2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10648241

A TRUSTWORTHY AUTHENTICATION AGAINST VISUAL MASTER FACE DICTIONARY ATTACKS (TRAUMA)

Muhammad Mohzary[1], *Baek-Young Choi*[2], *Sejun Song*[2]

[1]Department of Computer Science, Jazan University, Jazan, Saudi Arabia
[2]School of Science and Engineering, University of Missouri-Kansas City, MO, USA

ABSTRACT

Facial Recognition Systems (FRS) have become one of the most viable biometric identity authentication approaches in supervised and unsupervised applications. However, FRSs are known to be vulnerable to adversarial attacks such as identity theft and presentation attacks. The master face dictionary attacks (MFDA) leveraging multiple enrolled face templates have posed a notable threat to FRS. Federated learning-based FRS deployed on edge or mobile devices are particularly vulnerable to MFDA due to the absence of robust MF detectors. To mitigate the MFDA risks, we propose a trustworthy authentication system against visual MFDA (Trauma). Trauma leverages the analysis of specular highlights on diverse facial components and physiological characteristics inherent to human faces, exploiting the inability of existing MFDAs to replicate reflective elements accurately. We have developed a feature extractor network that employs a lightweight and low-latency vision transformer architecture to discern inconsistencies among specular highlights and physiological features in facial imagery. Extensive experimentation has been conducted to assess Trauma's efficacy, utilizing public GAN-face detection datasets and mobile devices. Empirical findings demonstrate that Trauma achieves high detection accuracy, ranging from 97.83% to 99.56%, coupled with rapid detection speeds (less than 11 ms on mobile devices), even when confronted with state-of-the-art MFDA techniques.

Index Terms— Master Face Dictionary Attacks, Trustworthy, GAN-generated Faces, Facial Recognition Systems, Vision Transformer.

1. INTRODUCTION

With significant advancements in vision-based artificial intelligence (AI) technologies, Facial Recognition Systems (FRS) have emerged as one of the most practical and viable authentication approaches. The utilization of FRS has been gaining traction in various sectors, such as payment, access control, and security, due to their quick authentication processes and contactless and uninterrupted user interaction. However, FRS's trustworthiness is known to be vulnerable to various adversarial attacks, such as identity theft, spoofing, and presentation attacks. Recently, Generative Adversarial Networks (GAN) generated master face dictionary attacks (MFDA)[1, 2] pose a significant risk to FRS with the reasonably high matching ratio (40%) to multiple enrolled face templates. MFDA is especially damaging to FRS applications with the federated learning environment on the edge and mobile devices due to the lack of computationally effective master face detectors. While [1] has suggested that GAN face detection methods may be able to detect MFDA, there is currently no widely accepted or implemented MFDA countermeasure available for edge and mobile applications. Hence, developing a lightweight and real-time MFDA detector optimized for the edge computing environment could significantly enhance spoofing detection capabilities, ultimately enabling AI-based FRS's more widespread and secure deployment in edge computing applications.

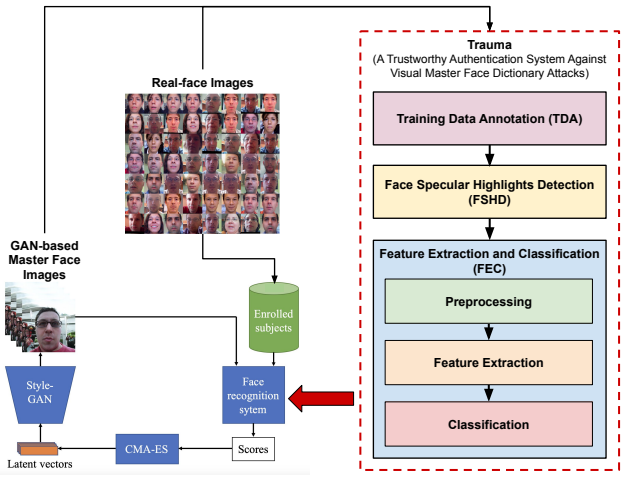

Fig. 1: An Overview of Trauma Authentication System Against Visual Master Face Dictionary Attacks (MFDA) [1].

This paper presents a novel, **tr**ustworthy **au**thentication system against visual **M**FD**A** (**Trauma**). Trauma takes specular reflections on different facial parts (e.g., eyes, cheeks, nose, chin, forehead, etc.) to extract their physiological characteristics, such as intensity and shape. We hypothesize that the existing MFDAs fail to coordinate their counterfeits with the

reflective elements on each facial component and demonstrate noticeable physiological flaws on different facial parts. Instead of assessing particular facial attributes or features, we have developed a streamlined and sensible feature extraction network based on Vision Transformer (ViT) technology [3]. This network can identify incongruences between specular highlights and physiological traits by analyzing non-overlapping, minuscule segments of a facial image. Our lightweight and low-latency approach renders it an efficient and practical solution for facial recognition tasks. The Trauma model leverages the strengths of Convolutional Neural Networks (CNN) and ViT architectures to incorporate and process local and global information, ultimately improving the representation learning process from facial images with fewer parameters. By fusing these two methods, Trauma effectively encodes the spatially localized features captured by CNNs with the global context awareness capabilities of ViT. As illustrated in Figure 1, Trauma comprises Training Data Annotation (TDA), Face Specular Highlights Detection (FSHD), and Feature Extraction and Classification (FEC) modules. We create a new Trauma dataset with high-resolution images from real and master faces (MF). The *TDA* annotates the Face Specular Highlight (FSH) regions with a range of environmental parameters to enable more accurate and precise analysis. For a given input image, the *FSHD* module can identify various FSHs across different regions of the face by analyzing the HSV (hue, saturation, value) color space of the pixels within the image. The *FEC* module employs a lightweight ViT-based backbone model to extract features effectively from the *FSH* images. The extracted features are employed to classify the input image, distinguishing between MFDA-generated and authentic facial representations. To gauge the effectiveness of Trauma, we have conducted comprehensive experiments, assessing its performance using publicly available GAN-face detection datasets. The empirical results show that Trauma achieves high accuracy, ranging from 97.83% to 99.56% against state-of-the-art (SOTA) MFDA and fast detection speed (less than 11 ms) on mobile devices. Further, the modular design of Trauma renders itself a complementary MFDA detection module for any existing FRS. The main contributions of this work include:

- Generating and annotating a new Trauma dataset with MF and real face images for MFDA detection.
- Exploiting reflective elements of the human face to detect physiological flaws effectively.
- Designing a lightweight, modular, and real-time approach to render a complementary MFDA detection module for edge and mobile FRSs.

The remainder of this paper is organized as follows. Section 2 summarizes the existing MFDA and GAN-face detection methods. Section 3 describes our design of Trauma. Section 4 discusses the experiment setups and results. Section 5 concludes the paper.

2. RELATED WORK

We briefly discuss the existing MFDA and GAN-face detection methods. The authors in [2] used StyleGAN [4] to generate MFDAs resulting in 40 percent of the 5,749 people in the Labeled Faces in the Wild (LFW) dataset [5]. They use three facial recognition models: Dlib, FaceNet, and SphereFace. The authors in [2] presented a comprehensive comparison between the different latent variable evaluation strategies (e.g., (LM-MA-ES) [6], (DE) [7], etc.) for the MFDA generation task. [1] used the StyleGAN face generator and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [8] to generate high-resolution MFDA.

The existing GAN-face detection methods comprise deep learning-based, physical-based, and physiological-based methods [9]. The deep learning-based GAN-face detection methods trained DNN, such as VGG-16 [10], one-shot learning [11], incremental learning [12], and attention-based [13] models, to learn deep hierarchical features and the classifiers jointly in an end-to-end manner to identify fake faces. On the other hand, the physical-based GAN-face detection methods [14, 15] detected GAN-synthesised faces through the inconsistency of the corneal specular highlights between the two synthesized eyes. Furthermore, the physiological-based GAN-face detection methods utilized noticeable artifacts in generated faces, such as asymmetric faces [16], eyes' inconsistent iris color [14], and irregular pupil shapes [17], to spot GAN-faces. However, physical and physiological-based methods cannot generalize against highly realistic MFDA because they only consider single artifacts of eyes, such as iris color, pupil shapes, or similarity of corneal reflections on both eyes. In addition, such artifacts may only sometimes be available due to the limitations of the images with blurriness, low-quality images, or occlusions.

The proposed Trauma system represents the pioneering effort to detect MFDA by leveraging specular reflection highlights. Its uniqueness and efficacy stem from its ability to discern physiological attributes across diverse facial elements (e.g., eyes, nose, cheeks, etc.) through the analysis of specular reflection highlights.

3. ARCHITECTURE

Trauma consists of Training Data Annotation (TDA), Face Specular Highlights Detection (FSHD), and Feature Extraction and Classification (FEC) modules to analyze the semantic aspect of the MFDAs using inconsistencies among specular highlights on various facial parts and physiological flaws of the MF images.

3.1. Training Data Annotation (TDA)

A large-scale benchmark dataset for evaluating GAN-based MFDA detection still needs to be improved. First, we created a

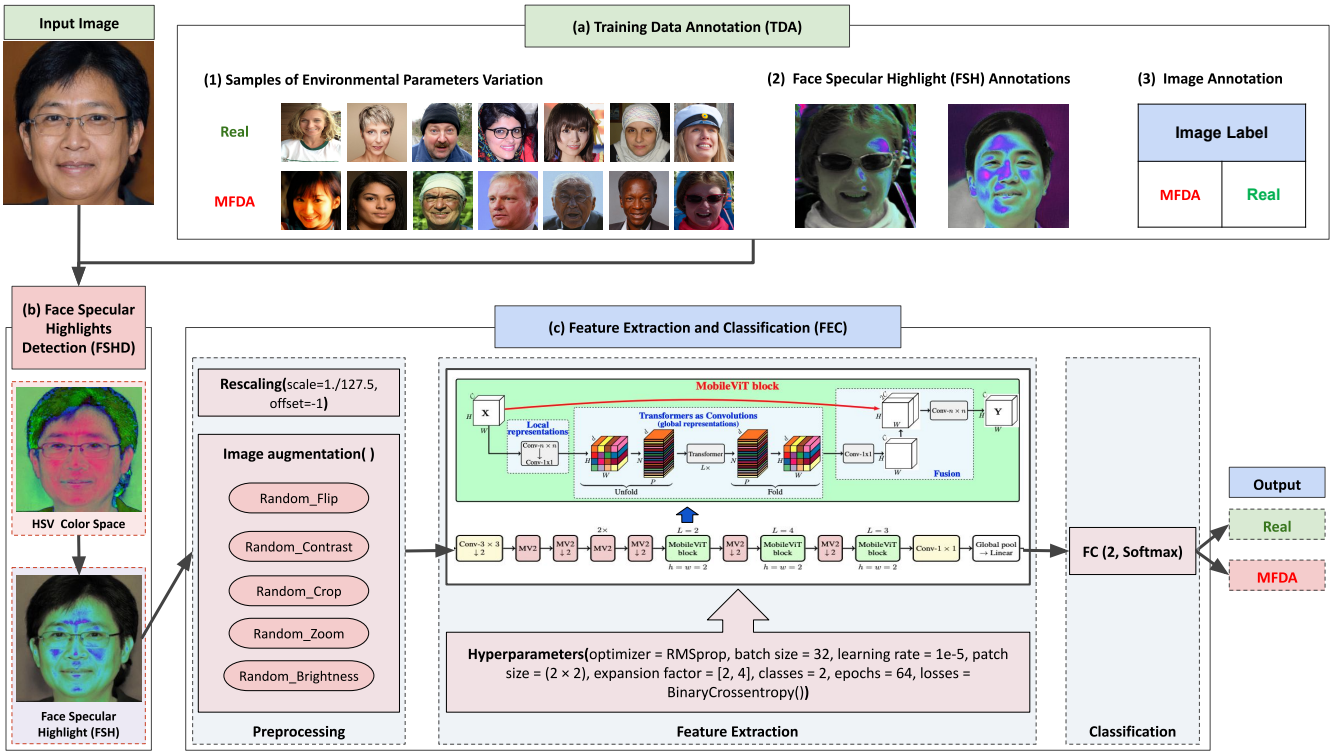

Fig. 2: The Trauma Architecture Block-diagram.

Trauma dataset [18] by collecting and annotating real-face and GAN-based MF images. The Trauma dataset contains 28,620 (29.40 GB) high-resolution facial images. Since MFDA detection is a binary classification problem, we collected 14,310 MF images using various SOTA GAN models, including 6,580 images from StyleGAN [4], 6,580 images from StyleGAN2 [19], and 1,150 images from StyleGAN3 [20]. In addition, we also collected 14,310 real-face images from diverse datasets, including 4,770 images from the FFHQ dataset [4], 4,770 images from the CelebA-HQ dataset [21], and 4,770 from the CelebA dataset [22]. Second, as presented in Figure 2 (a-1), the *TDA* module extracts environmental parameters, including illumination conditions, background colors, indoor or outdoor settings, face pose orientations, age, ethnicity, and appearances (e.g., wearing makeup and accessories) from the Trauma dataset images (1). Then, TDA annotates dataset images in a couple of different types. The *Face Specular Highlight (FSH) annotation* in Figure 2 (a-2) identifies various highlight patterns from the reflective facial regions. The *image annotation* in Figure 2 (a-3) labels the images either Real or MFDA. TDA also resizes all images to the same 256×256 images. TDA applies various augmentations, including horizontal flip, crop, and adjusting brightness and saturation to increase the diversity of the training set.

3.2. Face Specular Highlights Detection (FSHD)

The *FSHD* module in Figure 2 (b) detects the FSH patterns from various face parts by taking a 256×256 RGB image as an input. First, it remaps the primary colors of the input images into HSV (Hue (H), Saturation (S), Value (V)) color space dimensions. H specifies the angle of the color from 0 to 360 degrees, S controls the used color amount from 0 to 100 percent, and V maintains the brightness of the color from 0 (black) to 100. Then, *FSHD* identifies the FSH regions by performing backward conversion from HSV to BGR (RGB, revered) to highlight the brightest point with high and low saturation values.

3.3. Feature Extraction and Classification (FEC)

As illustrated in Figure 2 (c), the *FEC* module conducts image preprocessing, deep hierarchical feature extraction from the *FSH* images, and classification by employing a lightweight ViT-based backbone model [23].

We built an input processing pipeline that standardizes the input images by rescaling their values from the [0, 255] range to the [-1, 1] range. Then, it applies random augmentation transforms during training, including contrast, brightness, horizontal flip, crop, and zoom.

The *FEC* also leverages the strengths of both CNN and ViT architectures by fusing the spatially-localized features captured

Table 1: Classification performance with three different Trauma backbones and datasets.

Backbones	Trauma (StyleGAN): A dataset with 13,160 images (50% real & 50% MF), training 10,000, validation 2,100, and testing 1,060.				Trauma (StyleGAN2): A dataset with 13,160 images (50% real & 50% MF), training 10,000, validation 2,100, and testing 1,060.				Trauma (StyleGAN3): A dataset with 2,300 images (50% real & 50% MF), training 1,600, validation 500, and testing 200.			
	Acc↑	**Loss↓**	**FAR↓**	**FRR↓**	**Acc↑**	**Loss↓**	**FAR↓**	**FRR↓**	**Acc↑**	**Loss↓**	**FAR↓**	**FRR↓**
Trauma (S)	**99.25%**	**0.056**	**0.56%**	**0.94%**	98.87%	0.082	0.37%	1.88%	**99.56%**	**0.010**	**0.37%**	**0.50%**
Trauma (XS)	98.87%	0.071	**0.56%**	1.69%	**99.34%**	**0.030**	0.37%	**0.94%**	99.43%	0.011	0.50%	0.62%
Trauma (XXS)	97.83%	0.116	1.69%	2.64%	99.25%	0.032	**0.18%**	1.32%	98.81%	0.044	0.87%	1.75%

Table 2: Inference time with three different Trauma backbones on CPU and GPU.

Backbones	Size (MB)	Params	Inf. Time (ms)	
			CPU Batch Size (1)	**GPU** Batch Size (32)
Trauma (S)	81.6 MB	7,040,002	10.55 ms	194 ms
Trauma (XS)	32.8 MB	2,774,890	7.56 ms	146 ms
Trauma (XXS)	**16 MB**	**1,306,658**	**3.93 ms**	**126 ms**

by CNNs with the global context awareness capabilities of ViT, ultimately improving the representation learning process from facial images with fewer parameters. The *FEC* architecture is comprised of six blocks. The first block consists of a strode 3×3 standard convolution, followed by one MobileNetV2 (MV2) [24] inverted residual block. The second block contains three inverted residual MV2s for downsampling the resolution of the intermediate feature maps. The 3 to 5 blocks comprise a sequence of one inverted residual MV2 for downsampling and a MobileViT that captures local features through convolutional layers and global elements from the small patches using a transformer block [25] with different lengths. The 6th block comprises a global average pooling (GAP) and fully connected layers. The GAP layer performs downsampling, and the fully connected layer (predication layer) returns a probability distribution with two nodes and a softmax activation function for binary classification. A binary cross-entropy probabilistic loss function is used to compute the cross-entropy loss between actual and predicted labels and to measure the model's accuracy during training and testing. Eventually, it creates a binary classification result (either MF or real).

All images were pre-processed and scaled between -1 and 1. We used the Glorot normal initializer from the Keras library for the default weight initialization. We trained all three models on the GPU environment using the Google Colab Compute Engine (GCE) VM backend with (NVIDIA Tesla-P100-PCIE-16GB) model for 64 iterations with an Adam optimizer, batch size of 32, a learning rate of 1e-5, and patch size of 2×2 for the transformer blocks. In MV2s, we used an expansion factor of 4 for Trauma (S) and Trauma (XS), except for Trauma (XXS), we used an expansion factor of 2.

4. EVALUATIONS

We conducted *inference time* and *classification performance* tests on Trauma modules trained on three different backbone architectures (Trauma (S), Trauma (XS), and Trauma (XXS) in Table 2) with three distinct datasets (Trauma (StyleGAN), Trauma (StyleGAN2), and Trauma (StyleGAN3) in Table 1). Our evaluation study aimed to determine whether particular combinations of backbones and datasets produce better inference time results on resource-constrained devices and to evaluate the effectiveness of the Trauma modules in classifying images. Our experimental setup measured the *inference time* of different Trauma backbones (size and parameters) with CPU and GPU environments. A batch of 32 images was used for GPU, while one image per batch is used for an 8-core CPU. We evaluated its *classification performance* on predefined image classes, including binary cross-entropy loss function, a dense layer of two nodes, and Softmax activation at the top of every network.

The *inference time* tests results are presented in Table 2. Trauma (XXS), with 1.3 M parameters and 16 MB size, is the fastest network (within 4 ms) across all devices. On the other hand, Trauma (S), with 7 M parameters and 81.6 MB size, is the slowest. All models can evaluate within 200 ms for the typical batch size of 32 images with GPU and within 11 m for a single batch with CPU. The results have important implications for developing and deploying lightweight Trauma modules to detect MFDA in practical edge and mobile environments.

The outcome of *classification performance* tests, including the classification accuracy, loss, false acceptance rate (FAR), and false rejection rate (FRR), are summarized in Table 1. We observed that the Trauma (S), the largest backbone, consis-

tently outperformed the XS and XXS backbones in all metrics on Trauma (StyleGAN) and Trauma (StyleGAN3) datasets. However, the smaller backbones, Trauma (XS) or Trauma (XXS), result in better performance on Trauma (StyleGAN2) dataset. Overall, our findings indicate that the *classification performance* is very effective (e.g., higher than 97.83 % accuracy) regardless of the choice of backbone and dataset.

5. CONCLUSIONS

We proposed a novel, trustworthy authentication against visual MFDA (Trauma) that generates and annotates a new dataset with MFDA and real images to detect MFDA. It uses reflective elements to extract characteristics for MFDA detection that enable the detection of noticeable physiological flaws. We designed it as a lightweight, modular, real-time system to render a complementary MFDA detection module for edge and mobile FRSs. The empirical results show that Trauma achieves detection accuracy ranging from 97.83% to 99.56% and rapid detection speed (less than 11 ms) against the current SOTA MFDAs.

6. REFERENCES

[1] Huy H Nguyen, Junichi Yamagishi, Isao Echizen, and Sébastien Marcel, "Generating master faces for use in performing wolf attacks on face recognition systems," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[2] Ron Shmelkin, Tomer Friedlander, and Lior Wolf, "Generating master faces for dictionary attacks with a network-assisted latent space evolution," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–08.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[5] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[6] Ilya Loshchilov, Tobias Glasmachers, and Hans-Georg Beyer, "Limited-memory matrix adaptation for large scale black-box optimization," *arXiv preprint arXiv:1705.06693*, 2017.

[7] Rainer Storn and Kenneth Price, "Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341, 1997.

[8] Nikolaus Hansen and Andreas Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary computation*, vol. 9, no. 2, pp. 159–195, 2001.

[9] Xin Wang, Hui Guo, Shu Hu, Ming-Ching Chang, and Siwei Lyu, "Gan-generated faces detection: A survey and new perspectives," *arXiv preprint arXiv:2202.07145*, 2022.

[10] Nhu-Tai Do, In-Seop Na, and Soo-Hyung Kim, "Forensics face detection from gans using convolutional neural network," *ISITC*, vol. 2018, pp. 376–379, 2018.

[11] Hadi Mansourifar and Weidong Shi, "One-shot gan generated fake face detection," *arXiv preprint arXiv:2003.12244*, 2020.

[12] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva, "Incremental learning for the detection and classification of gan-generated images," in *2019 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2019, pp. 1–6.

[13] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, "Robust attentive deep neural network for detecting gan-generated faces," *IEEE Access*, vol. 10, pp. 32574–32583, 2022.

[14] Falko Matern, Christian Riess, and Marc Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2019, pp. 83–92.

[15] Shu Hu, Yuezun Li, and Siwei Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2500–2504.

[16] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu, "Exposing gan-synthesized faces using landmark locations," in *Proceedings of the ACM workshop on information hiding and multimedia security*, 2019, pp. 113–118.

[17] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu, "Eyes tell all: Irregular pupil shapes reveal gan-generated faces," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2904–2908.

[18] "Master face dictionary attacks via reflection-based identification (trauma) dataset," `https://github.com/READFake/DARI-MFDA-Detection-Dataset`, 03 2023, (Accessed on 03/01/2023).

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[20] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Alias-free generative adversarial networks," in *Proc. NeurIPS*, 2021.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, vol. abs/1710.10196, 2017.

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, no. 2018, pp. 11, 2018.

[23] Sachin Mehta and Mohammad Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations*, 2021.

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.