OPEN SET FACE FORGERY DETECTION VIA DUAL-LEVEL EVIDENCE COLLECTION (SUPPLEMENTARY MATERIALS)

Anonymous authors
Paper under double-blind review

A DATASET SETUP DETAILS

Table 1: Forgery configuration of each fake category in our OSFFD setting. This table details the specific deepfake forgeries included in each fake category. Certain forgeries are used in both the training and testing phases, while others are reserved exclusively for testing.

	FS	FR	EFS	FE	SM
Train & Test	FSGAN, FaceSwap, SimSwap, InSwapper, BlendFace	FOMM, FS_vid2vid, Wav2Lip, MRAA, OneShot	VQGAN, HStyleGAN-XL, SD-2.1, StyleGAN2, StyleGAN3	\	\
Test Only	UniFace, MobileSwap, e4s, FaceDancer, DeepFaceLab	LIA, TPSMM, HyperReenact, DaGAN, SadTalker, MCNet, PIRender, HeyGen	PixArt-alpha, MinJourney6, DDPM, WhichisReal, RDDM, DiT-XL/2, SiT-XL/2	StarGAN, StarGANv2, e4e, CollabDiff, StyleCLIP	DeepFakes-StarGan-Stack, StarGan-BlendFace-Stack

Here, we provide the data configuration details about our setup in our Open Set Face Forgery Detection (OSFFD) experiments. We utilize the DF40 (Yan et al., 2024) dataset, which comprises four fake categories, and further incorporate two Stacked Manipulation forgeries from the ForgeryNet benchmark (He et al., 2021), resulting in five distinct fake categories in total. Specifically, the five fake categories are: Face Swapping (FS), Face Reenactment (FR), Entire Face Synthesis (EFS), Face Editing (FE), and Stacked Manipulation (SM). Due to the limited amount of fake data available for FE and SM, these two categories are treated as novel fake categories. The remaining categories FS, FR, and EFS are considered as either seen or unseen categories across different experimental configurations. As outlined in the main paper, we initially adopted a leave-one-out strategy, in which one of the three categories was withheld during training and regarded as an unseen category during testing. In a subsequent setting, all three categories (FS, FR, and EFS) were included as known classes, while FE and SM were used as novel fake categories during evaluation. During the training phase, the data consists of faces form real class and known fake categories, whereas the testing phase additionally includes samples from novel fake categories.

Since the number of forgeries varies across the fake categories FS, FR, and EFS, we select five forgeries from each category for use in both training and testing, while reserving the remaining forgeries for testing only. We show the forgery configuration in Table 1. For detailed information on the methods included in FS, FR, EFS, and FE, along with the corresponding data quantities, sources, and sub-types, please refer to Table 2 in the DF40 paper (Yan et al., 2024). The methods categorized under SM can be found in Figure 3 of the ForgeryNet benchmark paper (He et al., 2021). We followed the original training and testing splits for each specific deepfake forgery and performed dataset balancing by randomly selecting the largest possible subset such that each fake category contained an equal number of samples, and the overall number of real and fake face images was balanced.

B IMPLEMENTATION DETAILS

Network Modules. Our Dual-Level Evidential face forgery Detection (DLED) method employs CLIP (Radford et al., 2021) as the backbone for both the spatial and frequency branches, using the ViT-B/16 (Dosovitskiy, 2020) encoder with half-precision (fp16) data type. The same CLIP configuration is also used for all CLIP-based baseline methods to ensure fair comparisons. For the CLIP Zero-Shot baseline, we used the text prompt “a [CLASS] photo.” as the input, where “[CLASS]” is replaced with the predefined class names corresponding to the seen classes. For CNN-based baseline methods and one-class detectors in Two-stage baselines, all backbones utilize the official implementations provided by the respective authors. These baseline models are trained from scratch for the OSFFD experiments. Specifically, SPSL (Liu et al., 2021) and UCF (Yan et al., 2023) utilized the Xception backbone (Rossler et al., 2019), while SBI (Shiohara & Yamasaki, 2022) and SIA (Sun et al., 2022) are based on EfficientNet-B4 (Tan & Le, 2019). NPR (Tan et al., 2024) is a modified version of ResNet (He et al., 2016), and OC-FakeDetect (Khalid & Woo, 2020) incorporates a Variational Autoencoder. (Kingma & Welling, 2013). Xception (Rossler et al., 2019) is used directly without modification.

Concrete Implementation. OSFFD requires the model to simultaneously detect novel fake categories and perform multiclass classification. Accordingly, we use multiclass classification accuracy (Acc) to evaluate performance on classification, and Detection Rate (DR), defined as the recall of the unseen fake categories, to assess the model's ability to discover novel fake. In our experiments, the batch size was set to 32 for training and 100 for testing across all models. Our DLED model employed the softplus function as its confidence function $h(\cdot)$. Both DLED and CLIP-based baseline models are optimized using SGD with a learning rate of 0.002, while the remaining baseline methods use the Adam optimizer with a learning rate of $1e^{-4}$. All models, except for the CLIP Zero-Shot baseline, are trained for 50 epochs. We randomly selected a seed and evaluated the model's performance at the final epoch in a single run. All methods are implemented in PyTorch, and experiments are conducted on an RTX 6000 Ada GPU. Our code is available at https://anonymous.4open.science/r/NovelDFD-BF78.

AI Disclosure. We used generative AI to edit portions of this paper for clarity, grammar, and word choice.

C DLED Algorithm

In this section, we provide a detailed description of the inference procedure for our DLED model. We denote the spatial and frequency branches as $\mathcal{F}^s$ and $\mathcal{F}^f$, respectively, such that $\mathcal{F} = \{\mathcal{F}^s, \mathcal{F}^f\}$. Given the retrieved uncertainty thresholds $\{\tau_k\}_{k=1}^K$ for each known class k, the DLED inference procedure for an input face image x is described in Algorithm 1, where $h(\cdot)$ denotes the evidence function and $\mathrm{FFT}(\cdot)$ represents the Fast Fourier Transform.

Algorithm 1 DLED Inference Procedure

```
Require: Input image $x$; uncertainty thresholds $\{\tau_k\}_{k=1}^K$ for each known class.
Obtain the frequency map: $x^f = \mathrm{FFT}(x)$
Extract dual-level evidence: $e^s = h(\mathcal{F}^s(x))$, $e^f = h(\mathcal{F}^f(x^f))$
Calculate belief and uncertainty for each branch:
    $b_k = e_k/S, \quad u = K/S, \quad S = \sum_{k=1}^K (e_k + 1)$
Fuse belief and uncertainty:
    $\widetilde{b}_k = \gamma(b_k^s b_k^f + b_k^s u^f + b_k^f u^s),$
    $\widetilde{u} = \gamma u^s u^f, \quad \gamma = 1/\left(1 - \sum_{i \neq j} b_i^s b_j^f\right)$        [Eq. 5]
Calculate fused evidence and parameters:
    $\widetilde{S} = K/\widetilde{u}, \quad \widetilde{e}_k = \widetilde{b}_k \times \widetilde{S}, \quad \widetilde{\alpha}_k = \widetilde{e}_k + 1$        [Eq. 6]
Calculate predictive distribution and predicted label:
    $\widetilde{p}_k = \widetilde{\alpha}_k/\widetilde{S}, \quad \widetilde{y} = \arg\max_k \widetilde{p}_k$
Improve uncertainty estimation:
    $\hat{u} = 1/\left(\max\{\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_K\}\right)$        [Eq. 8]
Novel category detection and final prediction:
    $\hat{y} = \begin{cases} \widetilde{y}, & \text{if } \hat{u} \leq \tau_{\widetilde{y}} \\ K+1, & \text{if } \hat{u} > \tau_{\widetilde{y}} \end{cases}$
return $\hat{y}$ and $\hat{u}$
```

D Influence of Uncertainty Threshold

In this section, we present additional experimental results analyzing the impact of the uncertainty threshold on the OSFFD problem. Uncertainty threshold is calculated from the training data such

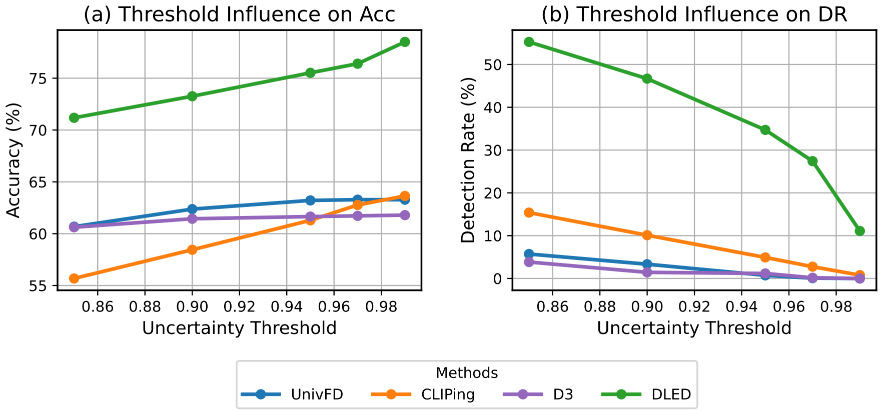

Figure 1: Accuracy and Detection Rate across different uncertainty thresholds for UnivFD, CLIPing, D^3, and DLED.

that a certain ratio of the samples in each class are marked as known. Our experiments on the OSFFD problem followed the widely adopted open set recognition evaluation protocol (Wang et al., 2022; Zhao et al., 2023), which sets the uncertainty threshold at 95%. A higher threshold indicates that more training samples are regarded as reliable, thereby imposing a stricter criterion for novel fake discovery. To further investigate the impact of the uncertainty threshold on the OSFFD problem, we conducted a series of experiments using various threshold values [0.85, 0.90, 0.95, 0.97, 0.99] on DLED, as well as three CLIP-based baseline methods: UnivFD (Ojha et al., 2023), CLIPing (Khan & Dang-Nguyen, 2024), and D^3 (Yang et al., 2025).

In this experiment, FS and FR are treated as seen fake categories, while EFS is designated as the only novel fake category. The results are presented in Figure 1. It can be observed that as the uncertainty threshold increases, classification accuracy improves, whereas the novel fake discovery rate declines. This indicates that a stricter threshold reduces the detection of novel samples but also decreases the misclassification of known samples. These findings highlight the critical role of the uncertainty threshold in balancing known class classification and novel fake discovery. Moreover, across all threshold settings, our DLED method consistently achieves higher Acc and DR compared to the three CLIP-based baselines, demonstrating its robustness and superior performance.

REFERENCES

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4360–4369, 2021.

Hasam Khalid and Simon S Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 656–657, 2020.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 772–781, 2021.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, 2022.

Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2022.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22412–22423, 2023.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23850–23859, 2025.

Chen Zhao, Dawei Du, Anthony Hoogs, and Christopher Funk. Open set action recognition via multi-label evidential learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22982–22991, June 2023.