

Supplementary Materials: Diffusion Facial Forgery Detection

Anonymous Authors

A ABSTRACT

The content of this supplementary material is divided into three main sections. Section B provides a detailed description of the thirteen synthetic methods we employed. Section C offers additional evaluation results, along with tables for the values in Figure 5 presented in the main manuscript. Section D presents comprehensive visualizations, including an overview of the overall pipeline for generating forged images, as well as numerous non-cherry-picked samples from the thirteen methods.

B DIFFUSION APPROACHES

We have employed a total of 13 approaches¹ to generate forged images, and a detailed description of these methods is as follows,

B.1 Text-to-Image

- **Midjourney** [38] is one of the largest online-accessible AI art creation service providers. We directly invoke the ‘/imagine’ command to generate images from prompts.
- **Stable Diffusion XL (SDXL)** [42]. SDXL is the drastically improved version of stable diffusion models [48], which encompasses three times of the UNet parameters of the previous ones. We utilize the official open-source code and parameters to generate images.
- **FreeDoM_T** [70] uses off-the-shelf pre-trained models to construct the energy function [73] and generates images with various conditions. In the T2I subset, we utilize CLIP [44] as the text encoder to guide face synthesis.
- **HPS** [65] trains a human preference classifier with the collected dataset and derive a human preference score to adapt Stable Diffusion to better align with human preferences through a human preference classifier.

B.2 Image-to-Image

- **Low-Rank Adaption (LoRA) Diffusion** [20]. LoRA is to freeze the pre-trained model weights and inject trainable layer into the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. With the I2I setting, we train the LoRA layer with images of one specific identity, which enables the model to learn the appearance of that identity, and output faces visually align closely with this person.
- **DreamBooth** [51]. DreamBooth offers another fine-tuning manner to personalize existing diffusion models such that it learns to bind a unique identifier to specific subject, by leveraging the semantic prior embedded with a new auto-genous class-specific prior preservation loss.
- **SDXL Refiner** [42]. The refiner model is designed to de-noise the small noise of T2I-generated images from SDXL [37], making them smoother and more realistic.

- **FreeDoM_I**. It is flexible to run FreeDoM with visual conditions. As a result, we utilize the real image with its corresponding facial appearance, sketches, landmarks, and segmentations for I2I generation.

B.3 Face Swapping

- **DiffFace** [26] is one of the first efforts to apply diffusion model in face swapping task, which utilizes the facial expert models to transfer source identity while preserving target attributes faithfully.
- **DCFace** [27] combines the source and target as two conditions, *i.e.*, appearance (ID) and external factor (style), respectively. These two conditions provide guidance to the dual-condition diffusion model and produce face images of the same subject under different styles.

B.4 Face Editing

- **Imagic** [25] produces text embeddings that align with both the input image and the desired edit, and fine-tunes the pre-trained diffusion model to operate editing. In the implementation, we construct new text prompts to guide image editing by replacing key entities in each textual prompt, such as gender, skin color, or age. Each prompt is executed four times, ultimately yielding over 40,000 synthesized images.
- **Collaborative Diffusion (CoDiff)** [22] employs pre-trained uni-modal diffusion models collaborate to achieve multi-modal face generation and editing without re-training. In facial modification applications, we alter facial expressions and movements by modifying the facial segmentation masks.
- **Cycle Diffusion (CycleDiff)** [64] is a method for unpaired image-to-image translation, which presents a reconstructable encoder for stochastic diffusion probabilistic models (DPMs). We apply the modified prompts in Imagic as desired edit conditions to generate outcomes with CycleDiff.

C MORE BENCHMARKS

In this section, we first presented the specific values for Figure 5, followed by additional experimental results. These include a further evaluation of the DiFF dataset and a new benchmark of the utilization of EGR.

C.1 Exact Values in Figures

Table S1 and Table S2 correspond to the specific values of Figure 5, respectively.

C.2 Results on Full DiFF

Trained on single subset. In Table S3, we presented the test results on the complete DiFF dataset (*i.e.*, across all subsets) following

¹Adjustments to the watermark implementation in released models are made to prevent detectors from discerning authenticity via watermarks.

Table S1: AUC (%) comparison among re-training detectors correspond to Figure 5. Each row represents the performance of the model trained on a specific subset and tested on all four subsets.

Method	Train	Testing			
		T2I	I2I	FS	FE
Xception	T2I	93.32	86.85	34.65	23.28
F ³ -Net		99.70	88.50	45.07	71.06
EfficientNet		99.89	89.72	21.49	49.63
DIRE		95.04	84.07	35.15	50.86
Xception	I2I	87.82	98.92	36.82	33.39
F ³ -Net		87.23	99.50	40.62	46.19
EfficientNet		84.39	99.80	19.47	27.46
DIRE		86.20	99.88	41.51	42.01
Xception	FS	23.17	24.47	99.95	10.17
F ³ -Net		35.43	30.39	99.98	20.79
EfficientNet		16.88	22.17	99.87	10.21
DIRE		16.80	36.27	99.09	32.68
Xception	FE	80.84	79.12	70.81	99.95
F ³ -Net		82.32	76.92	56.27	99.60
EfficientNet		80.41	63.06	66.62	99.24
DIRE		56.70	59.22	43.78	99.87

Table S2: AUC (%) of Xception with different training strategies correspond to Figure 5b. Each row represents the performance of the model trained on a specific subset and tested on all four subsets.

Strategy	Train	Testing			
		T2I	I2I	FS	FE
Re-training	T2I	93.32	86.85	34.65	23.28
Linear Probing		71.36	74.75	65.83	66.41
Fine-Tuning		93.66	88.94	36.10	37.51
Re-training	I2I	87.82	98.92	36.82	33.39
Linear Probing		79.88	85.88	87.68	76.23
Fine-Tuning		97.79	98.76	61.64	48.33
Re-training	FS	23.17	24.47	99.95	10.17
Linear Probing		60.77	66.45	92.93	60.78
Fine-Tuning		18.04	24.71	99.44	16.48
Re-training	FE	80.84	79.12	70.81	99.95
Linear Probing		56.52	65.54	82.52	78.13
Fine-Tuning		89.74	74.10	75.96	99.93

training on a single subset. Furthermore, we reported on the performance of the model when trained on one subset and tested on a combination of other subsets (e.g., trained on T2I and tested on I2I+FS+FE). It can be observed that the EGR method significantly increases the detection performance of the models, which is consistent with the observations from Table 9.

Trained on full DiFF. We also evaluated the models' performance when trained and tested on the complete DiFF dataset, with the experimental results depicted in Table S4. We observed that when the distributions of the training and test sets are identical, the models are capable of achieving satisfactory performance, and the implementation of EGR can enhance the model to a certain extent. However, it is noteworthy that this evaluation strategy does not reflect the detectors' generalizability. In fact, drawing from previous experimental results (e.g., Table 9 and Table S3), we are aware that

Table S3: AUC (%) comparison among re-trained detectors when tested full DiFF (i.e., four subsets) or full DiFF without training subset (i.e., three subsets).

Method	Train	Testing Subset	
		Full	Others
Xception	×	71.23	50.18
Xception	✓	81.07	68.17
F ³ -Net	×	72.41	58.56
F ³ -Net	✓	80.97	66.50
EfficientNet	×	79.36	60.02
EfficientNet	✓	80.37	62.87
DIRE	×	66.85	53.28
DIRE	✓	72.38	54.37
Xception	×	78.01	73.78
Xception	✓	84.26	81.10
F ³ -Net	×	78.76	72.78
F ³ -Net	✓	79.28	76.22
EfficientNet	×	75.61	70.73
EfficientNet	✓	88.85	86.75
DIRE	×	76.54	73.57
DIRE	✓	84.94	82.16
Xception	×	37.47	22.58
Xception	✓	68.45	60.93
F ³ -Net	×	46.36	33.62
F ³ -Net	✓	77.51	66.32
EfficientNet	×	34.73	17.38
EfficientNet	✓	65.19	56.89
DIRE	×	69.23	58.35
DIRE	✓	74.67	69.62
Xception	×	87.97	85.78
Xception	✓	89.30	87.35
F ³ -Net	×	81.69	80.70
F ³ -Net	✓	87.21	84.91
EfficientNet	×	81.96	78.67
EfficientNet	✓	83.62	80.63
DIRE	×	81.06	78.35
DIRE	✓	88.79	83.26

Table S4: AUC (%) comparison among re-trained detectors. Each row represents the performance when trained and tested on full DiFF dataset.

Method	Train Dataset	Training Strategy	
		Baseline	EGR
Xception	DiFF	93.87	97.81
F ³ -Net		98.37	99.05
EfficientNet		94.34	99.26
DIRE		96.35	98.62

the generalizability of the current detectors presents a significant challenge. This issue lies at the heart of the current issues on the detection of forged images. Therefore, we advocate for the validation of detectors based on their performance when trained on a single subset and subsequently tested on multiple unseen subsets.

C.3 Detection with Post-processing Methods

Results of EGR with post-processing approaches. We also evaluated the performance of models incorporating EGR when dealing with several post-processing approaches. As demonstrated in Table S5, EGR enhances the models' robustness for four different post-processing approaches. For instance, when the EfficientNet is

Table S5: AUC (%) comparison of detectors with and without our EGR method with different post-processing methods. Each row represents the average performance when tested on all four DiFF subsets. Better results are highlighted in bold. GN: Gaussian Noise; GB: Gaussian Blur; MB: Median Blur; JPEG: JPEG Compression.

Method	+EGR	Train Subset	Processing Method			
			GN	GB	MB	JPEG
Xception	×	T2I	47.65	15.02	56.59	58.69
Xception	✓		62.49	32.38	70.87	70.69
EfficientNet	×		40.09	53.62	65.35	54.98
EfficientNet	✓		42.10	62.47	69.51	62.28
Xception	×	I2I	19.70	54.09	58.07	63.66
Xception	✓		63.19	54.84	70.13	70.25
EfficientNet	×		27.76	54.75	52.39	51.01
EfficientNet	✓		43.13	77.32	68.13	59.39
Xception	×	FS	35.40	34.82	38.58	37.73
Xception	✓		58.59	48.49	58.30	66.36
EfficientNet	×		36.74	23.82	36.12	35.13
EfficientNet	✓		49.27	32.01	36.51	45.57
Xception	×	FE	39.69	24.15	79.35	81.19
Xception	✓		57.99	25.51	87.71	87.52
EfficientNet	×		51.95	39.65	71.10	71.14
EfficientNet	✓		61.04	39.74	74.30	72.11

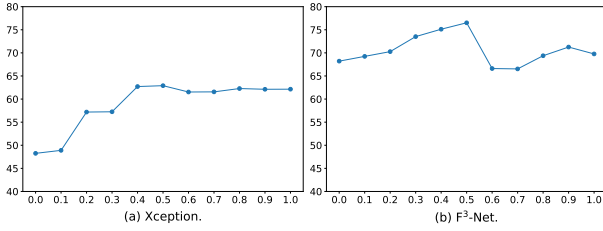


Figure S1: AUCs (%) of (a) Xception and (b) F³-Net. All of the models are trained on the T2I subset and tested on the other three subsets. Y-axis: Average AUC of detectors across three subsets.

trained on I2I, integrating EGR results in an average AUC improvement of 10%. This can be attributed to EGR guiding the models' attention to high-level features such as facial contours. These features offer more reliable evidence for distinguishing genuine from forged faces after post-processing.

C.4 More Ablation Study on EGR

Studies on parameter λ of EGR. We evaluated the impact of varying values of λ in Equation (3). Specifically, we utilized T2I as the training subset and assessed the average AUC across the remaining three subsets. Figure S1 shows that as the value of λ increases, the models' performance gradually improves and stabilizes when λ reaches 0.5. Beyond this value, there's a minor decline in AUC. There is a slight decrease in AUC. This may be attributed to the model excessively emphasizing edge graphs, potentially overlooking the color and texture features in pristine images.



Figure S2: More edge graphs from DiFF.

D VISUALIZATIONS OF DIFF

D.1 Edge Graphs

In Figure S2, we present more edge graphs. Furthermore, we offer a detailed explanation of the extraction for edge graphs. Specifically, we employ the Sobel operator to capture the edge graphs [59]. The Sobel operator is an influential technique in image processing. This operator functions by accentuating regions of high frequency, which correspond to edges. The core of the Sobel operator lies in its use of two distinct 3x3 kernels, each designed for detecting horizontal and vertical edges.

For horizontal edge detection:

$$M_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}. \quad (S1)$$

For vertical edge detection:

$$M_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}. \quad (S2)$$

Each pixel in the image is subjected to both kernels by superimposing them on the pixel and its immediate neighbors. This operation essentially computes a weighted average, accentuating changes in intensity across the specified directions. The gradient magnitude at each pixel is ascertained by integrating the results from both horizontal and vertical convolutions. This is typically computed using the formula:

$$G = \sqrt{G_x^2 + G_y^2}, \quad (S3)$$

where G_x and G_y represents the result of the horizontal and vertical convolution, respectively. Furthermore, the direction of the gradient can be calculated at each pixel, which offers insights into the orientation of edges. By establishing a threshold, gradients exceeding this value are classified as edges, thereby facilitating the generation of an edge graph.

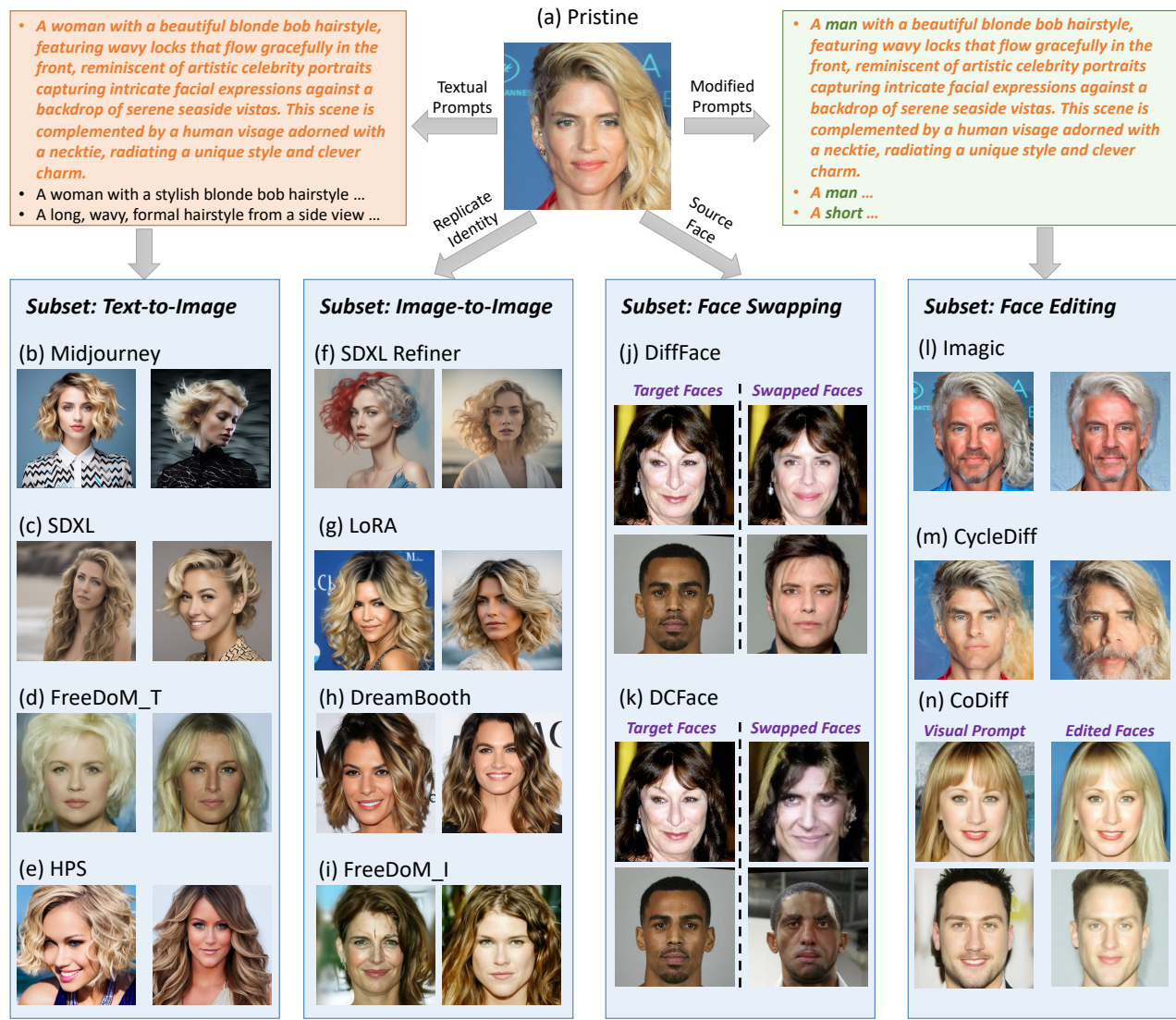


Figure S3: Visualization of one sample from DiFF. All synthesized images can be associated with the pristine images through textual or visual prompts.

D.2 Pristine and Forged Images

In Figure S3, we presented synthetic images corresponding to four major conditions of DiFF, along with their synthesis process. Specifically, each forged image is generated using specific prompts and maintains semantic consistency with these prompts. All of these prompts are collected from facial features in real images, thus establishing an association between each synthetic image and the pristine images.

Subsequently, Figure S5 – Figure S17 showcase large collections of synthetic images in DiFF from different synthesized methods, respectively.



Figure S4: Visualization results of pristine images.



Figure S5: Visualization results of Midjourney (T2I subset).



Figure S6: Visualization results of SDXL (T2I subset).



Figure S7: Visualization results of Freedom_T (T2I subset).

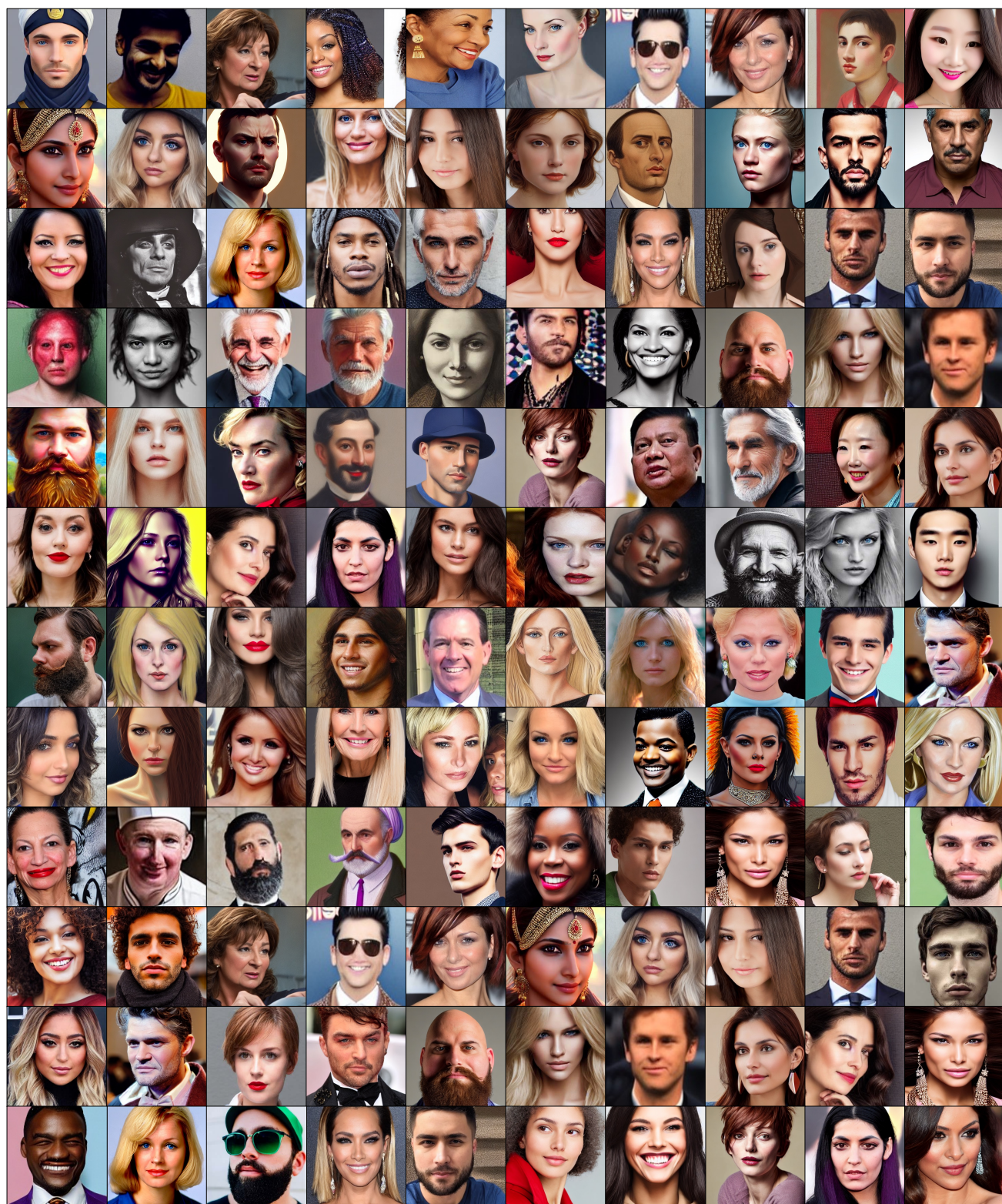


Figure S8: Visualization results of HPS (T2I subset).



Figure S9: Visualization results of SDXL Refiner (I2I subset).



Figure S10: Visualization results of LoRA (I2I subset).



Figure S11: Visualization results of DreamBooth (I2I subset).

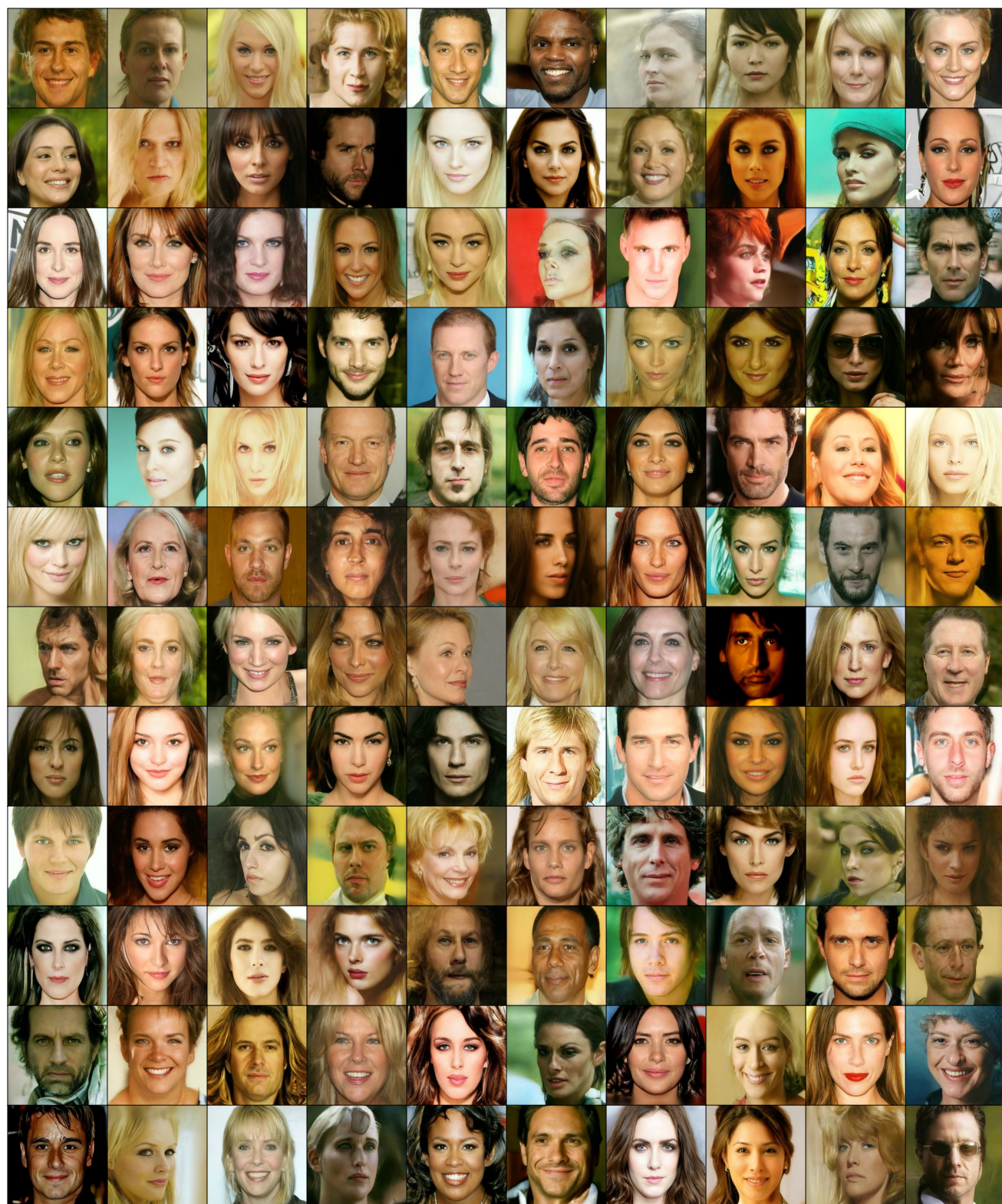


Figure S12: Visualization results of FreeDoM_I (I2I subset).

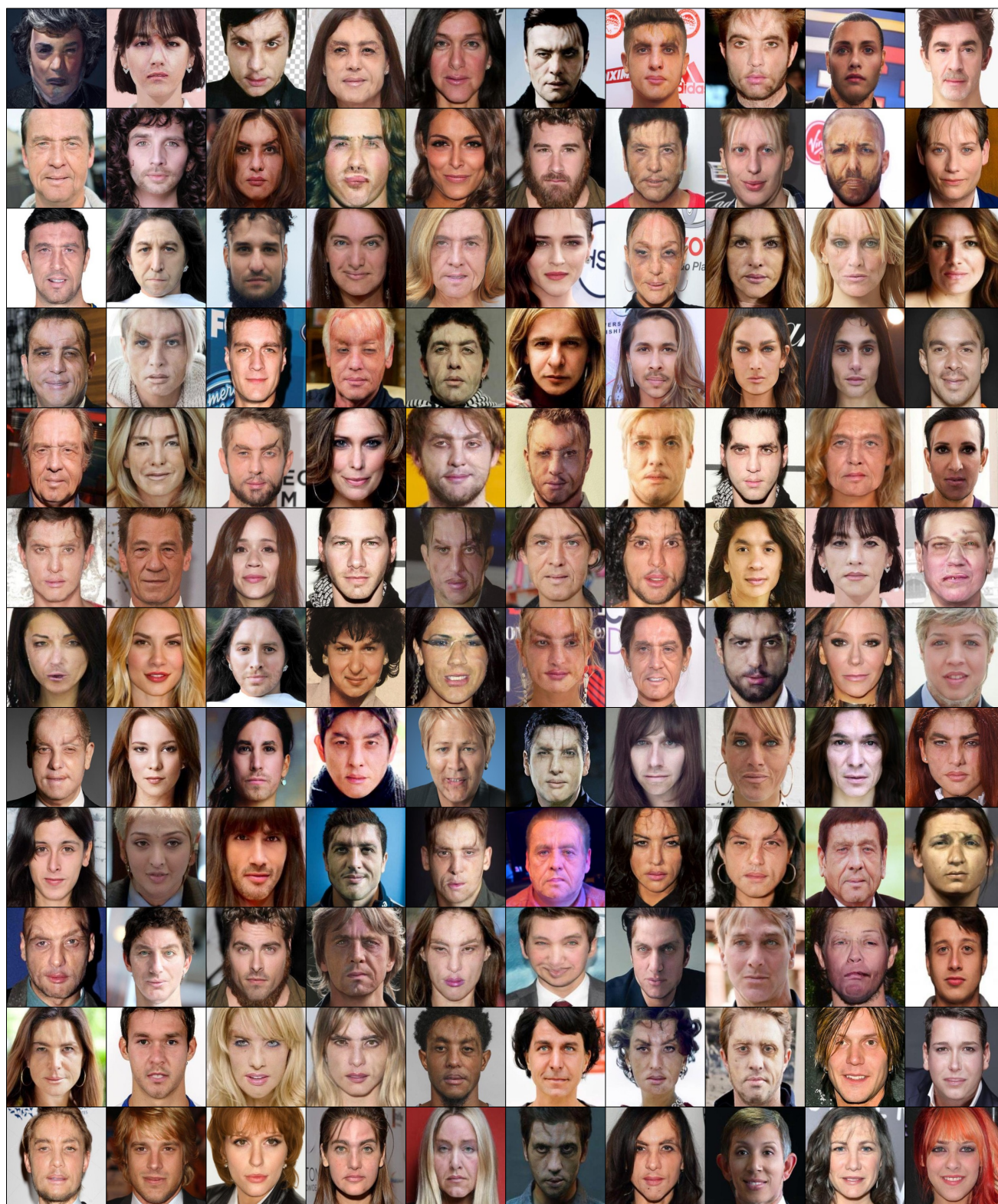


Figure S13: Visualization results of DiffFace (FS subset).



Figure S14: Visualization results of DCFace (FS subset).



Figure S15: Visualization results of Imagic (FE subset).



Figure S16: Visualization results of CoDiff (FE subset).



Figure S17: Visualization results of CycleDiff (FE subset).

REFERENCES

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *CVPR*. 18187–18197.
- [2] Yeqi Bai, Tao Ma, Lipo Wang, and Zhenjie Zhang. 2022. Speech Fusion to Face: Bridging the Gap Between Human's Vocal Characteristics and Facial Imaging. In *ACM MM*. 2042–2050.
- [3] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. 2022. Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models. In *ICLR*. 1–12.
- [4] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. 2022. Unleashing Transformers: Parallel Token Prediction with Discrete Absorbing Diffusion for Fast High-Resolution Image Generation from Vector-Quantized Codes. In *ECCV*. 170–188.
- [5] Ali Borji. 2022. Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. *CoRR* (2022), 1–4.
- [6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In *CVPR*. 18689–18698.
- [7] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. In *ICLR*. 1–9.
- [8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*. 1086–1090.
- [9] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On The Detection of Synthetic Images Generated by Diffusion Models. In *ICASSP*. 1–5.
- [10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Reverse Stable Diffusion: What prompt was used to generate this image? *CoRR* (2023), 1–13.
- [11] Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*. 8780–8794.
- [12] Brian Dolhansky, Joanna Bitton, Ben Pfau, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. 2020. The DeepFake Detection Challenge Dataset. *CoRR* (2020), 1–13.
- [13] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *ICML*. 3247–3258.
- [14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *CVPR*. 10686–10696.
- [15] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical Fine-Grained Image Forgery Detection and Localization. In *CVPR*. 3155–3165.
- [16] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2022. Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection. In *CVPR*. 14930–14942.
- [17] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In *CVPR*. 4360–4369.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*. 1–12.
- [19] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *JMLR* 23 (2022), 47:1–47:33.
- [20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*. 1–13.
- [21] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *CVPR*. 4490–4499.
- [22] Ziqi Huang, Kelvin C. K. Chan, Yuming Jiang, and Ziwei Liu. 2023. Collaborative Diffusion for Multi-Modal Face Generation and Editing. In *CVPR*. 6080–6090.
- [23] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In *ECCV*. 106–124.
- [24] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2Human: text-driven controllable human image generation. *ACM Transactions on Graphics* 41, 4 (2022), 162:1–162:11.
- [25] Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *CVPR*. 6007–6017.
- [26] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. 2022. DiffFace: Diffusion-based Face Swapping with Facial Guidance. *CoRR* (2022), 1–11.
- [27] Minchul Kim, Feng Liu, Anil K. Jain, and Xiaoming Liu. 2023. DCFace: Synthetic Face Generation with Dual Condition Diffusion Model. In *CVPR*. 12715–12725.
- [28] Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu. 2022. BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis. In *ICLR*. 1–12.
- [29] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *CVPR*. 5548–5557.
- [30] Haodong Li, Weiqi Luo, and Jiwei Huang. 2017. Localization of Diffusion-Based Inpainting in Digital Images. *IEEE TIFS* 12, 12 (2017), 3050–3064.
- [31] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-Ray for More General Face Forgery Detection. In *CVPR*. 5000–5009.
- [32] Yixuan Li, Chao Ma, Yichao Yan, Wenhan Zhu, and Xiaokang Yang. 2023. 3D-Aware Face Swapping. In *CVPR*. 12705–12714.
- [33] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2022. Pseudo Numerical Methods for Diffusion Models on Manifolds. In *ICLR*. 1–11.
- [34] Yaqi Liu, Chao Xia, Xiaobin Zhu, and Shengwei Xu. 2022. Two-Stage Copy-Move Forgery Detection With Self Deep Matching and Proposal SuperGlue. *IEEE TIP* 31 (2022), 541–555.
- [35] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *CVPR*. 11451–11461.
- [36] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-Branch Recurrent Network for Isolating Deepfakes in Videos. In *ECCV*. 667–684.
- [37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*. 1–14.
- [38] Midjourney. 2022. <https://www.midjourney.com>. <https://www.midjourney.com/>
- [39] Shivansh Mundra, Gonzalo J. Aniano Porcile, Smit Marvaniya, James R. Verbus, and Hany Farid. 2023. Exposing GAN-Generated Profile Photos from Compact Embeddings. In *CVPRW*. 884–892.
- [40] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. In *ICML*. 8162–8171.
- [41] Nikita Pavlichenko and Dmitry Ustulov. 2023. Best Prompts for Text-to-Image Models and How to Find Them. In *SIGIR*. 2067–2071.
- [42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR* (2023), 1–21.
- [43] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *ECCV*. 86–103.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*. 8748–8763.
- [45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *ICML*. 8821–8831.
- [46] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. 2022. Towards the Detection of Diffusion Model Deepfakes. *CoRR* (2022), 1–11.
- [47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. 10674–10685.
- [48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*. IEEE, 10674–10685.
- [49] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *ICCV*. 1–11.
- [50] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. In *CVPR*. 10219–10228.
- [51] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *CVPR*. 22500–22510.
- [52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*. 1–15.
- [53] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2022. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *CoRR* (2022), 1–14.
- [54] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. D2C: Diffusion-Decoding Models for Few-Shot Conditional Generation. In *NeurIPS*. 12533–12548.
- [55] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*. 2256–2265.

- [56] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *ICLR*. 1–12.
- [57] Andreas Stöckl. 2022. Evaluating a Synthetic Image Dataset Generated with Stable Diffusion. *CoRR* (2022), 1–13.
- [58] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*. 6105–6114.
- [59] O Rebecca Vincent, Olusegun Folorunso, et al. 2009. A descriptive algorithm for sobel image edge detection. In *InSITE*. 97–107.
- [60] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *CVPR*. 8692–8701.
- [61] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. 2023. Dynamic Graph Learning with Content-guided Spatial-Frequency Relation Reasoning for Deepfake Detection. In *CVPR*. 7278–7287.
- [62] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. In *ICCV*. 22445–22455.
- [63] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. 2023. AltFreezing for More General Video Face Forgery Detection. In *CVPR*. 4129–4138.
- [64] Chen Henry Wu and Fernando De la Torre. 2023. A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance. In *ICCV*. 7378–7387.
- [65] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Better Aligning Text-to-Image Models with Human Preference. In *ICCV*. 2096–2105.
- [66] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. 2020. SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features. In *ICASSP*. 2952–2956.
- [67] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. 2022. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In *ICLR*. 1–15.
- [68] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. ReCo: Region-Controlled Text-to-Image Generation. In *CVPR*. 14246–14255.
- [69] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *CoRR* (2015), 1–9.
- [70] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. 2023. FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. *ICCV* (2023), 23174–23184.
- [71] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S. Jaakkola, and Shiyu Chang. 2023. Towards Coherent Image Inpainting Using Denoising Diffusion Implicit Models. In *ICML*. 41164–41193.
- [72] Qingsheng Zhang and Yongxin Chen. 2023. Fast Sampling of Diffusion Models with Exponential Integrator. In *ICLR*. 1–12.
- [73] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. 2022. EGSD: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations. In *NeurIPS*. 1–14.
- [74] Yipin Zhou and Ser-Nam Lim. 2021. Joint Audio-Visual Deepfake Detection. In *ICCV*. 14800–14809.
- [75] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. HiDDeN: Hiding Data With Deep Networks. In *ECCV*. 682–697.
- [76] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. *CoRR* (2023), 1–11.