PUDD: Towards Robust Multi-modal Prototype-based Deepfake Detection

Alvaro Lopez Pellcier*, Yi Li*, Plamen Angelov
School of Computing and Communications, Lancaster University

Abstract

Deepfake techniques generate highly realistic data, making it challenging for humans to discern between actual and artificially generated images. Recent advancements in deep learning-based deepfake detection methods, particularly with diffusion models, have shown remarkable progress. However, there is a growing demand for real-world applications to detect unseen individuals, deepfake techniques, and scenarios. To address this limitation, we propose a Prototype-based Unified Framework for Deepfake Detection (PUDD). PUDD offers a detection system based on similarity, comparing input data against known prototypes for video classification and identifying potential deepfakes or previously unseen classes by analyzing drops in similarity. Our extensive experiments reveal three key findings: (1) PUDD achieves an accuracy of 95.1% on Celeb-DF, outperforming state-of-the-art deepfake detection methods; (2) PUDD leverages image classification as the upstream task during training, demonstrating promising performance in both image classification and deepfake detection tasks during inference; (3) PUDD requires only 2.7 seconds for retraining on new data and emits 10^5 times less carbon compared to the state-of-the-art model, making it significantly more environmentally friendly.

1. Introduction

Deepfakes, created through digitally manipulated techniques, convincingly replace one person's likeness with another's [1, 16]. A recent study highlights the challenge in distinguishing real from AI-generated images, with only 61% of participants accurately identifying them—falling short of the expected 85% [18]. This difficulty stems from advancements in deep learning models, notably autoencoders [26, 37], Generative Adversarial Networks (GANs) [11], diffusion models [32] and neural style transfer (NST) [13]. Manipulating audio, video, and image data poses serious consequences, including security vulnerabilities, safety concerns, ethical dilemmas, and erosion of public trust [28]. One famous example of a deepfake involved a video purportedly showing the Ukrainian president urging soldiers to surrender to Russia. The video circulated on social media and appeared on a Ukrainian news website before being debunked and removed [7]. Consequently, deepfake detection has become a critical area of research, drawing increasing attention from researchers.

In recent years, significant advancements in deep learning techniques have greatly improved their effectiveness in detecting deepfakes, resulting in notable performance gains [30]. Deepfake detection methods can be broadly categorized into two aspects: artifact-specific [19] and undirected approaches [21], depending on the data and deepfake techniques involved. For instance, artifact-specific approaches focus on detecting unnatural areas in deepfake human faces by leveraging edges and optical flow. Chintha et al. employed a combination dataset comprising visual frames, edge maps, and dense optical flow maps as inputs to a recurrent XceptionNet [8]. By learning a fused representation of these features, the model achieves accurate predictions. On the other hand, undirected approaches eschew specific artifacts or predefined feature sets, instead training a general-purpose classifier to autonomously analyze the entire input data and learn relevant features. However, these undirected deepfake detection methods suffer from three main drawbacks.

The majority of recent deepfake detection techniques [30, 36, 39] struggle with robustness, which refers to the ability of the detector to maintain high accuracy when processing unseen deepfakes—those generated using techniques and models different from those used in training. Robustness is essential for the practical application of these systems in real-world scenarios. Secondly, training these deepfake detection models is time-consuming due to the large scale of the network models. For example, Zhao et al. utilize two parallel Vision Transformer-Large (ViT-L) networks with several Xception blocks [9, 10] to extract spatial and temporal features from deepfake videos [39], resulting in over 600 million parameters. Retraining such models for new individuals is therefore exceedingly time-consuming. Thirdly, many of these detection techniques

*These authors contributed equally to this work
Emails: {a.lopezpellicer, y.li154, p.angelov}@lancaster.ac.uk. (*Corresponding author: Yi Li.*)

lack interpretability due to their complex network architecture and black box nature. They often make detection decisions based on high-dimensional feature maps, limiting their explainability.

To overcome these drawbacks, our contributions are summarized as follows:

• As the core idea of our contribution, we propose a Prototype-based Unified Framework for Deepfake Detection (PUDD) framework. Prototypes are clustered to learn representations for the upstream task, i.e., video classification. This robust representation allows deepfakes generated by unseen deepfake techniques to be returned unedited, maintaining their visual integrity and preserving their latent space representation.

• We propose integrating state-of-the-art techniques from sim-DNN [34] and xClass [3] for deepfake detection. Our approach includes a prototype learning layer that is easily trained and significantly enhances detection accuracy without necessitating the retraining of the entire framework. Additionally, it significantly reduces CO2 emissions, computational and power requirements compared to other large detection and classification models making our approach significantly more environmentally friendly.

• We provide interpretability to understand the prototype-based classification as the degree to which a human can consistently predict the model's output.

• We demonstrate the efficiency and effectiveness of our proposed methods by comparing them to state-of-the-art deepfake detection models across multi-modal data, i.e., deepfake images and videos.

2. Related Works

2.1. Deepfake Generation

Deepfake generation involves the use of deep learning techniques to create convincing image, audio and video hoaxes. There are several methods for creating deepfakes, but the most widely used methods are Variational Autoencoders (VAEs) [12, 35], Generative Adversarial Networks (GANs) [20, 23, 33], and diffusion models [15]. To generate a deepfake with VAEs, FaceSwap encodes both the source and target faces into the latent space using the trained encoder [12]. Then, it swaps the latent representations of the faces, effectively transferring the facial features of the target face onto the source face. Moreover, as a common used deepfake technique, style-based GAN (StyleGAN) [20] facilitates an automatically learned, unsupervised separation of high-level attributes, e.g., pose and identity when trained on human faces, and stochastic variation in the generated images, e.g., freckles and hair. It also allows for intuitive, scale-specific control of the synthesis.

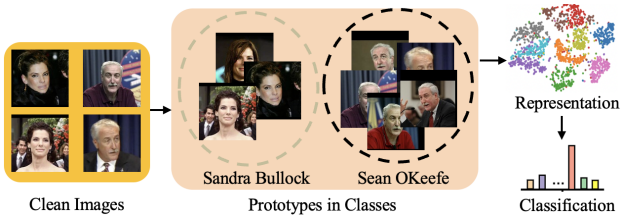

Figure 1. Prototype learning-based image classification with original images.

2.2. Deepfake Detection

The recent literature [4, 17, 24, 25] confirms the critical need for detecting deepfakes to protect the reputations and credibility of public figures, particularly politicians, who are vulnerable to manipulation and misinformation campaigns. Deepfakes have the potential to propagate false narratives and undermine trust in democratic processes. Robust detection methods are therefore essential to prevent the dissemination of deceptive content. By investing in deepfake detection technologies, we can mitigate the risks posed by malicious actors intent on exploiting digital media for political gain, thus safeguarding the integrity of public discourse

Reiss et al. introduce a state-of-the-art deepfake detection technique based on the concept of 'fact checking', adapted from fake news detection [31]. This approach verifies that claimed facts (e.g., identity as Biden) align with observed media (e.g., is the face truly Biden's?), allowing differentiation between real and fake media. Similar with our upstream video classification task, Haliassos et al. propose self-supervised representation learning across visual and auditory modalities to capture factors such as facial movements, expression, and identity [14]. These learned representations serve as targets predicted by the detector alongside the traditional binary forgery classification task.

2.3. Prototype Learning

As depicted in Figure 1, prototype-based deepfake detection methods [2, 34] calculate the local peaks of the density for each individual, essentially identifying the most representative data samples in each class from the training set as prototypes.

These methods then evaluate the similarity between new data samples and autonomously selected prototypes to classify images as either deepfake or original data samples. Bouter et al. simplify the complexity of working with spatio-temporal prototypes and enable their replacement to achieve greater interpretability [6]. Aghasanli et al. calculate similarity scores (Euclidean distance in feature space) between an input image and all identified prototypes to derive rules for each specific sample [2]. However, a common limitation of prototype-based deepfake detection methods is the time-consuming nature of retraining detectors for new

classes or individuals.

3. Proposed Method

Our proposed solution is built on a series of novel contributions that collectively form the deepfake detection framework. As illustrated in Figure 2, these innovations include the introduction of the Prototype Learning layer (3.3) to cluster prototypes from input data and calculate their similarity to established prototypes. Additionally, the Classification layer (3.4) is developed to classify images based on the estimated similarity scores obtained from the Prototype Learning layer.

3.1. Pre-processing

In our experiments, we evaluate our framework on two public datasets [5, 27]. Firstly, for the Celeb-DF dataset, we sample one frame every two seconds from the videos. These frames are then cropped to extract smaller patches containing only the face regions. In the training stage, we utilize all 59 celebrities available in the Celeb-DF dataset [27]. However, different from conventional methods that rely on paired data, we only consider frames from original videos for training and frames from deepfake videos for inference. Secondly, for the CIFAKE dataset, no preprocessing of clean images is required prior to training.

3.2. Feature Extraction

The proposed PUDD extracts features from a pre-trained model for prototype learning and the upstream task, i.e., video classification, eliminating the need for fine-tuning. To achieve this, we choose DINOV2 [29] as the feature extractor due to its ability to effectively correct non-uniformities in images and its promising performance in image classification tasks. After feature extraction, the prototypes are calculated and learned from these features.

3.3. Prototype Learning Layer

Given a training set $X = \{x_1, x_2, ..., x_n\}$ of n image samples with C classes, we aim to learn the prototypes $P = \{p^{1_1}, p^{1_2}, ..., p^{c_m}\}$ of original videos for video classification. For example, p^{c_m} refers to the m-th prototype in the in the c-th class. Particularly, the most representative data samples in each class of the dataset are selected as prototypes. We show the prototype clustering result in Figure 3.

In Figure 3, it is depicted that the prototypes from original videos (ID13, Id23, and Id24) are clustered, while two outliers, i.e., deepfake videos are kept away from these clusters. Specifically, even though the hair style, presence of a mustache, and apparent age vary across data samples within ID13, the prototypes are clustered effectively to enable successful classification of the celebrity. This indicates that the proposed PUDD framework can capture and leverage subtle yet discriminative features to distinguish between different individuals, even amidst significant variations in appearance. These prototypes facilitate a reasoning process based on the similarity (proximity in feature space) between a data sample and a prototype. In this case, prototypes are identified as the local density peaks [34], essentially the most representative samples from the training set. Therefore, only a limited number of samples from the training dataset are chosen as prototypes, ensuring the system's efficiency and compatibility with a broad range of devices. We define N_c and M_c are the numbers of samples and prototypes in the c-th class, respectively.

As the core idea of our contribution, the Prototype Learning layer serves to cluster prototypes and calculate the similarity and for each image associated prototype. Inspired from a specific Cauchy equation in [34], we define a similarity score between n-th data x_n and prototypes in c-th class by using Euclidean distance to identify how closely new data aligns with known data patterns drawn from the extracted features:

$$S_{x_n}(p^c) = \frac{\sum_{m=1}^{M_c} \|x_n - p^{c_m}\|^2}{1 + \frac{\|x-\mu\|^2}{\|\sigma\|^2}} \tag{1}$$

where μ and σ are the mean and variance of data samples, respectively. This step evaluates the proximity of data samples within the feature space, utilizing Euclidean distance as metrics. After calculating all the similarity scores to prototypes in all classes, we compare the minimum similarity score against the mean and variance of data by using m-σ rule.

3.4. Classification Layer

The proposed classification layer makes the decision on whether an input belongs to an existing class or a deepfake. The m-σ rule is applied to detect potential attacks, which can be depicted through an inequality condition:

$$\begin{aligned} \text{IF} \quad & min(S_{x_n}) > (\bar{\mu} - m\sigma) \\ \text{THEN} \quad & x_n \in \text{Potential deepfake video or image} \\ \text{ELSE} \quad & x_n \in \text{Classification label} \end{aligned} \tag{2}$$

where $\bar{\mu}$ refers to the recursive mean of data samples. If this condition is met, it suggests that the system has recognized a divergence or a new data concept, distinct from the established data patterns used to generate the prototypes. If not, it indicates no significant change in the data concept, allowing the algorithm to continue with its standard classification process. This mechanism enables PUDD to adaptively respond to new data and effectively identify potential deepfake detection.

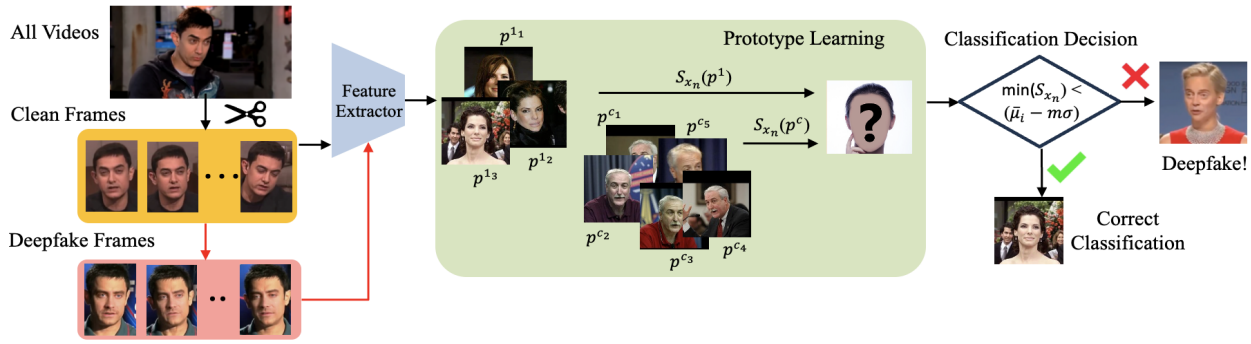

Figure 2. The proposed prototype learning-based framework. We extract frames from raw videos and crop them into small patches. The *red* lines only refer to the inference stage.

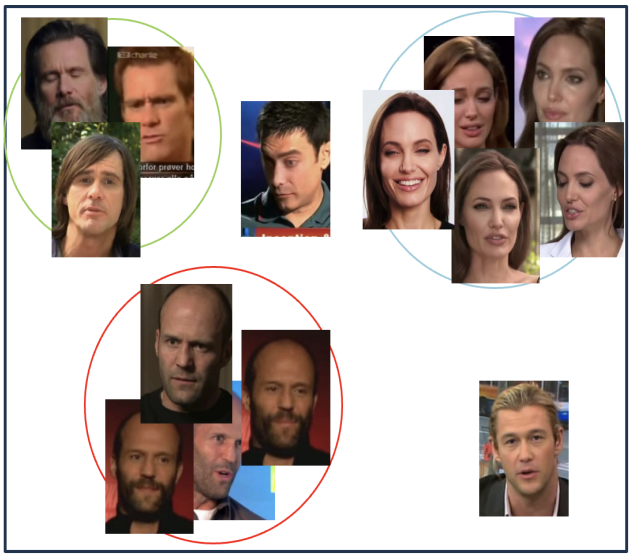

Figure 3. Prototype clustering visualizations.

4. Experiments

4.1. Data and Deepfakes

4.1.1 Celeb-DF

In the Celeb-DF dataset [27], there are 590 real videos featuring 59 celebrities of diverse genders, ages, and ethnic groups, collected from publicly available sources such as YouTube. Additionally, the dataset includes 5,639 deepfake videos generated using improved synthesis methods, including color mismatch, inaccurate face masks, and temporal flickering. Consequently, the overall visual quality of the synthesized deepfake videos in Celeb-DF is significantly enhanced compared to existing datasets, with notably fewer visual artifacts.

In this work, we use all 590 real videos and 5,639 deepfake videos for the training and inference stages, respectively. We extract one frame per every two seconds of the videos and then crop these frames into smaller patches containing only the face regions. Specifically, the training stage comprises 11,723 cropped frames from the original videos, while the inference set consists of 60,847 cropped frames from the deepfake videos.

4.1.2 CIFAR-10

Different from Celeb-DF, CIFAKE [5] is designed to encompass non-human classes such as birds, cars, and ships. The dataset comprises 60,000 synthetically-generated images and an equal number of real images collected from CIFAR-10 [22]. The synthetic images are generated using a fine-tuned Stable Diffusion Model [32].

In our study, we train the model using prototypes learned from 50,000 original images in the training set. Subsequently, we evaluate the proposed method using 50,000 deepfake images from the inference set.

4.2. Competitors and Implementation

The proposed method is evaluated and compared to state-of-the-art competitor models. We reproduce three state-of-the-art deepfake detection techniques [30, 36, 40], utilizing the best-reported implementations available in the literature. For example, Aghasanli et al. achieve superior results by fine-tuning only the multilayer perceptron (MLP) head in the original Vision Transformer (ViT) [2]. Therefore, we fine-tune this model with our dataset to serve as a competitor in our comparison experiments. Secondly, we reproduce four prototype learning methods, including both for deepfake detection [2, 6] and adversarial attack detection [34, 38], i.e., similar to deepfakes.

In this paper, the proposed prototype learning is implemented on the detector and further studies on feature extractor are out of scope of this paper. In the comparison experiments (5.1 & 5.2), we exploit DINOV2 [29] as the feature extractor. All the experiments are run on Tesla V100 GPUs.

Table 1. Deepfake video detection comparison on the Celeb-DF dataset. Para. and Acc are trainable parameters and detection accuracy, respectively.

	Algorithm & Network			Computational Cost		Acc (%)
Method	Pre-training	Prototype	Backbone	Para.	Training Time (s)	
MPC-CA [38]	✓	✓	BERT + MLP	6.4 M	399.5	79.5
Sim-DNN [34]	✓	✓	VGG16 + DNN	**2.8 M**	109.8	88.4
IPD [2]	✓	✓	ViT-L-32 + MLP Head	3.7 M	254.7	89.2
ProtoExplorer [6]	✗	✓	DPNet	3.8 M	258.0	92.5
NoiseDF [36]	✓	✗	RIDNet + Attention	9.9 M	278.6	70.1
FTCN [40]	✗	✗	FTCN	26.6 M	7482.6	86.9
MMtrace [30]	✗	✗	MLP	4.7 M	209.5	92.9
PUDD	✓	✓	DINOV2 + xDNN	7.6 M	**2.7**	**95.1**

5. Results

5.1. Celeb-DF

The proposed PUDD is evaluated on deepfake detection task over the Celeb-DF dataset [27]. Table 1 shows the results, each of them is the average of 60,847 deepfake frames.

From Table 1, it can be observed that: (1) In all the evaluated models, the proposed PUDD achieves *95.1*% for deepfake video detection, which offers the best effectiveness. (2) PUDD demonstrates state-of-the-art efficiency in the training stage compared to state-of-the-art models due to its rapid calculation of simple prototypes. This feature enables swift retraining for unseen celebrities, making it highly practical for real-world applications.

5.2. CIFAKE

We compare the deepfake image detection performance over the CIFAKE dataset [5]. The results are presented in Table 2, each result is average of 50,000 deepfake images.

Table 2. Deepfake image detection comparison on the CIFAKE dataset.

Method	Training Time (s)	Acc (%)
MPC-CA [38]	483.2	79.5
Sim-DNN [34]	142.3	88.4
IPD [2]	249.9	89.2
ProtoExplorer [6]	261.7	92.5
NoiseDF [36]	300.6	70.1
FTCN [40]	8358.3	86.9
MMtrace [30]	252.7	92.9
PUDD	**2.8**	**94.6**

From Table 2, the proposed PUDD outperforms the state-of-the-art models [2, 6, 30, 34, 36, 38, 40] on both accuracy and training time. there are three main differences between the Celeb-DF and CIFAKE datasets. Firstly, Celeb-DF comprises video data, whereas CIFAKE consists of image data. Secondly, the deepfake generation techniques used in these datasets differ, with Celeb-DF employing improved generation techniques and CIFAKE utilizing the Stable Diffusion Model. Thirdly, while Celeb-DF only includes human classes, CIFAKE encompasses 10 non-human classes such as cars, birds, ships, and cats. Therefore, the robust deepfake detection performance observed across these two datasets validates the effectiveness of PUDD across diverse scenarios.

5.3. Visualization

In this section, we make some visualizations to confirm the effectiveness of the proposed PUDD framework. Firstly, Figure 4 illustrates different similarity scores when different deepfakes generated from original videos are considered.

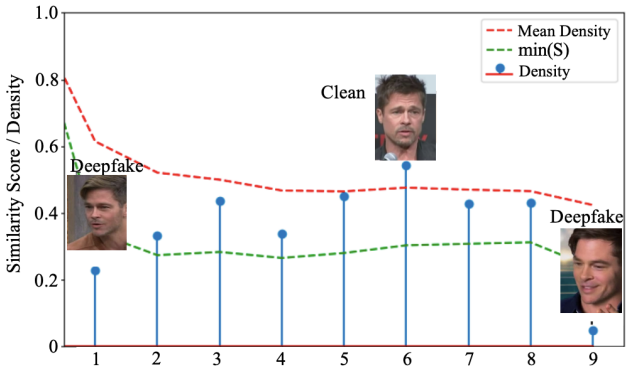

Figure 4. Similarity/Density score drop in deepfake videos.

From Figure 4, it is apparent that for each incoming sample, we can calculate and plot their density score relative to the existing prototypes, the mean density of our prototypes, and the minimum of similarity score. By considering these values, we can visually discern that a clean image will exhibit a density score higher than the minimum of similarity score, whereas DeepFaked images will be flagged as abnormal and will have a value below the minimum of similarity score.

Secondly, we present some qualitative result in Figure 5 to show the effectiveness of PUDD. The detection results of PUDD and MMtrace are denoted by green/red and black, respectively.

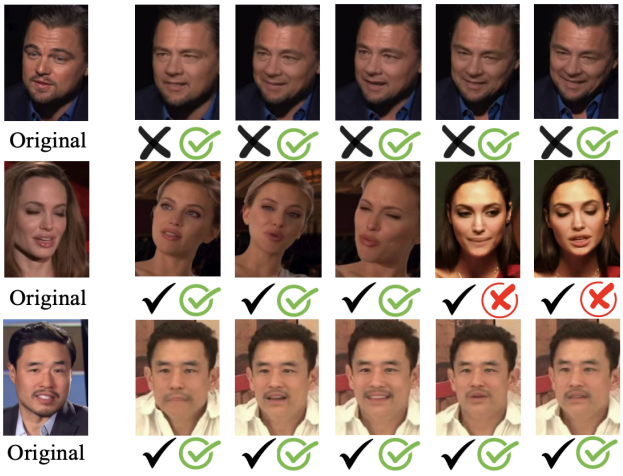

Figure 5. Challenging deepfakes in Celeb-DF. Black and green/red marks refer to the detection prediction from the MMtrace and PUDD, respectively.

As qualitative analysis, Figure 5 presents the deepfake detection results by using MMtrace and PUDD. We can observe that: (1) Both PUDD and MMtrace successfully detect the third celebrity as it is relatively easy to distinguish; (2) PUDD outperforms MMtrace in detecting the first celebrity by classifying them as an unseen class, leading to a more accurate decision; (3) PUDD fails to detect the second video of the second celebrity. This failure may be due to the celebrity frequently closing their eyes throughout the majority of the video, making prototype recognition more challenging.

5.4. Image Classification

As aforementioned, we estimate the prototypes for image classification as the upstream task, demonstrating promising performance of PUDD in both image classification and deepfake detection tasks during inference. In this experiment, we compare the image classification accuracy over original videos and images in the Celeb-DF [27] and CIFAKE datasets [5], respectively. The results are presented in Table 3.

It can be observed from Table 3 that PUDD achieves best image classification accuracy on both datasets, i.e., 92.7% and 96.4%, respectively. These results affirm the promising performance of PUDD across both tasks.

5.5. Interpretability

As aforementioned, the proposed PUDD learns prototypes from the data samples to provide interpretability. We calculate the similarity score (as described in Eq. 1) between an input image and all identified prototypes, so, we were able to extract a rule-based linguistic representation for each specific sample to explain the model's behavior as described:

Table 3. Image classification comparison on Celeb-DF and CIFAKE.

Method	Celeb-DF	CIFAKE
MPC-CA [38]	79.2	84.1
IPD [2]	87.4	87.9
ProtoExplorer [6]	92.0	94.6
NoiseDF [36]	90.1	94.7
FTCN [40]	92.3	93.6
MMtrace [30]	92.5	93.9
PUDD	**92.7**	**96.4**

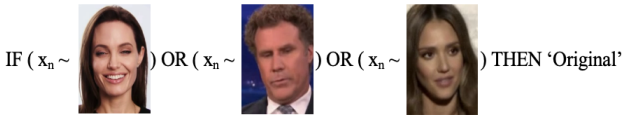

Figure 6. Linguistic rule-based representation of the prototypes for PUDD Interpretability on 'Original' class of Celeb-DF with top 3 closest prototypes on the feature space.

5.6. Environmental Impact

As aforementioned, PUDD only requires a limited number of parameters for retraining due to efficient prototype learning. We report the potential carbon emission of retraining a PUDD in Table 4. All models are trained on a single V100 GPU with a power consumption of 300 W.

Table 4. Carbon footprint of reproducing models. tCO_2eq refers to the tonnes of CO_2 equivalent.

Method	Total Power Consumption	tCO_2eq
NoiseDF [36]	23.2 kWh	1.2×10^{-2}
FTCN [40]	624.4 kWh	0.3
MMtrace [30]	17.5 kWh	8.7×10^{-3}
PUDD	0.2 Wh	10^{-7}

For comparison, retraining a MMtrace or PUDD would require 17.5 kWh and 0.2 Wh, respectively, if run in the same data center. This is 10^5 more carbon emission.

6. Discussion and Conclusion

The advantages of our proposed method are listed bellow:

1. In the training stage, our approach only require the access to original data. Therefore, different from conventional deepfake detection methods, we do not rely on paired

training data, which includes both original and deepfake samples. This characteristic of our method streamlines the training process and eliminates the need for paired samples, simplifying the data collection and labeling process.

2. The proposed PUDD exploits prototype information derived from original data, thereby rendering it agnostic to the specific deepfake generation techniques and models present in the inference data. Consequently, it exhibits the capability to effectively detect unseen deepfakes generated using different techniques and models than those encountered during the training stage. The experimental results further confirm the effectiveness of PUDD.

3. PUDD can be easily implemented with various feature extractors to detect deepfakes across diverse data modalities, including video and image. Moreover, PUDD offers flexibility for researchers to select a suitable feature extractor tailored to the specific characteristics of their target class.

4. The rapid retraining capability of PUDD, taking only 2.7 seconds, significantly accelerates its adoption in new domains compared to conventional deepfake detection methods. This speed makes PUDD highly feasible for potential real-world applications, enhancing its practicality and versatility.

5. The PUDD framework exploits image classification as the upstream task, enabling it to achieve promising performance in image classification despite being primarily designed for deepfake detection and trained on a deepfake dataset. This demonstrates the adaptability and robustness of PUDD across various tasks and datasets.

6. Prototype learning aids in understanding prototype-based classification by quantifying the extent to which a human can reliably predict the model's output.

7. Due to efficient prototype clustering and simplified calculations, PUDD requires 10^5 times less carbon emission than the state-of-the-art model, making it much more environmentally friendly.

Overall, we have proposed a Prototype-based Unified Framework for Deepfake Detection (PUDD) framework for deepfake video and image detection, offering an effective alternative to conventional competitors. Different from these conventional methods, we learned most representative prototypes in classes to efficiently detect deepfake samples and provide interpretability. Our evaluation with multimodal datasets has demonstrated the robust performance of the proposed method on both deepfake images and videos. Additionally, PUDD required only 2.7 seconds on new data, making it feasible for potential real-world applications.

Acknowledgment

This work is supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible.

References

[1] A. Aghasanli, D. Kangin, and P. Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[2] A. Aghasanli, D. Kangin, and P. Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 4, 5, 6

[3] P. Angelov and E. Soares. Detecting and learning from unknown by extremely weak supervision: exploratory classifier (xclass). *Neural Computing and Applications*, 33:15145–15157, 2021. 2

[4] M. Appel and F. Prietzel. The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4):1 – 13, 2022. 2

[5] J. J. Bird and A. Lcifi. CIFAKE: image classification and explainable identification of AI-generated synthetic images. *IEEE Access*, page 99, 2024. 3, 4, 5, 6

[6] M. L. Bouter, J. L. Pardo, Z. Geradts, and M. Worring. ProtoExplorer: interpretable forensic analysis of deepfake videos using prototype exploration and refinement. *arXiv preprint arXiv:2309.11155*, 2023. 2, 4, 5, 6

[7] S. Burgess. Ukraine war: deepfake video of Zelenskyy telling Ukrainians to 'lay down arms' debunked. *Sky News*, 2023. 1

[8] A. Chintha, A. Rao, S. Sohrawardi, K. Bhatt, M. Wright, and R. Ptucha. Leveraging edges and optical flow on faces for deepfake detection. *Proceedings of IEEE International Joint Conference on Biometrics (IJCB)*, 2020. 1

[9] F. Chollet. Xception: deep learning with depthwise separable convolutions. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings*

of International Conference on Learning Representations (ICLR), 2021. 1

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Conference on Neural Information Processing Systems (NeurIPS)*, 2014. 1

[12] Y. Guo, W. He, J. Zhu, and C. Li. A light autoencoder networks for face swapping. *Proceedings of International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, 2018. 2

[13] D. Gutiérrez and M. Mendoza. Bimodal neural style transfer for image generation based on text prompts. *Proceedings of International Conference on Human-Computer Interaction*, 2023. 1

[14] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[16] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao. Evading deepFake detectors via adversarial statistical consistency. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[17] J. Ice. Defamatory political deepfakes and the first amendment. *Case Western Reserve Law Review*, 70(2):417 – 455, 2019. 2

[18] R. Jones. Real person or deepfake? can You tell? *University of Waterloo*, 2023. 1

[19] E. Josephs, C. Fosco, and A. Oliva. Artifact magnification on deepfake videos increases human detection and subjective confidence. *Journal of Vision*, 23:5327, 2023. 1

[20] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[21] A. Khormali and J.-S. Yuan. Self-supervised graph Transformer for deepfake detection. *arXiv preprint arXiv:2307.15019*, 2023. 1

[22] A. Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis*, 2009. 4

[23] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. MaskGAN: towards diverse and interactive facial image manipulation. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[24] Y. Li, P. Angelov, and N. Suri. Domain generalization and feature fusion for cross-domain imperceptible adversarial attack detection. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2023. 2

[25] Y. Li, P. Angelov, and N. Suri. Rethinking self-supervised learning for cross-domain adversarial sample recovery. *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2024. 2

[26] Y. Li, Y. Sun, K. Horoshenkov, and S. M. Naqvi. Domain adaptation and autoencoder based unsupervised speech enhancement. *IEEE Transactions on Artificial Intelligence*, 3(1):43 – 52, 2021. 1

[27] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-DF: a large-scale challenging dataset for deepfake forensics. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4, 5, 6

[28] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh. DF-Platter: multi-face heterogeneous deepfake dataset. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[29] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: learning robust visual features without supervision. *arXiv preprint arXiv: 2304.07193*, 2023. 3, 4

[30] M. A. Raza and K. Malik. Multimodaltrace: deepfake detection using audiovisual representation learning. *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 4, 5, 6

[31] T. Reiss, B. Cavia, and Y. Hoshen. Detecting deepfakes without seeing any. *arXiv preprint arXiv:2311.01458*, 2023. 2

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021. 1, 4

[33] P. Sharma, M. Kumar, and H. K. Sharma. A GAN-based model of deepfake detection in social media. *Procedia Computer Science*, 218:2153–2162, 2023. 2

[34] E. Soares, P. Angelov, and N. Suri. Similarity-based deep neural network to detect imperceptible adversarial attacks. *Proceedings of IEEE Symposium Series on Computational Intelligence (SSCI)*, 2022. 2, 3, 4, 5

[35] D.-C. Stanciu and B. Ionescu. Autoencoder-based data augmentation for deepfake detection. *Proceedings of Annual ACM International Conference on Multimedia Retrieval (ICMR)*, 2023. 2

[36] T. Wang and K. Chow. Noise based deepfake detection via multi-head relative-interaction. *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 2023. 1, 4, 5, 6

[37] M. Zendrana and A. Rusiecki. Swapping face images with generative neural networks for Deepfake technology – experimental study. *Proceedings of International Conference on Knowledge-Based and Intelligent Information Engineering Systems*, 2021. 1

[38] F. Zhang, S. Tian, L. Yu, and Q. Yang. Multi-channels prototype contrastive learning with condition adversarial attacks for few-shot event detection. *Neural Processing Letters*, 56(31):30 – 31, 2024. 4, 5, 6

[39] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang. ISTVT: interpretable spatial-temporal video Transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1335 – 1348, 2023. 1

[40] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4, 5, 6