
Supplementary for \mathcal{X}^2 -DFD: A framework for e \mathcal{X} plainable and e \mathcal{X} tendable Deepfake Detection

Anonymous Author(s)

Affiliation

Address

email

A More experiments on different types of face forgery

We evaluated a broader range of facial forgery types using the DiFF [4] dataset, a comprehensive collection of diffusion-generated facial images. This dataset encompasses Text-to-Image (T2I), Image-to-Image (I2I), Face Swapping (FS), and Face Editing (FE). The results presented in Table 1 and Table 2 demonstrate that our model exhibits strong generalization performance.

Text-to-Image Detection. We test four state-of-the-art methods—Midjourney [6], Stable Diffusion XL (SDXL) [7], FreeDoM T [16], and HPS [15]—for this subset. The first two are leading web services, accessed via their official APIs. The latter two are text-to-image (T2I) models, for which we employ pre-trained models. In detecting images generated by these models, Our detection model demonstrates superior generalization, achieving a 14.15% performance improvement over compared methods.

Image-to-Image Detection. We test four methods: LoRA [1], DreamBooth [10], SDXL Refiner [7], and FreeDoM I [16]. LoRA and DreamBooth fine-tune diffusion models for specific facial features. SDXL Refiner improves SDXL outputs, while FreeDoM I uses a visual encoder for face reconstruction. These methods use common forgery techniques like model fine-tuning, image enhancement, and visual prompt-guided synthesis to create realistic fake faces. In detecting images generated by these models, our detection model achieves a 14.03% improvement in generalization performance.

Diffusion Based Face Swapping Detection. In this subset, we apply DiffFace [4] and DCFace [5] for the face swapping task. Compared to several detection methods, in detecting images generated by these models, our detection model exhibits enhanced generalization, with a 10.07% performance improvement.

Face Editing Detection. This subset focuses on detecting images generated by three editing algorithms: Imagic [3], Cycle Diffusion (CycleDiff) [14], and Collaborative Diffusion (CoDiff) [2]. Our detection model outperforms compared detection methods, achieving a 21.53% improvement in generalization performance.

Our model demonstrates excellent generalization across four types of forgery tasks—text-to-image, diffusion-based face swapping, face editing, and image-to-image—encompassing 13 distinct forgery methods. Compared to other detectors, our approach consistently achieves superior performance, respectively, and robust performance across diverse datasets.

Table 1: Comparison of universal deepfake detection methods based on Frame-Level AUC on the DiFF dataset. The symbol ‘†’ denotes models specifically designed for deepfake detection, while ‘‡’ indicates models developed for general detection of generated face forgeries.

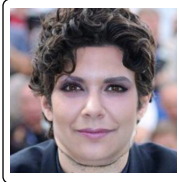
Method	Test Subset			
	T2I	I2I	FS	FE
Xception [†] [9]	62.43	56.83	85.97	58.64
F ³ -Net [†] [8]	66.87	67.64	81.01	60.60
EfficientNet [†] [11]	74.12	57.27	82.11	57.20
DIRE [‡] [13]	44.22	64.64	84.98	57.72
OURS	88.27	81.67	96.04	82.13

Table 2: Comparison of universal deepfake detection methods based on Frame-Level AUC on the DiFF dataset. The notation ‘†’ denotes models specifically designed for deepfake detection, while ‘‡’ indicates models developed for general detection of generated face forgeries. All experiments were trained on the c23 version of the FF++ dataset.

Method	Test Subset												
	FE			FS		I2I				T2I			
	Cycle	CoDiff	Imagic	DiFace	DCFace	Dream	SDXL_R	FD_I	LoRA	Midj	FD_T	SDXL	HPS
F ³ -Net† [8]	36.14	35.00	32.56	25.61	53.29	55.45	65.04	40.90	-	-	45.00	61.64	68.96
EfficientNet† [11]	56.51	38.38	48.50	64.45	89.13	71.64	65.04	59.93	-	-	69.67	64.94	74.63
CNN_Aug‡ [12]	50.31	46.84	75.95	43.10	80.69	58.75	60.10	43.65	-	-	47.90	61.95	60.56
OURS	88.10	57.36	95.00	94.13	96.69	82.87	82.87	64.98	90.52	85.52	83.22	92.10	89.16

36 B More Visual Sample

37 Our model effectively detects and explains artifacts associated with diffusion-generated images, as
 38 shown in Figures 1 to 3. Figures 1 and 2 illustrate cases where the Specific Feature Detector (SFD)
 39 do not detect forgery-related evidence, yet the model still correctly identifies the forgery based on the
 40 image analysis. In contrast, Figure 3 demonstrates that when SFD detects evidence of forgery in the
 41 image, the model provides a corresponding explanation.



Prompt by adding SFD information: Is this image real or fake? And the blending feature score of content is 0.0058.

Answer: This image is **fake**, the **eyebrows are uneven** with one side looking **slightly distorted in shape**, the neck doesn't match the face with a **weird color and transition**, and the **face contour has sharp, unnatural lines around the jaw and chin**.

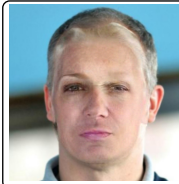
Figure 1: From the test dataset on DIFF, we show the sample of diffusion generated image detection and explanation, where the specific feature detector found no evidence of a particular forgery in the image.



Prompt by adding SFD information: Is this image real or fake? And the blending feature score of content is 0.0019.

Answer: This image is **fake**, the **nostrils are uneven**, with one larger than the other, making the nose look unnatural. The **skin tone** on the cheeks is patchy and unnatural. The forehead is too smooth and shiny, appearing rigid and edited.

Figure 2: From the test dataset on DIFF, we show the sample of diffusion generated image detection and explanation, where the specific feature detector found no evidence of a particular forgery in the image.



Prompt by adding SFD information: Is this image real or fake? And the blending feature score of content is 0.6951.

Answer: This image is **fake**, as the eyes are **unnaturally symmetrical**, the nose appears **misaligned** and **unnaturally shaped**, with **uneven contours around the nostrils**. Furthermore, the facial contour near the **jawline and cheeks shows irregular transitions**. And the image contains obvious **blending artifact**.

Figure 3: From the test dataset on DIFF, we show the samples of diffusion generated image detection and explanation, where the specific feature detector found evidence of a particular forgery in the image.

References

- [1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [2] Ziqi Huang, Kelvin C. K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *CVPR*, pages 6080–6090, 2023.
- [3] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023.
- [4] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *CoRR*, pages 1–11, 2022.
- [5] Minchul Kim, Feng Liu, Anil K. Jain, and Xiaoming Liu. Dcfac: Synthetic face generation with dual condition diffusion model. In *CVPR*, pages 12715–12725, 2023.
- [6] Midjourney. <https://www.midjourney.com>, 2022.
- [7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, pages 1–21, 2023.
- [8] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.
- [9] 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 Conference on International Conference on Computer Vision*, 2019.
- [10] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.
- [11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [12] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.
- [13] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.
- [14] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, pages 7378–7387, 2023.
- [15] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.
- [16] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *ICCV*, pages 23174–23184, 2023.