Uncertainty-guided Test-time Training for Face Forgery Detection

Shenyuan Huang[1,5], Huaibo Huang[5], Zi Wang[1], Nan Xu[5], Aihua Zheng[2,3,4], and Ran He[5(✉)]

[1] School of Computer Science and Technology, Anhui University, Hefei, China
[2] Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei, China
[3] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei, China
[4] School of Artificial Intelligence, Anhui University, Hefei, China
[5] NLPR, CRIPAC, CASIA, Beijing, China
{hsywatchingu, ziwang1121, ahzheng214}@foxmail.com, huaibo.huang@cripac.ia.ac.cn, xunan2015@ia.ac.cn, rhe@nlpr.ia.ac.cn

Abstract. Face forgery detection is becoming increasingly important in computer vision as facial manipulation technologies cause serious concerns. Recent works have resorted to the frequency domain to develop face forgery detectors and achieved better generalization achievements. However, there are still unignorable problems: a) the role of frequency is not always sufficiently effective and generalized to different forgery technologies; and b) the network trained on public datasets is unable to effectively quantify its uncertainty. To address the generalization issue, we design a Dynamic Dual-spectrum Interaction Network (DDIN) that allows test-time training with uncertainty guidance. RGB and frequency features are first interacted in multi-level by using a Frequency-guided Attention Module (FAM). Then these multi-modal features are merged with a Dynamic Fusion Module (DFM). Moreover, we further exploit uncertain perturbations as guidance during the test-time training phase. The network can dynamically fuse the features with quality discrepancies, thus improving the generalization of forgery detection. Comprehensive evaluations of several benchmark databases corroborate the superior generalization performance of DDIN.

Keywords: Face Forgery Detection · Frequency Domain · Test-time training · Generalization.

1 Introduction

Recent years have witnessed significant progress in the area of face forgery technology. The quality of fake media has been greatly improved with the development of deep learning technology. At the same time, these forged media may be abused for malicious purposes, causing severe trust issues and security concerns

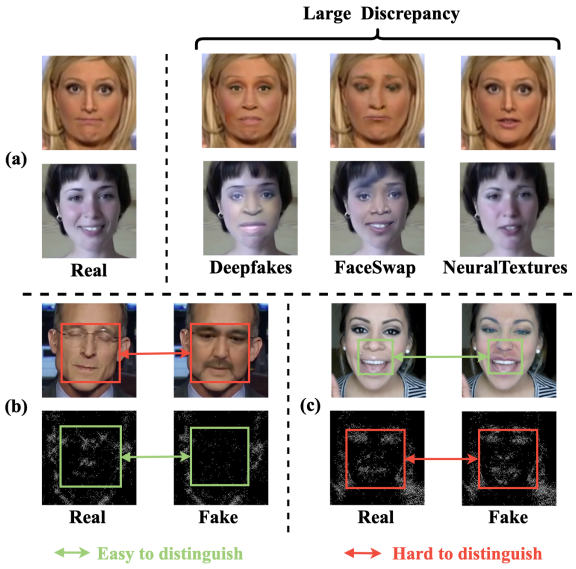

Fig. 1. In the first two rows, the real images are compared to images synthesized by different forgery techniques. In the last two rows, we show the quality difference in RGB and frequency domains. The red box indicates that the forgery traces are hard to recognize, while the green box shows that it is easy to distinguish.

in our society. Therefore, it is critical to developing effective forgery detection methods.

Most of the existing methods focus on within-database detection [2,20], where forged images in the training set and testing set are manipulated by the same forgery technique. As shown in Fig. 1 (a), the styles of the synthesized images from various forgery techniques are quite different. Thus, an ongoing issue of face forgery detection is generalization under out-of-distribution (OOD) data [12]. As shown in Fig. 1 (b), the frequency distributions of real and fake images differ significantly in some datasets, but it is difficult to distinguish between the two in the RGB domain. Recent methods [17,29,23,26,39,5] introduced the face forgery frequency network to mine forgery traces in the frequency domain. Chen et al.[5] proposed a similarity model using frequency features to improve the model's performance in unseen domains, and Luo et al.[23] assumed that the high-frequency noise of images can remove colour texture and mine forgery traces and utilize image noises to boost the generalization ability. However, the role of the frequency domain is not always sufficiently effective, and RGB features also contain discriminative forgery information, as shown in Fig. 1 (c). The quality discrepancies between frequency and RGB features are less addressed [29].

To alleviate the effects of feature quality discrepancies and model uncertainty, we design a Dynamic Dual-spectrum Interaction Network (DDIN) that allows test-time training with uncertainty guidance. First, in order to explore the forgery region, frequency features are more effective than discriminating only in the RGB domain. We propose a Dynamic Fusion Module (DFM) to

use the quality distinction between RGB and frequency domain in an adaptive evaluation. Secondly, to increase model generalization on unseen data, we further fine-tune the trained network by estimating the uncertainty in the test-time training phase.

Spectrum transformation on an RGB image is used to obtain its corresponding frequency image based on *Discrete Cosine Transform* (DCT), and then these RGB and frequency images are input into the transformer-based network. Second, in the multi-level interaction stage, we use a Frequency-guided Attention Module (FAM) to direct the RGB modality from a frequency perspective, allowing us to attach more forgery traces. Thirdly, in the multi-modal fusion stage, we use the Cross-modal Attention Module (CAM) to fuse the features of the dual-stream network's output in order to enrich the information of the forged area. We further propose a Dynamic Fusion Module (DFM) for the dynamic enhancement of this multi-modal information to boost the generalization ability.

Moreover, to learn a more generalizable face forgery detector, we propose Uncertainty-guided Test-time Training (UTT). The key idea is to fine-tune the dynamic fusion module by estimating and exploiting the uncertainty of unseen test data. Specifically, we apply uncertainty-guided perturbations to different branches. The uncertain perturbation causes the network to predict quality weights in a probabilistic manner, and we fine-tune the network based on the distribution bias caused by this uncertainty. The distribution of predictions for forgery features on the training and test sets can be narrowed. Thus, it results in more robust predictions, particularly when the test set contains OOD data.

In brief, the main contributions are as follows:

- We propose a Dynamic Dual-spectrum Interaction Network (DDIN) that allows test-time training with uncertainty guidance to alleviate the effects of feature quality discrepancies and model uncertainty.
- We propose a Frequency-guided Attention Module (FAM) and Dynamic Fusion Module (DFM) in Dynamic Dual-spectrum Interaction Network (DDIN) that can be used to make the model dynamically fuse the features with quality discrepancies.
- We propose an Uncertainty-guided Test-time Training (UTT) by adding uncertain perturbation during the test-time training phase to improve the network generalization of forgery detection.
- Extensive experiments and visualizations demonstrate the effectiveness of our method against state-of-the-art competitors.

2 Related Work

2.1 Face Forgery Detection

In the past few years, face forgery detection has made significant strides, with a number of forgery-spotting models being successively proposed to meet the application's practical requirements. In the earlier stage, methods[2,18,1,7,28,13] are built with a significant emphasis on spotting semantic visual artefacts with sophisticated model designs. Recently, several works have focused on solving

the generalizing problem. For instance, methods[16,5] both notice the content-independent low-level features that can uniquely identify their sources and the identity swapping will destroy the origin consistency. Li et al.[16] suggest identifying those subtle features across the facial boundary and Chen et al.[5] turn to discover the spatial-local contradictions. Methods[25,10,3] fuse the low-level frequency pattern learning with CNN to improve the generalizability. Despite the fact that these techniques frequently work, the low-level artefacts they rely on are sensitive to post-processing techniques that differ across datasets, putting their generalizability at risk. Despite the possibility that these features will lead to some advancements, it is very likely that deepfake algorithms will be created in the future in order to create more realistic fakes and pose a bigger threat to social security. Different from existing works, we propose a novel dynamic dual-spectrum interaction network that allows test-time training with uncertainty guidance.

2.2 Test-time Training Strategy

The concept of test-time training was first presented in [33] for generalization to out-of-distribution test data. In this method, the main classification task is combined with a self-supervised rotation prediction task during training, and only the self-supervised task is used to help improve the visual representation during inference, which indirectly improves semantic classification. Li et al.[19] propose a reconstruction task within the main pose estimation framework, which can be trained by contrasting the reconstructed image with the ground truth gleaned from other frames. Chen et al.[4] proposed one-shot test-time training specially designed for the generalizable deepfake detection task. Nevertheless, despite some positive findings, current TTT methods aim to choose empirical self-supervised tasks, which carry a significant risk of degrading performance when the tasks are not properly chosen[22]. Instead, our UTT method is easy to implement and can avoid the tedious work of selecting an effective self-supervised task, which can significantly boost the deepfake detector's generalization performance and outperform existing solutions in a variety of benchmark datasets.

3 Proposed Method

3.1 Spectrum Transformation

As shown in the left-top in Fig. 2, we apply spectrum transformation that decomposes the input RGB image into frequency components, assisting the network in mining the distinction between real and forged regions.

Without loss of generality, let $X_{rgb} \in \mathbb{R}^{H \times W \times 3}$ denote the RGB input, where H and W are the height and width. First, we apply the Discrete Cosine Transform (DCT) to transform X_{rgb} from RGB into the frequency domain. DCT places low-frequency responses in the top-left corner and high-frequency responses in the bottom-right corner. Qian et al. [29] show that the low-frequency band is the first 1/16 of the spectrum, the middle-frequency band is between 1/16 and 1/8

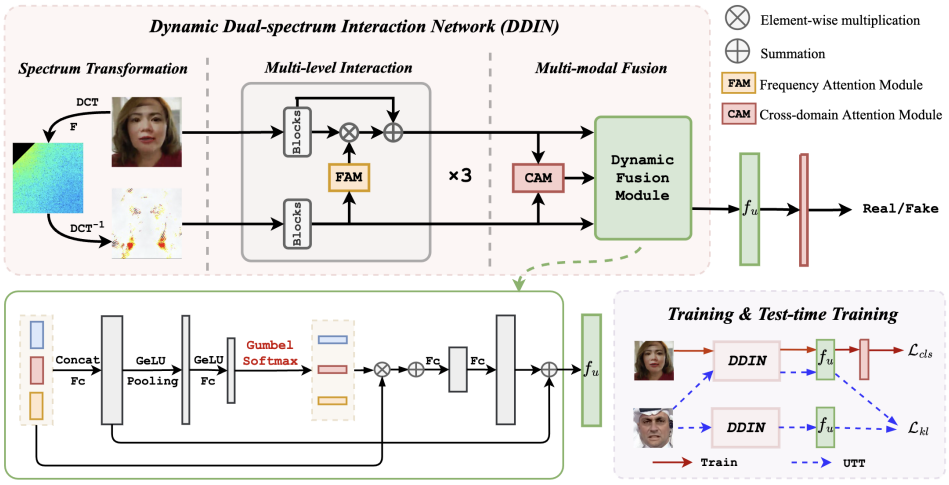

Fig. 2. The framework of our proposed DDIN and the pipeline of UTT.

of the spectrum, and the high-frequency band is the last 7/8 of the spectrum. To amplify subtle artefacts at high frequency, we filter out low and middle-frequency information by setting their frequency band to 0. To preserve shift-invariance and local consistency of natural images, we then invert the high-frequency spectrum back into RGB via IDCT to obtain the desired representation in the frequency domain, which can be formulated 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 IDCT. In this way, the original RGB input is decomposed and recombined frequency-aware data while maintaining the spatial relationship. Finally, we input both RGB and frequency images into the multi-level interaction phase to enhance the forged features.

3.2 Multi-level Interaction

RGB information is useful for locating anomalous textures in forged images, whereas frequency information amplifies subtly manipulated artefacts. To explore more forgery traces, we use a Frequency-guided Attention Module (FAM) based on CBAM[40]. While CBAM gains the attention weights from the RGB images, we exploit the frequency features to obtain the attention maps, to direct the RGB modality from a frequency perspective.

As shown in middle in DDIN at Fig. 2, let $X_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and $x_{freq} \in \mathbb{R}^{H\times W\times 3}$ denotes the RGB input and the frequency input. After feature extraction, we use FAM to derive the frequency attention map. That is:

$$\hat{f} = Conv_{3\times 3}(f_{freq}), \tag{2}$$

$$f_{att} = \sigma(Conv_{7\times7}(CAT(GAP(\hat{f}), GMP(\hat{f}))), \tag{3}$$

where f_{freq} denotes the frequency feature after feature extraction in Eq. (2), σ denotes the Sigmoid function, GAP and GMP represent global average pooling and global max pooling in Eq. (3), respectively. And CAT concatenates the features along with the depth. We finally choose a 7×7 convolution kernel to extract the forged traces in the frequency domain because it can detect edge information better and cover a larger area than three 3×3 convolution kernels. The attention map f_{att} contains subtle forgery traces in the frequency domain that are difficult to mine in the RGB features. Therefore, we implement f_{att} on the RGB feature f_{rgb}, directing f_{rgb} further mine forgery traces, that is:

$$f_{rgb} = f_{rgb} \oplus (f_{rgb} \otimes f_{att}), \tag{4}$$

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

In addition, there are three level stages to feature extraction, low-level, mid-level and high-level. The low-level features represent texture forgery information, while the high-level features extract more overall forgery traces. Therefore, we interact with RGB features and frequency features at multi-level obtaining a more comprehensive representation of forged features. Specifically, the frequency domain output f^i_{freq} of the i-th stage is used as the i+*1*-th stage input $\hat{f}^i_{freq}$, and the RGB input $\hat{f}^{i+1}_{rgb}$ is the RGB feature f^i_{rgb} previously guided in the frequency domain, which can be formulated as:

$$\hat{f}^{i+1}_{rgb} = f^i_{rgb} \oplus (f^i_{rgb} \otimes f^i_{att}), \quad \hat{f}^{i+1}_{freq} = f^i_{freq}. \tag{5}$$

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 thes output features.

3.3 Multi-modal Fusion

In recent years the attention mechanism has been broadly applied in natural language processing [37] and computer vision [9]. Inspired by these works, the resulting RGB features are combined with frequency features with a Cross-modal Attention Module (CAM). And considering the role of the frequency domain is not always sufficiently effective, which causes the quality discrepancies between the RGB and frequency feature, we designed a Dynamic Fusion Module (DFM) in the multi-modal fusion stage.

According to Sec. 3.2, the frequency modal should serve as a supporting component. Given RGB features f_{rgb} and frequency features f_{freq}, we implement CAM to perform a preliminary fusion of them into a unified representation by using the query-key-value mechanism. Specifically, we use 1×1 convolutions to embed f_{rgb} into Q and embed f_{freq} into K and V. Then we perform the attention mechanism by flatting them to 2D embeddings $\hat{Q}$, $\hat{K}$ and $\hat{V} \in \mathbb{R}^{\frac{h\times w}{16}\times c}$ along the channel dimension, which can be formulated as:

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

where f_{cam} denotes the preliminary fusion feature which aggregates the RGB and frequency information. Then, in order to effectively utilize the forged information in f_{rgb}, f_{cam} and f_{freq}, we design a dynamic fusion module in the multi-modal fusion stage. In more details, we input a set $S = \{f_{rgb}, f_{cam}, f_{freq}\}$ into DFM. DFM generates corresponding weights for each branch based on the quality information from all branches, which it then uses to weigh the combined information to combine from various branches.

To obtain the corresponding weights for each branch, we further integrate the three branch features by using two fully connected layers FC_1 and FC_2, global average pooling GAP and active layer GELU function δ, which can be formulated as:

$$f' = FC_1(CAT(f_{rgb}, f_{cam}, f_{freq})), \quad \hat{f} = FC_2(\delta_1(GAP(\delta_2(f')))), \tag{7}$$

where $f' \in \mathbb{R}^{h \times w \times 3c}$ and $\hat{f} \in \mathbb{R}^{1 \times 1 \times c}$. 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:

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

where $\alpha_i \in \mathbb{R}^{1 \times 1 \times C}$ represents the quality of each branch. Because different branches are contributed differently to mining forged clues, we weigh the fusion features based on the quality and use two linear mapping layers to restore the channel dimension of the dynamic fusion features. The output f_u can be formulated as:

$$f_u = f' + FC_4(FC_3(\sum_i^3 \alpha_i S_i)), i = 1, 2, 3. \tag{9}$$

3.4 Uncertainty-guided Test-time Training

The first three sections Sec. 3.1, Sec. 3.2 and Sec. 3.3 describe the DDIN network's mining of forgery traces' quality in the RGB and frequency domains and the dynamic fusion of forgery features to discriminate based on quality differences. However, the trained network's prediction weight is biased towards the quality discrepancies of different modalities of the training set data during dynamic fusion, leading to a bias in the fusion weight of uncertain unseen data, which affects the model's generalization. To improve the network's generalization of forgery detection, we further use uncertain perturbation as guidance during the test-time training phase.

To accomplish this, 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 $g = -\log(-\log(u))$[11]. We implement g to the DFM to influence the network's perception of intra-modal quality. The uncertain g modifies the quality weight

slightly, making it probabilistic rather than deterministic. And the uncertain quality weight β is given by the Gumbel softmax function[14]:

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

where τ is the softmax temperature. The β replaces the α in Eq. (9) and results in an uncertain distribution of the fused feature f_u. In contrast to f_u during training, the distribution of f_u in UTT is uncertain and related to the test set. Based on this uncertainty, we design a self-supervised task in the UTT stage. Specifically, we sample an image x from test-set $\mathbf{D}$ and input it twice to the pre-trained detector $f(\cdot, \theta)$, where the θ is the model parameter. Then we can gain two uncertain fused features f_u^1 and f_u^2. The distributions of these two features match the actual distribution of the model output, but they are influenced by the test set, and the perturbations make them uncertain. The KL loss is used to evaluate the distribution shift caused by uncertainty and to update the model parameters by narrowing the feature distribution gap, which encourages the model to perform well on the test set.

3.5 Loss Function

In the training phase, we flatten the f_u and pass it through the fully connected layer and sigmoid function to obtain the final predicted probability $\hat{y}$. And the classification loss is defined as:

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

where y is set to 1 if the image has been manipulated, otherwise it is set to 0.

In the test-time training phase, We add uncertain perturbations to the DFM 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

4.1 Experimental Setup

Datasets. We evaluate our proposed method and existing approaches on FaceForensics++ (FF++)[30], CelebDF[21] and DFDC[8]. FF++ is a face forgery detection dataset consisting of 1000 original videos with real faces, in 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 which are crafted by the improved DeepFake algorithm[36]. DFDC is a largescale dataset which contains 128,154 facial videos of 960 IDs.

Evaluation Metrics. To evaluate our method, we apply the Accuracy score (Acc) and Area Under the Receiver Operating Characteristic Curve (AUC) as

Table 1. Quantitative results on Celeb-DF dataset and FaceForensics++ dataset with different quality settings. The best results are shown in **blod**, and the second results are shown in blue.

Method	FF++(C23)		FF++(C40)		Celeb-DF	
	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)
MesoNet[1]	83.10	-	70.47	-	-	-
Multi-task [25]	85.65	85.43	81.30	75.59	-	-
Xception [6]	95.73	96.30	86.86	89.30	97.90	99.73
Face X-ray[17]	-	87.40	-	61.60	-	-
Two-branch[24]	96.43	98.70	86.34	86.59	-	-
RFM[38]	95.69	98.79	87.06	89.83	97.96	**99.94**
F3-Net [29]	97.52	98.10	90.43	93.30	95.95	98.93
Add-Net [42]	96.78	97.74	87.50	91.01	96.93	99.55
FDFL [16]	96.69	99.30	89.00	92.40	-	-
MultiAtt [41]	97.60	99.29	88.69	90.40	97.92	**99.94**
PEL [10]	97.63	99.32	90.52	94.28	-	-
DDIN	97.59	99.31	90.41	94.47	98.02	99.83
DDIN (+UTT)	**97.69**	**99.39**	**90.84**	**94.80**	**98.20**	99.93

our evaluation metrics. To ensure a fair comparison, the results of all comparison methods are taken from their paper.

Implementation Details. We modify MOA-Transformer[27] pre-trained on ImageNet as the backbone network. We use the DLIB[31] for face extraction and alignment. The input shape of images is resized to 224×224 with the data augmentation of randomly 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. We only update the parameters in the DFM and freeze the parameters of other layers during the test-time training phase.

4.2 Experimental Results

Intra-testing. In this section, we first compare our method with state-of-the-art face forgery detection methods on widely used datasets FF++ and Celeb-DF datasets. As shown in Table 1, for the FF++ dataset, our proposed method consistently outperforms all compared opponents by a considerable margin. For example, compared with the state-of-the-art method PEL[10], the AUC of our method exceeds it by 0.10% and 0.52% at all the two quality settings(c23 and c40), and this performance gain is also obtained under Acc. To explain, DDIN considers the quality of the auxiliary discriminant information contained in each branch in the multi-modal fusion stage and improves the network's discriminative ability on the test set in the UTT stage. The above results demonstrate the effectiveness of the proposed DDIN framework and UTT strategy. The above results demonstrate the effectiveness of our proposed method.

Table 2. Cross-testing in terms of AUC (%) by training on FF++. The best results are shown in **blod**.

Method		Xception[6]	RFM[38]	Add-Net[42]	F3-Net[29]	MultiAtt[41]	PEL[10]	DDIN	DDIN(+**UTT**)
Test	**CelebDF**	61.80	65.63	65.29	61.51	67.02	69.18	68.35	**69.32**
	DFDC	63.61	66.01	64.78	64.60	68.01	63.31	68.80	**70.10**

Table 3. Cross-manipulation evaluation on the subsets of FF++(C40) in terms of AUC(%). Grey background indicates intra-dataset results and Cross Avg. means the average of cross-method results. The best results are shown in **blod**.

Method	*Train*	**DF**	**F2F**	**FS**	**NT**	**Cross Avg.**
FDFL[16]		98.91	58.90	66.87	63.61	63.13
MultiAtt[41]		99.51	66.41	67.33	66.01	66.58
DDIN	**DF**	99.71	61.99	78.08	67.02	69.03
DDIN (+UTT)		**99.78**	66.62	**78.81**	**67.83**	**71.08**
FDFL[16]		67.55	93.06	55.35	66.66	63.19
MultiAtt[41]		73.04	97.96	**65.10**	71.88	70.01
DDIN	**F2F**	73.85	98.01	64.25	72.49	70.19
DDIN (+UTT)		**77.10**	**98.09**	64.42	**74.71**	**72.07**
FDFL[16]		75.90	54.64	98.37	49.72	60.09
MultiAtt[41]		82.33	61.65	98.82	54.79	66.26
DDIN	**FS**	88.20	62.13	98.80	56.63	68.98
DDIN (+UTT)		**89.18**	**62.56**	**98.83**	**58.44**	**70.06**
FDFL[16]		79.09	74.21	53.99	88.54	69.10
MultiAtt[41]		74.56	80.61	60.90	93.34	72.02
DDIN	**NT**	78.15	81.34	62.67	93.34	74.05
DDIN (+UTT)		**79.57**	**85.73**	**63.22**	**94.17**	**76.04**

Cross-testing. To evaluate the generalization ability of our method on unknown forgeries, we conduct cross-dataset experiments by training and testing on different datasets. Specifically, we train the models on FF++ and then test them on Celeb-DF and DFDC, respectively. As shown in Table 2, we observe that our method outperforms the other methods well on the unseen dataset. For example, when testing on the DFDC dataset, the AUC score of most previous methods drops to around 70%. The performance mainly benefits from the proposed DDIN framework and UTT fine-turning which focus on quality differences between different modalities, while the uncertainty perturbation guides the model to learn more distribution discrepancies between the train-set and test-set. Instead of overfitting with specific forged patterns as in existing methods, our method fine-tunes the trained model by using the unlabeled test set data to achieve better generalizability.

We further conduct fine-grained cross-testing by training on a specific manipulation technique and testing on the others listed in FF++(C40). As shown in Table 3, we compare our method with approaches that focus on specific forgery patterns like FDFL[16] and MultiAtt[41]. Our method generally outperforms others on unseen forgery types. In comparison, our pre-trained detector can adapt to the test samples via UTT, thus being more effective than the two methods for the generalizable deepfake detection task.

4.3 Ablation Study

Components. As shown in Table 4, 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 frequency-guided attention module used in the multi-level interaction stage can enhance the performance, and it can be improved by adding the proposed frequency-guided attention module or dynamic fusion module, reaching better performance when they are applied to the overall DDIN framework. The results verify that the frequency input is distinct and complementary to the RGB information and the quality discrepancies are negligible.

Table 4. Ablation study of the proposed method on FF++.

Ablation Study	***Modules***			**FF++(C40)**	
	FAM	**DFM**	**UTT**	**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**

Furthermore, uncertainty-guided test-time training is also extremely effective to boost performance. And the best performance is achieved when combining all the proposed components with Acc and AUC of 90.84% and 94.80%, respectively. In addition, before the dynamic fusion module, we investigate the preliminary feature fusion method. Table 5 shows the results of three different scenarios. By comparing a and b, we can observe that the initial feature fusion is required. The result of c shows that the CAM module we use can effectively supplement the RGB and frequency domain feature information.

4.4 Visualization

To gain a better understanding of our method's decision-making mechanism, we provide the Grad-CAM[32] visualization on FF++ as shown in Fig. 3. It

Table 5. Ablation study of fusion method before dynamic fusion.

Ablation Study	*Fusion Method*		FF++(C40)	
	Concat	CrossAtt	AUC(%)	Acc(%)
(a)	-	-	93.47	89.79
(b)	✓	-	94.26	90.33
(c)	-	✓	**94.80**	**90.84**

can be observed that the baseline method MOA-Transformer[27] tends to overlook forged traces in fake faces, particularly those that are concealed 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[36] and NeuralTextures[34] technologies, our method focuses on the edge contours of artefacts, which are difficult to detect in the RGB domain. And our method is more sensitive to the abnormal texture information forged by FaceSwap[15] in the eyes region and the inconsistent information forged by Face2Face[35] in the mouth region.

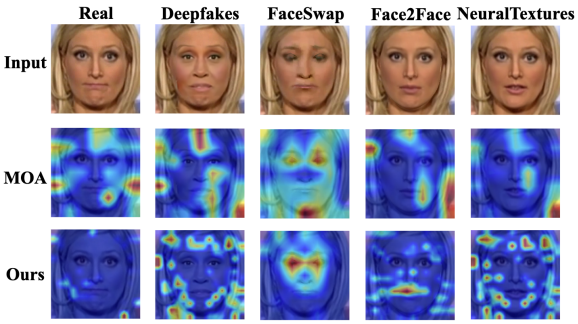

Fig. 3. The Grad-CAM [32] of visualization.

5 Conclusion

In this paper, we have proposed a Dynamic Dual-spectrum Interaction Network (DDIN) that allows test-time training to alleviate the effects of feature quality discrepancies and model uncertainty. The frequency-guided attention module used in multi-level interaction and the dynamic fusion module applied in multi-modal fusion can make the network dynamically fuse the features with quality discrepancies. Meanwhile, the uncertainty-guided test time training is introduced to fine-tune the trained detector by adding uncertain perturbation, which improves the network generalization. Extensive experiments and visualizations demonstrate the effectiveness of our method against its state-of-the-art competitors. In the future, we will explore the use of uncertainty to design self-supervised tasks in other related fields such as forensic attribution.

6 Acknowledgment

This research is partly supported by National Natural Science Foundation of China (Grant No. 62006228), Youth Innovation Promotion Association CAS (Grant No. 2022132) and the University Synergy Innovation Program of Anhui Province (No. GXXT-2022-036). The authors would like to thank Tong Zheng (AHU) and Jin Liu (ShanghaiTech) for their valuable discussions.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: WIFS (2018)
2. Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: CVPRW (2019)
3. Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: End-to-end reconstruction-classification learning for face forgery detection. In: CVPR (2022)
4. Chen, L., Zhang, Y., Song, Y., Wang, J., Liu, L.: Ost: Improving generalization of deepfake detection via one-shot test-time training. In: NeurIPS (2022)
5. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: AAAI (2021)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR (2017)
7. Das, S., Seferbekov, S., Datta, A., Islam, M., Amin, M., et al.: Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. In: ICCV (2021)
8. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R.: Exploiting fine-grained face forgery clues via progressive enhancement learning. In: AAAI (2022)
11. Gumbel, E.J.: Statistical theory of extreme values and some practical applications: a series of lectures, vol. 33. US Government Printing Office (1954)
12. Guo, H., Wang, H., Ji, Q.: Uncertainty-guided probabilistic transformer for complex action recognition. In: CVPR (2022)
13. He, R., Zhang, M., Wang, L., Ji, Y., Yin, Q.: Cross-modal subspace learning via pairwise constraints. IEEE TIP **24**(12), 5543–5556 (2015)
14. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
15. Kowalski, M.: Faceswap. https://github.com/marekkowalski/faceswap
16. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: CVPR (2021)
17. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR (2020)
18. Li, X., Hou, Y., Wang, P., Gao, Z., Xu, M., Li, W.: Trear: Transformer-based rgb-d egocentric action recognition. TCDS **14**(1) (2021)

19. Li, Y., Hao, M., Di, Z., Gundavarapu, N.B., Wang, X.: Test-time personalization with a transformer for human pose estimation. In: NeurIPS (2021)
20. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In: WIFS (2018)
21. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: CVPR (2020)
22. Liu, Y., Kothari, P., van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? In: NeurIPS (2021)
23. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: CVPR (2021)
24. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: ECCV (2020)
25. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: BTAS (2019)
26. Nirkin, Y., Wolf, L., Keller, Y., Hassner, T.: Deepfake detection based on discrepancies between faces and their context. IEEE TPAMI (2021)
27. Patel, K., Bur, A.M., Li, F., Wang, G.: Aggregating global features into local vision transformer. arXiv preprint arXiv:2201.12903 (2022)
28. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: ICPR (2021)
29. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV (2020)
30. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV (2019)
31. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. IVC **47** (2016)
32. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
33. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: ICML (2020)
34. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM TOG **38**(4) (2019)
35. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
36. Tora: Deepfakes. https://github.com/deepfakes/faceswap/tree/v2.0.0
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
38. Wang, C., Deng, W.: Representative forgery mining for fake face detection. In: CVPR (2021)
39. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.G., Li, S.N.: M2tr: Multi-modal multi-scale transformers for deepfake detection. In: ICMR (2022)
40. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV (2018)
41. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: CVPR (2021)
42. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: ACM MM (2020)