D^3: Scaling Up Deepfake Detection by Learning from Discrepancy

Supplementary Material

In the appendix, we add some experiments to delve deeper into the robustness of the differentiated features. We conduct a sensitivity analysis on the only variable of the patch-shuffle operation, the patch size, in Sec. A; we use different classification heads to demonstrate the robustness of the differentiated features to the classification heads in Sec. B; we showcase the samples that were correctly identified additionally by using differentiated features compared to the baseline UFD in Sec. C; we detailly report the test results of our method and the baseline across various generators in Sec. D; finally, we visualize the focus of UFD and D^3 to demonstrate D^3's ability of capturing more comprehensive and universal artifacts in Sec. E.

A. Sensitivity of Discrepancy Features to Patch Size

Patch size	Validataion					
	In-domain		Out-of-domain		Total	
	Mean acc.	AP	Mean acc.	AP	Mean acc.	AP
1	0.958	0.992	0.837	0.934	0.885	0.960
14	0.967	0.995	0.859	0.944	0.904	0.968
28	0.966	0.995	0.871	0.939	0.909	0.965
56	0.962	0.998	0.871	0.943	0.907	0.965
112	0.949	0.989	0.858	0.942	0.895	0.964
224	0.889	0.964	0.829	0.935	0.853	0.946

Table 1. **Results of different patch sizes on the validation set.** The ablated patch sizes range from 1 to 224 (the image size). The significant improvement brought by the switch from 224 to 112 shows the effectiveness of introducing discrepancy. Patch sizes 14, 28, and 56 achieve similarly high performance. But patch size 1's performance drops for the over-destruction of local artifacts.

We study how the patch size affects the learning of discrepancy signals. A tradeoff exists between increasing the discrepancy between features and mining the universal local artifacts, i.e. the smaller patch size offers more discrepancy but retains fewer local artifacts. Therefore, given the original image size of 224, we conduct validations with different patch sizes, ranging from 1 to 224, to see the changing trend. These experiments adhere to the previous setting, with only patch size being adjusted. As shown in Tab. 1, changing the patch size from 224 to 112 brings a significant improvement of 6.0 points in ID accuracy and 2.9 points in OOD accuracy, suggesting that additional discrepancy in features helps in expanding the representation of features and extracting universal artifacts. The patch sizes 14, 28, and 56 yield similarly high overall performance, showing the introduced discrepancy's robustness in different patch sizes. Note that when the patch size is 1, the local artifacts of the shuffled image are significantly affected, resulting in a drop in model performance compared to patch size 14. In our SOTA version, we directly opt for a patch size of 14 to align with the backbone CLIP:ViT-L/14 [34] while introducing the highest discrepancy in features.

B. Different Classifier Heads

Architecture	Validataion					
	In-domain		Out-of-domain		Total	
	Mean acc.	AP	Mean acc.	AP	Mean acc.	AP
FC	0.918	0.977	0.835	0.938	0.868	0.954
MLP	0.960	0.995	0.865	0.932	0.903	0.963
Self-Attention	0.967	0.995	0.859	0.944	0.904	0.968
Transformer	0.965	0.995	0.872	0.952	0.909	0.973

Table 2. **Results of different classifier heads on the validation set.** We evaluate four classifier network architectures and find that a network that learns the correlations between features will perform better.

We investigate how different classifier heads influence the model's performance to verify the effectiveness of artifact invariance learning. We evaluate four architectures: (i) **FC**: a single fully connected layer, (ii) **MLP**: A two-layer non-linear perceptron network with ReLU activation and a hidden layer dimension of 2×1024 neurons, (iii) **Self-Attention**: a network consisting of a self-attention layer [14] and a single fully connected layer, and (iv) **Transformer**: A network composed of two transformer encoder layers with 4 attention heads and a forward dimension of 4×1024 [14] and one fully connected layer.

Tab. 2 presents the results of these variants in our proposed experimental setting. The findings show that the results of MLP, Self-Attention, and Transformer are significantly improved compared to FC. This means establishing the correlations between the two discrepancy features helps learn universal artifacts. In addition, the performances of Self-Attention, MLP, and Transformer don't show an obvious gap, which demonstrates that our discrepancy features are highly distinguishable for deepfake detection.

C. Comparative Analysis of Uniquely Detected Samples

We take a further step to explore how our method outperforms. We present a selection of samples from both in-domain and out-of-domain. These samples are accurately classified by our approach, yet erroneously classified by the UFD [31], with a discrepancy in classification confidence

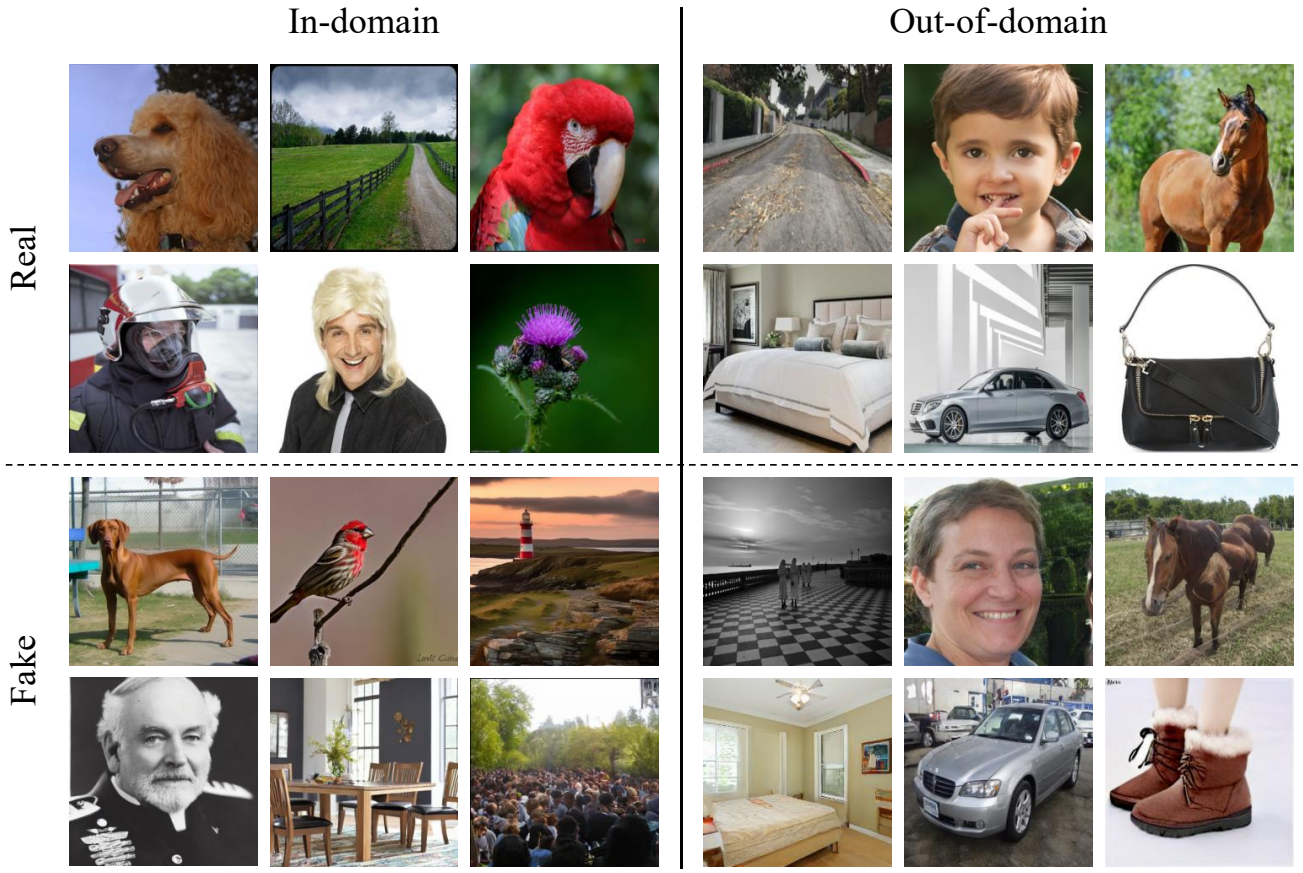

Figure 1. **Visualization of uniquely detected samples.** We present a selection of samples from both in-domain and out-of-domain, that are accurately classified by our approach, yet erroneously classified by the UFD [31], with a discrepancy in classification confidence exceeding 0.8.

	In-domain								Out-of-domain														
methods	ADM	Big GAN	GLI DE	Mid jour ney	LDM	VQ DM	wu kong	Pro GAN	Cycle GAN	Style GAN	Style GAN2	Gau GAN	Star GAN	Deep fakes	which-face isreal	SI TD	SAN	CRN	IM LE	DAL L·E	ID	OOD	Total
CNNDet	89.6	79.1	98.2	97.3	93.7	97.0	95.7	95.6	72.0	75.7	80.0	56.4	98.2	80.2	47.3	49.7	72.4	56.0	56.0	94.7	93.3	69.9	79.2
Patchfor	99.8	85.1	99.6	99.6	99.7	99.7	99.7	99.9	93.2	98.1	94.4	59.1	99.8	90.7	61.6	65.4	84.5	50.0	50.0	99.6	97.9	78.9	86.5
LNP	91.0	72.5	95.0	94.4	92.2	89.4	92.8	87.3	71.0	89.9	85.0	68.8	83.7	65.5	53.9	52.0	55.8	50.1	62.6	84.6	89.3	68.6	76.9
DIRE	99.9	82.6	99.9	99.8	99.9	100	100	98.4	60.3	94.2	95.1	55.3	88.6	67.9	50.0	50.0	59.7	50.0	50.0	99.7	97.6	68.4	80.1
UFD	83.2	92.0	86.3	80.2	85.4	89.0	85.6	91.1	74.8	79.4	82.4	95.5	89.0	75.9	79.8	73.1	61.9	87.9	90.0	86.6	86.6	81.4	83.5
Ours	94.8	98.5	95.0	96.8	94.4	96.7	97.1	99.4	92.7	94.9	95.7	98.1	96.0	67.7	83.1	73.8	62.6	88.1	95.0	92.8	96.6	87.6	90.7

Table 3. **Detailed mean accuracy results of comparisons with the state-of-the-art on the testing set.** We report the mean accuracy per generator in the percentage form. The results of generators with the same architecture but with different parameters are averaged.

exceeding 0.8. As shown in Fig.1, these samples are challenging to discern with the naked eye. This compellingly demonstrates that our method is capable of learning deeper and more universal artifacts, thereby retaining its effectiveness even when confronted with such challenging samples.

D. Detailed Mean Accuracy Results of Comparisons with the State-of-the-arts

In this section, we report the detailed mean accuracy results of comparisons with the state-of-the-art in Sec. 3, as a supplement to Table 1 in Sec 4.5, including ADM [13], BigGAN [4], GLIDE [30], Midjourney [2], LDM [36], VQDM [18], wukong [3], ProGAN [22], CycleGAN [51], StyleGAN [23], StyleGAN2 [24], GauGAN [32], StarGAN [8], Deepfakes [37], whichfaceisreal [1], SITD [6], SAN [11], CRN [7], IMLE [27], and DALL·E [35]. Results of generators with the same architecture but different parameters are averaged. For example, the result of BigGAN [4] in this table is the average of BigGAN in UFD [31] and BigGAN in GenImage [52].

Figure 2. Visualizations of fake image samples along with the corresponding occlusion maps [48] for each detector. These images represent cases accurately identified by D^3 where UFD failed.

E. Visualization of the Detector's Attention

D^3 effectively learns more universal artifacts through discrepancies introduced by image transformations, achieving superior performance. We employ the occlusion technique [48] to visually demonstrate this to identify regions of interest during inference, as illustrated in Fig. 2. Our analysis reveals that UFD focuses on limited regions, often leading to misjudgments due to its narrow attention. In contrast, D^3 exhibits a significantly broader attention range, further validating that the introduced discrepancy enables D3 to capture more comprehensive and universal artifacts.