_c^1,F_c^2 and F_c^3) and softmax function to generate quality weights α_i for each branch, which can be formulated as:

$$\alpha_i = \frac{exp(F_c^i(f''))}{\sum_j^3 exp F_c^j(f'')}, i = 1, 2, 3, \tag{8}$$

where $\alpha_i \in \mathbb{R}^{1\times 1\times C}$ represents the quality of each branch. Given that different branches make varying contributions to the extraction of forged clues, we weigh the fusion features according to their quality. To achieve this, we employ two linear mapping layers FC_4 and FC_3 that help restore the channel dimension of the dynamically fused features. This results in the output f_u,

$$f_u = f' + FC_4\left(FC_3\left(\sum_{i=1}^{3}\alpha_i \otimes S_i\right)\right). \tag{9}$$

where the S_i represents the $S = \{f_{rgb}, f_{cam}, f_{freq}\}$ in DFM. Based on the output f_u we can obtain a classifier in real or fake results through the loss function, and we will replace the α_i in the test-time training phase.

3.2 Test-Time Training with Prompt Learning

The first three sections outline the ability of the DDIN network to discern the quality of forgery traces in both RGB and frequency domains, as well as its dynamic fusion process for distinguishing based on quality disparities. To improve the robustness of the forgery detection model, when training and test data come from different distributions, we leverage teat-time training (Sun et al., 2020), which is a new take on the generalization that learns from them at test time. However, when handling uncertain or unseen data, these schemes still can not be generalized effectively. To address this issue, we employ uncertain perturbations as a guiding mechanism during the test-time training phase. Additionally, we introduce a spatial-frequency prompt learning method to significantly enhance the performance of the pre-trained DDIN model when presented with unseen data.

3.2.1 Uncertainty Guidance

To incorporate uncertainty into training, we introduce a perturbation g drawn from Gumbel(0, 1) in the test-time training phase. The Gumbel(0, 1) distribution can be sampled using inverse transform sampling by drawing $u \sim$ Uniform(0, 1) and computing $g = -\log(-\log(u))$ (Gumbel, 1954). We implement g into the Dynamic Fusion Module to impact the evaluation of the network in intra-modal quality. The presence of uncertainty in g results in slight modifications to the quality weight, rendering it probabilistic instead of deterministic. Specifically, the uncertain quality weight β can be determined using the Gumbel softmax function, as described in (Jang et al., 2017):

$$\beta_i = \frac{exp((log(F_c^i(\hat{f})) + g_i)/\tau)}{\sum_j^3 exp((log(F_c^j(\hat{f})) + g_j)/\tau)}, i = 1, 2, 3, \tag{10}$$

where τ represents the softmax temperature. The β value replaces α in Eq. (9), leading to an uncertain distribution of the fused feature f_u. Thus, different from early training, in the test-time training phase, the distribution of f_u becomes uncertain by the weight value β in Eq. (10), and is more dependent on the characteristics of the testing set.

3.2.2 Spatial-Frequency Prompt Learning

In NLP (Brown et al., 2020; Sanh et al., 2022; Houlsby et al., 2019) and computer vision (Jia et al., 2022; Wang et al., 2022b; Dong et al., 2023a) communities, prompt learning methods have gained significant attention for efficiently fine-tuning large foundational models with fewer parameters. To enhance the performance of the pre-trained DDIN on the testing set efficiently, we incorporate the proposed spatial-frequency prompt learning method into the test-time training process.

Considering that our DDIN encompasses both the spatial (i.e., X_{rgb}) and frequency (i.e., X_{freq}) domains of the input, we introduce two prompts, namely $P_s \in \mathbb{R}^{H\times W\times 3}$ for spatial and $P_f \in \mathbb{R}^{H\times W\times 3}$ for frequency. These prompts have the same dimensions as X_{rgb} and X_{freq}, respectively. This format seamlessly integrates with various domains, offering abundant information when used as a prompt. These prompts are used to better leverage knowledge within the forgery detection model. They serve as trainable parameters to enhance the model's performance in a lightweight manner.

During the test-time training phase, P_s and P_f are directly added to the spatial and frequency domains of the input. This simple and effective visual prompt format can be seamlessly incorporated with our DDIN, offering valuable auxiliary information during uncertainty-guided test-time training. In the following experiment, we name spatial-frequency prompt learning with uncertainty-guided test-time training as PUTT.

There are various ways to initialize visual prompts. Considering that our spatial and frequency visual prompts employ the additive way to their corresponding input, both P_s and P_f

are initialized to zeros. The parameters of P_s and P_f gradually expand from zeros to optimized parameters through a learned manner. Through this initialization approach, the early phases of network training remain undisturbed, resulting in a training process that is both accelerated and stable.

3.2.3 Test-Time Training

Building upon the principles of uncertainty guidance and visual prompts, we formulate a self-supervised task within the uncertainty-guided test-time training phase. Specifically, during testing, we sample a testing image x and then input it twice to the pre-trained detector $f(\cdot, \theta)$, where θ represents the model parameters. This process yields two uncertain fused features, namely f_u^1 and f_u^2. These feature distributions align with the actual model output but are influenced by the testing set, featuring perturbations that introduce uncertainty. To measure the distribution shift brought about by uncertainty and update the model parameters, we employ the Kullback–Leibler (KL) divergence loss. This loss serves to narrow the distribution gap between the two features, thereby encouraging the model to perform optimally on the testing set.

3.3 Loss Function

During the early training phase, we flatten the f_u and pass it through a fully connected layer, followed by a sigmoid function to obtain the final predicted probability $\hat{y}$. The classification loss takes the following form,

$$\mathcal{L}_{cls} = y \log \hat{y} + (1 - y) \log(1 - \hat{y}), \tag{11}$$

where y is set to 1 if the image has been manipulated, otherwise it is set to zero.

In the test-time training phase, we add uncertain perturbations to the Dynamic Fusion Module and obtain two uncertain features denoted as f_u^1 and f_u^2. To narrow the two feature distributions, we use KL divergence loss as follows,

$$\mathcal{L}_{kl} = \mathcal{D}_{kl}(f_u^1 \| f_u^2) = f_u^1 \log(f_u^1) - f_u^1 \log(f_u^2). \tag{12}$$

4 Experiments

In this section, we evaluate the proposed method against the state-of-the-art methods on multiple datasets. First, we introduce experimental settings and datasets. Then, we conduct experiments using the close-set and open-set settings under the Celeb-DF, FaceForensics++, DFDC, and our proposed FaceDiffusionForensics datasets. Finally, we present ablation studies and visualization results.

4.1 Experimental Setting

4.1.1 Implementation Details

The backbone network is modified from MOA-Transformer (Patel et al., 2022), which is pre-trained on ImageNet. The structure of the MOA Transformer unfolds across three stages, where each stage features a patch embedding/merging layer and a local transformer block. MOA applies a global attention-based module in the local window-based transformer after each stage. Additionally, a global multi-resolution overlapped attention module is incorporated after each stage, except for the final one. We use the DLIB (Sagonas et al., 2016) for face extraction and alignment. The face images are resized to 224×224 with random erase. The τ in Eq. (10) is set to 1, and the batch size of the training and test-time training phase are all set to 32. We use the Adam optimizer for optimizing the network with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rates for the training and test-time training phase are set to $1e{-}5$ and $1e{-}4$, respectively. During the test-time training phase, we only update the parameters in DFM and freeze the parameters of other layers.

Evaluation Metrics. We apply the Accuracy score (Acc) and Area Under the Receiver Operating Characteristic Curve (AUC) as evaluation metrics. The results of the comparison methods are taken from their own paper unless otherwise specified.

4.2 Datasets

We evaluate our proposed method on commonly used datasets, including FaceForensics++ (FF++) (Rössler et al., 2019b), CelebDF (Li et al., 2020b), and DFDC (Dolhansky et al., 2019). FF++ is based on four known manipulations including DeepFakes, Face2Face, FaceSwap and NeuralTextures. FF++ consists of 1000 videos, of which 720 videos are used for training, 140 videos are reserved for validation, and 140 videos for testing. CelebDF includes 590 real videos and 5,639 high-quality fake videos generated by the improved DeepFake algorithm (Tora, 2021). DFDC is a large-scale dataset that contains 19,197 real videos and 100,000 fake videos produced with several Deepfake, GAN-based, and non-learned methods.

In addition, we build a dataset, dubbed FaceDiffusionForensics (FDF), using the newly proposed techniques based on diffusion models. Diffusion-based methods (Liu et al., 2022b; Zhang & Chen, 2023; Rombach et al., 2022) bring new challenges for forgery detection but are less explored in the existing literature. They can generate high-resolution images with fine details. More importantly, the diffusion models update rapidly due to the efforts from the open source communities. That means the forgery detection methods have to recognize those images generated by unseen techniques.

Table 1 Composition of the proposed FaceDiffusion Forensics dataset

Dataset split	Source	Type	Forgery technique	Number of images
Training	FFHQ	Real	–	10k
		Image2Image	PNDM (Liu et al., 2022b)	10k
			DEIS (Zhang & Chen, 2023)	10k
			IBD (Gao et al., 2021)	10k
		Text2Image	PNDM (Liu et al., 2022b)	10k
			DEIS (Zhang & Chen, 2023)	10k
Testing	FFHQ	Real	–	60k
		Image2Image	DEIS (Zhang & Chen, 2023)	20k
			IBD (Gao et al., 2021)	20k
		Text2Image	PNDM (Liu et al., 2022b)	20k
	CelebA-HQ	Real	–	30k
		Image2Image	PNDM (Liu et al., 2022b)	10k
			DEIS (Zhang & Chen, 2023)	10k
			IBD (Gao et al., 2021)	10k

The training set consists of 10k real pictures and 50k synthetic images. The testing set consists of 90k real pictures and 90k synthetic images

Table 2 Quantitative results on Celeb-DF dataset and FaceForensics++ dataset with different quality settings

Method	FF++ (C23)		FF++ (C40)		Celeb-DF	
	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)
MesoNet (Afchar et al., 2018)	83.10	–	70.47	–	–	–
Multi-task (Nguyen et al., 2019a)	85.65	85.43	81.30	75.59	–	–
Xception (Rössler et al., 2019a)	95.73	–	81.00	–	–	–
Face X-ray (Li et al., 2020a)	–	87.40	–	61.60	–	–
Two-branch (Masi et al., 2020)	96.43	98.70	86.34	86.59	–	–
RFM (Wang & Deng, 2021)	95.69	98.79	87.06	89.83	97.96	**99.94**
F3-Net (Qian et al., 2020)	97.52	98.10	90.43	93.30	95.95	98.93
Add-Net (Zi et al., 2020)	96.78	97.74	87.50	91.01	96.93	99.55
FDFL (Li et al., 2021a)	96.69	99.30	89.00	92.40	–	–
MultiAtt (Zhao et al., 2021)	97.60	99.29	88.69	90.40	97.92	**99.94**
HFI-Net (Miao et al., 2022)	91.87	97.07	85.69	88.40	–	–
LiSiam (Wang et al., 2022a)	95.51	99.13	87.81	91.44	–	–
F2Trans (Miao et al., 2023)	96.09	99.18	86.98	89.60	–	–
PEL (Gu et al., 2022)	*97.63*	*99.32*	*90.52*	94.28	–	–
DDIN (Ours)	97.59	99.31	90.41	*94.47*	*98.02*	99.83
DDIN (+PUTT) (Ours)	**97.91**	**99.59**	**91.23**	**95.16**	**98.60**	**99.94**

The best results are shown in bold, and the second best results are shown in italic. The **DDIN(+PUTT)** represents DDIN that allows test-time training with uncertainty-guided and spatial-frequency prompt learning

Therefore, the detection of diffusion-based forged images can be considered a significant challenge. In the following experiments, we use the new FDF dataset to verify the generalization of our approach and the competing methods. Below, we first describe the dataset composition and testing protocols.

We collect 100k real face images from two high-quality face datasets, Flickr-Faces-HQ (FFHQ) (Karras et al., 2021) and CelebA-HQ (Karras et al., 2018). The face images cover considerable real-world variation in terms of age, ethnicity, pose, and image background. We produce 140k synthesized images using two diffusion-based methods (PNDM (Liu et al., 2022b), DEIS (Zhang & Chen, 2023)) and an identity swapping method IBD (Gao et al., 2021).

Specifically, we perform the image-to-image and text-to-image settings to synthesize images. In the text-to-image setting, we use the positive text prompts to change the hairstyle and skin tone. We add negative prompts to make

synthesized results more realistic and less error-prone. In the image-to-image setting, the real image is an additional input to maintain the structure of the original image. The generate strength and guidance scale are set to 0.75 and 7.5, respectively.

As shown in Table 1, we split the FDF dataset into a training set and a testing set. This dataset is specifically designed for cross-testing, and the testing set is larger than the training set. To build the training set, we use 10k images from FFHQ to generate 50k synthetic images with IBD method (Gao et al., 2021) and diffusion models (PNDM (Liu et al., 2022b) and DEIS (Zhang & Chen, 2023)) by image-to-image synthesis and text-to-image synthesis. To build the testing set, we use the remaining 60k images from FFHQ and 30k images from the CelebA-HQ dataset. We use these images to synthesize 90k fake images. The testing set is divided into three subsets according to different manipulation models (PNDM, DEIS, and IBD) for comprehensive experiments. Specifically, each of the subsets contains 30K real images and 30K fake images. It's notable that the face identities in the training and testing sets do not overlap, which makes the forgery detection task considerably more challenging.

4.3 Experimental Results

In this section, we compare the proposed method against existing forgery detection methods on the above mentioned datasets (Rössler et al., 2019b; Li et al., 2020b; Dolhansky et al., 2019). We evaluate the performances under two settings, i.e., intra-testing and cross-testing. Concretely, in the former setting, the methods for forgery are known. By contrast, in the latter setting, fake samples are synthesized by some methods unknown to the detection models.

4.3.1 Intra-Testing

As shown in Table 2, our proposed method consistently outperforms all competing methods by a considerable margin on the FF++ dataset. For example, compared with the PEL method (Gu et al., 2022), the AUC of our method exceeds it by 0.27% and 0.88% under the two quality settings (C23 and C40), and the Acc also obtains considerable improvements. To explain, our method considers the quality of the auxiliary discriminant information contained and thus improves the discriminative ability of the network on the testing set. On the Celeb-DF dataset, the AUC of the state-of-the-art methods is close to 100%, but the Acc of our method still exceeds the second best by 0.64%. We can observe that DDIN alone falls short compared to PEL (Gu et al., 2022) in C23. However, when combined with PUTT, it surpasses PEL significantly. The above results demonstrate the effectiveness of the proposed framework (DDIN) and test-time strategy with uncertainty guidance and spatial-frequency prompt learning (DDIN+PUTT).

4.3.2 Cross-Testing

To evaluate the generalization ability of our method to unknown forgery types, we conduct fine-grained cross-testing. Specifically, we choose the training and testing sets from all possible combinations from four subsets (DF, F2F, FS, NT) in the FF++ dataset. The evaluation results are shown in Table 3. Notably, we compare our method against approaches that focus on specific forgery types like FDFL (Li et al., 2021a) and MultiAtt (Zhao et al., 2021). The comparison results show that our method generally outperforms the other methods.

The generalization performance of forgery detection models is also affected by the emergence of new manipulation methods. To further demonstrate the generalization ability of detection models when facing unseen manipulation types, we conduct performance on multi-source manipulation evaluation on FF++ (C23 and C40). In Table 4, it can be observed that the proposed DDIN+PUTT significantly outperforms other forgery detectors. Despite the different compression rates and manipulation methods of GID-DF and GID-FF, our method can also learn dynamic fusion features and generalize to detect unseen forged data with spatial-frequency prompt learning and uncertainty-guided test-time training. For instance, the proposed DDIN improves by 1.31%, while PUTT further boosts by 2.95% in terms of frame-level Acc compared with the current similar-sized SOTA F2Trans-S, on highly compressed GID-DF (C23).

We further conduct cross-dataset experiments by choosing training and testing sets from different datasets. Specifically, we train the models on the FF++ dataset and then test the trained models on the Celeb-DF and DFDC datasets. As shown in Table 5, we observe that our method outperforms the competing methods well in cross-testing. On the DFDC dataset, the AUC scores of the competing methods are about 70%, which degrade significantly compared with those under the intra-testing. However, benefiting from the proposed DDIN framework and test-time fine-turning strategy, our method achieves better generalizability and performs favorably against all the competing methods.

To evaluate the generalization capabilities of different methods to diffusion models, we conduct comparison experiments on the proposed FDF dataset. We evaluate our method on the three subsets (PNDM, DEIS and IBD) of the testing set separately and compute the average value. For each subset, a comparative analysis has been conducted, pitting our model against several established methods (F3Net, CNNDetection). Specifically, DDIN(+PUTT) and F3Net are pretrained on the FF++ dataset, while CNNDetection is pretrained on a dataset derived from the LSUN dataset using ProGAN. The

Table 3 Cross-manipulation evaluation on the subsets of FF++(C40) in terms of AUC(%)

Method	Train	DF	F2F	FS	NT	Cross Avg.
FDFL (Li et al., 2021a)	**DF**	*98.91*	58.90	66.87	63.61	63.13
MultiAtt (Zhao et al., 2021)		*99.51*	66.41	67.33	66.01	66.58
DDIN *(ours)*		*99.71*	61.99	78.08	67.02	69.03
DDIN (+PUTT) *(ours)*		***99.79***	**66.80**	**78.93**	**68.03**	**71.25**
FDFL (Li et al., 2021a)	**F2F**	67.55	*93.06*	55.35	66.66	63.19
MultiAtt (Zhao et al., 2021)		73.04	*97.96*	**65.10**	71.88	70.01
DDIN *(ours)*		73.85	*98.01*	64.25	72.49	70.19
DDIN (+PUTT) *(ours)*		**77.36**	***98.13***	64.80	**74.91**	**72.36**
FDFL (Li et al., 2021a)	**FS**	75.90	54.64	*98.37*	49.72	60.09
MultiAtt (Zhao et al., 2021)		82.33	61.65	***98.82***	54.79	66.26
DDIN (ours)		88.20	62.13	*98.80*	56.63	68.98
DDIN (+PUTT) *(ours)*		**89.38**	**62.71**	*98.81*	**58.82**	**70.30**
FDFL (Li et al., 2021a)	**NT**	79.09	74.21	53.99	*88.54*	69.10
MultiAtt (Zhao et al., 2021)		74.56	80.61	60.90	*93.34*	72.02
DDIN (ours)		78.15	81.34	62.67	*93.34*	74.05
DDIN (+PUTT) (ours)		**79.81**	**86.05**	**64.13**	***94.31***	**76.66**

Italic indicates intra-dataset results and Cross Avg. means the average of cross-method results. The best results are shown in bold

Table 4 Performance on multi-source manipulation evaluation

Method	GID-DF(C23)		GID-DF(C40)		GID-F2F(C23)		GID-F2F(C40)	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
EffcientNet (Tan & Le, 2019)	82.40	91.11	67.60	75.30	63.32	80.10	61.41	67.40
ForensicTransfer (Cozzolino et al., 2018)	72.01	–	68.20	–	64.50	–	55.00	–
Multi-task (Nguyen et al., 2019b)	70.30	–	66.76	–	58.74	–	56.50	–
MLDG (Li et al., 2018)	84.21	91.82	67.15	73.12	63.46	77.10	58.12	61.70
LTW (Sun et al., 2021)	85.60	92.70	69.15	75.60	65.60	80.20	65.70	72.40
DCL (Sun et al., 2022)	87.70	94.9	75.90	83.82	68.40	82.93	67.85	75.07
IID (Huang et al., 2023a)	88.21	95.03	76.90	84.55	69.36	84.37	67.99	74.80
F2Trans-S Miao et al. (2023)	89.64	97.47	77.86	86.92	81.43	90.55	66.79	76.52
DDIN	91.01	97.85	78.75	87.35	82.35	91.30	68.96	78.79
DDIN+PUTT	93.96	98.91	79.35	87.57	82.91	91.43	69.33	79.83

GID-DF means training on the other three manipulated methods of FF++ and testing on DeepFakes and the same for the others

Table 5 Cross-testing in terms of AUC (%) by training on FF++(C40)

Method	CelebDF	DFDC
Xception (Rössler et al., 2019b)	61.80	63.61
RFM (Wang & Deng, 2021)	65.63	66.01
Add-Net (Zi et al., 2020)	65.29	64.78
F3-Net (Qian et al., 2020)	61.51	64.60
MultiAtt (Zhao et al., 2021)	67.02	68.01
PEL (Gu et al., 2022)	69.18	63.31
DDIN *(ours)*	68.35	68.80
DDIN(+PUTT) (ours)	**69.68**	**71.31**

The best results are shown in bold

evaluation results are presented in Table 6. An interesting observation is that the performance of Xception (Rössler et al., 2019b) is even worse than a random guess. This could potentially be attributed to the pretraining on GAN-generated fake samples. The model excessively focuses on identifying these synthetic GAN samples, ultimately impeding its ability to detect counterfeit samples from the diffusion models. The results show that our model outperforms the others in terms of the average AUC and Acc. Furthermore, the utilization of the PUTT method exhibits promising potential to further enhance the performance of our model.

4.4 Ablation Study

In this section, we will analyze the different components of DDIN and the test-training scheme with or without the spatial-frequency prompt. The CAM module is analyzed in Sect. 4.4.2 to evaluate different feature fusion methods before dynamic fusion. Finally, we also evaluate the effectiveness of the frequency attention in our feature extraction process, which encompasses three distinct levels.

4.4.1 Components

As shown in Table 7, we develop several variants and conduct a series of experiments on the FF++ (C40) dataset to explore the influence of different components in our proposed method. The comparison results show that the AUC and Acc of the baseline model increase by 3.01% and 2.72%, respectively. Model performance can be improved by the FAM and further by the DFM as shown in (b),(c), and (d). In addition, according to the DFM integrates frequency domain information, by comparing (a) and (c), the results verify that the frequency information is distinct and complementary to the RGB information and the quality discrepancies are negligible. Furthermore, uncertainty-guided test-time training (UTT) is also effective in boosting performance as shown in (d) and (f). A noticeable decrease in performance is observed when FAM is not applied in (e) and (f). The best performance is achieved when combining all the proposed components with Acc and AUC of 91.23% and 95.16%, respectively. When spatial-frequency prompt learning is applied, there is almost a 0.4% increment.

4.4.2 Different Feature Fusion Methods

We investigate the preliminary feature fusion method before the dynamic fusion module, including concat fusion and cross-domain attention fusion, which can verify the effectiveness of the CAM module. Table 8 shows the results of three different scenarios. We observe that adding the **Concat** fusion method improves model performance, which shows that the initial feature fusion is required. The result of adding the **CrossAtt** fusion method shows that the CAM module can effectively supplement the RGB and frequency domain feature information.

4.4.3 Multi-Level Interaction

Moreover, in the multi-level interaction stage, our feature extraction process encompasses three distinct levels: low-level, mid-level, and high-level. Low-level features capture texture forgery information, while high-level features provide a more comprehensive view of the overall forgery traces.

Table 6 Quantitative results of different methods on the three testing subsets (PNDM, DEIS, IBD) of the proposed FaceDiffusionForensics (FDF) dataset

Method	PNDM		DEIS		IBD		Average	
	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)
Xception Rössler et al. (2019a)	40.94	49.76	37.84	49.99	46.38	46.38	41.72	48.71
CNNDetection Wang et al. (2020)	57.18	51.79	68.78	55.18	50.22	49.69	58.73	52.22
F3Net Qian et al. (2020)	55.72	57.79	43.51	48.89	44.65	44.65	47.96	50.44
SBIs Shiohara and Yamasaki (2022)	58.65	49.73	64.37	50.04	83.75	50.00	68.92	49.92
DDIN *(ours)*	63.82	59.97	71.14	64.61	71.66	63.09	68.87	62.56
DDIN+PUTT *(ours)*	64.22	60.59	71.57	65.38	71.68	63.14	**69.16**	**63.04**

We also report the average scores and the best results are shown in bold

Table 7 Ablation study of the proposed method on the FF++ dataset

Ablation Study	Modules				FF++ (C40)	
	FAM	DFM	UTT	PUTT	AUC(%)	Acc(%)
(a)	–	–	–	–	92.15	88.51
(b)	✓	–	–	–	93.43	89.73
(c)	–	✓	–	–	93.82	90.21
(d)	✓	✓	–	–	94.47	90.41
(e)	–	✓	✓	–	94.02	90.23
(f)	✓	✓	✓	–	94.80	90.84
(g)	✓	✓	✓	✓	95.16	91.23

The **UTT** represents the test-time training phase with uncertainty-guided, while the **PUTT** represents the training with uncertainty-guided and spatial-frequency prompt learning

Table 8 Ablation study of different feature fusion methods before dynamic fusion in the multi-modal fusion phas.

Ablation study	Fusion method		FF++(C40)	
	Concat	CrossAtt	AUC (%)	Acc (%)
(a)	–	–	93.47	89.79
(b)	✓	–	94.26	90.33
(c)	–	✓	95.16	91.23

Concat and CrossAtt represent concatenate the feature and cross-domain attention mechanism respectively

Table 9 Ablation study of the multi-level Interaction on FF++(C40), and the results are without taking spatial-frequency prompt learning

Ablation Study	Interaction level			FF++(C40)	
	Low	Mid	High	AUC (%)	Acc (%)
(a)	✓	–	–	94.13	90.25
(b)	–	✓	–	94.21	90.33
(c)	–	–	✓	94.25	90.36
(d)	✓	✓	–	94.36	90.41
(e)	✓	–	✓	94.44	90.52
(f)	–	✓	✓	94.46	90.60
(g)	✓	✓	✓	94.80	90.84

Consequently, we engage with both RGB and frequency features at multiple levels to obtain a holistic representation of forged features. As shown in Table 9, we seek a suitable interaction method by exploring FAM in different feature exaction stages and conducting a series of experiments on the FF++(C40) dataset, which include the frequency domain and RGB domain interactions for all three stages of the baseline. It can be seen that interacting at a single level can slightly improve performance and it turns out that interacting at every level achieves the best results.

4.5 Visualization

To better understand the decision-making mechanism of our method, we provide the Grad-CAM (Selvaraju et al., 2020) visualization on FF++ as shown in Fig. 4. It can be observed that the baseline method MOA-Transformer (Patel et al., 2022) tends to overlook forged traces in fake faces, particularly those that are hidden within the RGB domain. In contrast, even though it only uses binary labels for training, our method generates distinguishable heatmaps for real and fake faces, with the prominent regions varying in forgery techniques. For example, when detecting images forged with Deepfakes (Tora, 2021) and NeuralTextures (Thies et al., 2019a) technologies, our method focuses on the edge contours of artifacts, which are difficult to detect in the RGB domain. Our method is more effective for the abnormal texture information forged by FaceSwap (Kowalski, 2021) in the eyes region, and also effective for the inconsistent information forged by Face2Face (Thies et al., 2019b) in the mouth region. The results provide a decision-making perspective on the effectiveness of DINN.

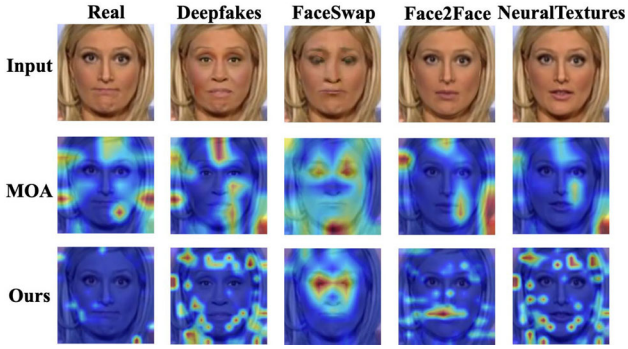

Fig. 4 Attention heatmaps on real and fake samples produced by our method and MOA (baseline architecture). We use the Grad-CAM (Selvaraju et al., 2020) method for visualization

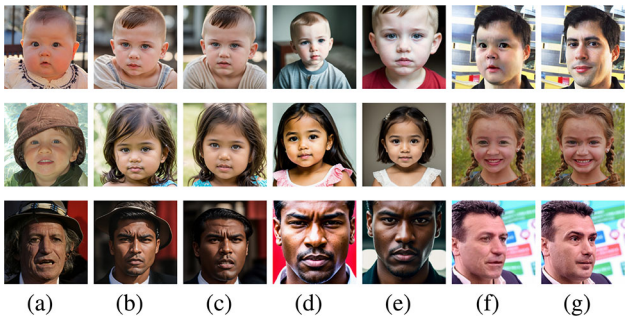

Fig. 5 Exemplify the real images (**a**) and synthesis using PNDM image-to-image (**b**), DEIS image-to-image (**c**), PNDM text-to-image (**d**), DEIS text-to-image (**e**), IBD method (**f**), another real picture for swapping faces (**g**)

In order to better demonstrate the distribution of the FaceDiffusionForensics dataset, Fig. 5 exemplifies the synthesis using the two diffusion methods (PNDM, DEIS) image-to-image, text-to-image, and IBD method. The prompt template of the diffusion method is "A very close-up portrait photo of a {person} with {hairstyle} hair, {skin_tone} skin tone, {expression} expression, with a {face_type}-shaped face, and with a background of {background_option}. (highly detailed skin:1.2), 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3", where {person}, {hairstyle}, {skin_tone}, {expression}, {face_type} and {background_option} attributes are selected randomly from custom candidate lists.

4.6 Discussion

The advent of deep learning has led to significant advancements in various computer vision tasks. One prominent domain is the detection of manipulated or forged facial images, which has gained importance due to the rise of AI-generated content. We leverage insights from spatial-frequency prompt and test-time training to learn intricate

patterns and artifacts introduced during the forgery process to enhance the model ability. By fusing information of spatial and frequency, the model can potentially identify subtle manipulations that might go unnoticed using traditional detection methods. The experimental results demonstrate the effectiveness of the test-time trained model. The model learned with spatial-frequency prompts outperforms baseline models on various benchmark datasets, enhancing the ability to make accurate forgery predictions. Especially in datasets containing both diffusion and non-diffusion images, it is evident that utilizing prompt learning enhances the performance of diffusion methods detection. This also paves the way for future research in more general intelligent detection.

However, since the proposed test-time training scheme needs to optimize the model parameters by traversing the testing set, this undoubtedly increases the inference time. Therefore, the network may take a long time to converge when the amount of test data is large. In the meanwhile, the detection accuracy of our method on the FDF dataset is not constantly stable. It is our future work to develop methods that aim for both GAN-synthesized and diffusion-synthesized fake images. Therefore, ongoing research and development are necessary for the effectiveness of the forgery detection model in real-world applications.

(1) Dataset diversity: Continuously updating and expanding the training datasets to include a diverse set of forgery techniques and variations.
(2) Explainability: Developing methods to interpret predictions of the model could enhance its trustworthiness and facilitate its adoption in critical applications.
(3) Model traceability: Ensuring the traceability of forgery detection models is a critical aspect of advancing the field of forgery detection and maintaining ethical standards in digital content verification.

5 Conclusion

In this paper, we have introduced a Dynamic Dual-spectrum Interaction Network for face forgery detection. DDIN is designed to address feature quality discrepancies and model uncertainty through test-time training. It incorporates a frequency-guided attention module for multi-level interaction and a dynamic fusion module for multi-modal fusion, enabling dynamic feature fusion to reconcile discrepancies in feature quality. Additionally, uncertainty-guided test-time training is proposed to fine-tune the trained detector by incorporating uncertain perturbations. Furthermore, we enhance network generalization through spatial-frequency prompt learning. We build a large-scale image forgery dataset that contains high-quality samples synthesized by diffusion models and non-diffusion models. Our extensive experimental results and visualizations demonstrate the efficacy of our approach over state-of-the-art methods and verify the benefits of test-time training guided by uncertainty and spatial-frequency prompt learning.

Acknowledgements This work is partially funded by the National Natural Science Foundation of China (Grant Nos. U21B2045, U20A20223, 32341009, 62206277), Youth Innovation Promotion Association CAS (Grant No. 2022132), and Beijing Nova Program (20230484276).

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: A compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, pp 1–7.

Bai, W., Liu, Y., Zhang, Z., Li, B., & Hu, W. (2023). Aunet: Learning relations between action units for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 24709–24719.

Bartler, A., Bühler, A., Wiewel, F., Döbler, M., & Yang, B. (2022). MT3: Meta test-time training for self-supervised test-time adaption. *International Conference on Artificial Intelligence and Statistics, 151*, 3080–3090.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., & Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 4103–4112.

Chen, A., Yao, Y., Chen, P., Zhang, Y., & Liu, S. (2022a). Understanding and improving visual prompting: A label-mapping perspective. arxiv:abs/2211.11635

Chen, L., Zhang, Y., Song, Y., Wang, J., & Liu, L. (2022b). OST: Improving generalization of deepfake detection via one-shot test-time training. In: Advances in Neural Information Processing Systems.

Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., & Ji, R. (2021). Local relation learning for face forgery detection. In *Thirty-Fifth Conference on Artificial Intelligence*, AAAI, pp 1081–1088.

Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., & Verdoliva, L. (2018). Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arxiv: abs/1812.02510

Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, AK. (2020). On the detection of digital face manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 5780–5789.

Das, S., Seferbekov, S.S., Datta, A., Islam, MS., & Amin, MR. (2021). Towards solving the deepfake problem : An analysis on improving deepfake detection using dynamic face augmentation. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp 3769–3778.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton-Ferrer, C. (2019). The deepfake detection challenge (DFDC) preview dataset. arxiv: abs/1910.08854

Dong, B., Zhou, P., Yan, S., & Zuo, W. (2023a). LPT: Long-tailed prompt tuning for image classification. In *The Eleventh International Conference on Learning Representations*.

Dong, C., Chen, X., Hu, R., Cao, J., & Li, X. (2023). Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(3), 3539–3553.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly,

S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Gao, G., Huang, H., Fu, C., Li, Z., & He, R. (2021). Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 3404–3413.

Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., & Yi, R. (2022). Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Thirty-Sixth Conference on Artificial Intelligence*, AAAI, pp 735–743.

Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., & Verdoliva, L. (2023). Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 20606–20615.

Gumbel, E.J. (1954). Statistical theory of extreme values and some practical applications: A series of lectures. 33.

Guo, H., Wang, H., & Ji, Q. (2022). Uncertainty-guided probabilistic transformer for complex action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp 20020–20029.

Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., & Liu, X. (2023). Hierarchical fine-grained image forgery detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 3155–3165.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In*Proceedings of the 36th International Conference on Machine Learning, 97*, 2790–2799.

Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., & Ye, D. (2023a). Implicit identity driven deepfake face swapping detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4490–4499.

Huang, Q., Dong, X., Chen, D., Zhang, W., Wang, F., Hua, G., & Yu, N. (2023b). Diversity-aware meta visual prompting. arxiv: abs/2303.08138

Huang, S., Huang, H., Wang, Z., Xu, N., Zheng, A., & He, R. (2023c). Uncertainty-guided test-time training for face forgery detection. In: Pattern Recognition—7th Asian Conference, ACPR, vol 14407, pp 258–272.

Jang, E., Gu, S., & Poole, B. (2017). Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

Jia, M., Tang, L., Chen, B., Cardie, C., Belongie, S.J., Hariharan, B., & Lim, S. (2022). Visual prompt tuning. arxiv: abs/2203.12119

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations*.

Karras, T., Laine, S., & Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(12), 4217–4228.

Kowalski, M. (2021). Faceswap. https://github.com/marekkowalski/faceswap, 2022, April 7

Kwon, M., Nam, S., Yu, I., Lee, H., & Kim, C. (2022). Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision, 130*(8), 1875–1895.

Li, D., Yang, Y., Song, Y., & Hospedales, T.M. (2018). Learning to generalize: Meta-learning for domain generalization. In McIlraith SA, Weinberger KQ (eds) *Thirty-Second AAAI Conference on Artificial Intelligence*, pp 3490–3497.

Li, D., Zhu, J., Wang, M., Liu, J., Fu, X., & Zha, Z. (2023). Edge-aware regional message passing controller for image forgery localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 8222–8232.

Li, J., Xie, H., Li, J., Wang, Z., & Zhang, Y. (2021a). Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp 6458–6467.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020a). Face x-ray for more general face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 5000–5009.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020b). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 3204–3213.

Li, Y., Hao, M., Di, Z., Gundavarapu, N.B., & Wang, X. (2021b). Test-time personalization with a transformer for human pose estimation. In *Advances in Neural Information Processing Systems*, pp 2583–2597.

Liu, D., Dang, Z., Peng, C., Zheng, Y., Li, S., Wang, N., & Gao, X. (2023). Fedforgery: Generalized face forgery detection with residual federated learning. *IEEE Transactions on Information Forensics and Security, 18*, 4272–4284.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. (2021a). Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp 772–781.

Liu, H., Wu, Z., Li, L., Salehkalaibar, S., Chen, J., & Wang, K. (2022a). Towards multi-domain single image dehazing via test-time training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 5821–5830.

Liu, L., Ren, Y., Lin, Z., & Zhao, Z. (2022b). Pseudo numerical methods for diffusion models on manifolds. In *The Tenth International Conference on Learning Representations*.

Liu, Y., Kothari, P., van Delft, B., Bellot-Gurlet, B., Mordan, T., & Alahi, A. (2021b). TTT++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, pp 21808–21820.

Luo, Y., Zhang, Y., Yan, J., & Liu, W. (2021). Generalizing face forgery detection with high-frequency features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp 16317–16326.

Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. (2020). Two-branch recurrent network for isolating deepfakes in videos. *European Conference on Computer Vision, 12352*, 667–684.

Miao, C., Tan, Z., Chu, Q., Yu, N., & Guo, G. (2022). Hierarchical frequency-assisted interactive networks for face manipulation detection. *IEEE Transactions on Information Forensics and Security, 17*, 3008–3021.

Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., & Yu, N. (2023). F^2trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security, 18*, 1039–1051.

Nguyen, H.H., Fang, F., Yamagishi, J., & Echizen, I. (2019a). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pp 1–8.

Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. *10th IEEE International Conference on Biometrics Theory* (pp. 1–8). BTAS: Applications and Systems.

Patel, K., Bur, A.M., Li, F., & Wang, G. (2022). Aggregating global features into local vision transformer. In *International Conference on Pattern Recognition*. pp 1141–1147.

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. *European Conference on Computer Vision, 12357*, 86–103.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10674–10685.
Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019a). Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision*, pp 1–11.
Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019b). Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision*, pp 1–11.
Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing, 47*, 3–18.
Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., & Dey, M., et al. (2022). Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision, 128*(2), 336–359.
Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 18699–18708.
Shu, M., Nie, W., Huang, D., Yu, Z., Goldstein, T., Anandkumar, A., & Xiao, C. (2022). Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*.
Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 6048–6058.
Sun, K., Liu, H., Ye, Q., Gao, Y., Liu, J., Shao, L., & Ji, R. (2021). Domain general face forgery detection by learning to weight. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp 2638–2646.
Sun, K., Yao, T., Chen, S., Ding, S., Li, J., & Ji, R. (2022). Dual contrastive learning for general face forgery detection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp 2316–2324.
Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., & Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning, 119*, 9229–9248.
Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B., & Khoreva, A. (2022). OASIS: Only adversarial supervision for semantic image synthesis. *International Journal of Computer Vision, 130*(12), 2903–2923.
Tan, M., & Le, Q.V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R (eds) *Proceedings of the 36th International Conference on Machine Learning*, ICML, vol 97, pp 6105–6114.
Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics, 66*(1–66), 12.
Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2019). Face2face: Real-time face capture and reenactment of RGB videos. *Communications of the ACM, 62*(1), 96–104.
Tora. (2021). Deepfakes. https://github.com/deepfakes/faceswap, 2022, March 5.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pp 5998–6008.
Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp 14923–14932.
Wang, D., Shelhamer, E., Liu, S., Olshausen, B.A., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.
Wang, J., Sun, Y., & Tang, J. (2022). Lisiam: Localization invariance siamese network for deepfake detection. *IEEE Transactions on Information Forensics and Security, 17*, 2425–2436.
Wang, S., Wang, O., Zhang, R., Owens, A., & Efros, A.A. (2020). Cnn-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 8692–8701.
Wang, Z., Zhang, Z., Lee, C., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J.G., & Pfister, T. (2022b). Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 139–149.
Wang, Z., Bao, J., Zhou, W., Wang, W., & Li, H. (2023). Altfreezing for more general video face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 4129–4138.
Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. *European Conference on Computer Vision, 11211*, 3–19.
Yang, Y., Tan, Z., Tiwari, P., Pandey, H. M., Wan, J., Lei, Z., Guo, G., & Li, S. Z. (2021). Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *International Journal of Computer Vision, 129*(10), 2731–2744.
Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T., & Sun, M. (2021). CPT: Colorful prompt tuning for pre-trained vision-language models. arxiv: abs/2109.11797
Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., & Finn, C. (2021). Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, pp 23664–23678.
Zhang, Q., & Chen, Y. (2023). Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*.
Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., & Li, H. (2022). Tip-adapter: Training-free adaption of CLIP for few-shot classification. *European Conference on Computer Vision, 13695*, 493–510.
Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2185–2194.
Zhou, K., Yang, J., Loy, C.C., & Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 16795–16804.
Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*(9), 2337–2348.
Zhu, X., Fei, H., Zhang, B., Zhang, T., Zhang, X., Li, S. Z., & Lei, Z. (2023). Face forgery detection by 3d decomposition and composition search. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(7), 8342–8357.
Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *The 28th ACM International Conference on Multimedia*, pp 2382–2390.