Harnessing Foundation Models for Image Anonymization

Luca Piano
Politecnico di Torino
Turin, Italy
luca.piano@polito.it

Pietro Basci
Politecnico di Torino
Turin, Italy
pietro.basci@polito.it

Fabrizio Lamberti
Politecnico di Torino
Turin, Italy
fabrizio.lamberti@polito.it

Lia Morra
Politecnico di Torino
Turin, Italy
lia.morra@polito.it

***Abstract*—Traditional deep learning pipelines involve multiple intricate steps, from data acquisition to model training, fine-tuning, and deployment. However, recent advancements in foundation models, particularly in text-to-image generation, offer a paradigm shift in addressing tasks without the need for these conventional processes. In this paper, we explore how foundation models can be leveraged to solve tasks, specifically focusing on anonymization, without the requirement for training or fine-tuning. By bypassing traditional pipelines, we demonstrate the efficiency and effectiveness of this approach in achieving anonymization objectives directly from the foundation model's inherent knowledge. Our findings underscore the transformative potential of foundation models in simplifying and accelerating deep learning tasks, paving the way for novel applications in various domains.**

I. Introduction

Deep learning (DL) has become a powerful paradigm for handling tasks in many fields, from image recognition to natural language processing. Solving tasks with DL generally is not a straightforward path, which includes data acquisition, preprocessing, model selection or design, training, fine-tuning, and deployment. Often, following these steps requires significant computational resources, time, and expertise to implement and optimize. Recent breakthroughs in foundation models, particularly those leveraging transformer architectures, have reshaped the landscape of deep learning. These models, trained on massive datasets and equipped with powerful language understanding capabilities, have demonstrated remarkable performance across a wide range of tasks, including text generation, image synthesis, and language translation. An interesting potential emerges in this context: can we solve a task with low resource commitment? By exploiting the latent knowledge embedded within these models, it may be feasible to directly address specific tasks, such as anonymization, without the need for explicit training or fine-tuning. This work explores this novel paradigm, focusing specifically on the anonymization task, and investigates the potential of leveraging foundation models to achieve this objective without traditional training pipelines. Image anonymization involves hiding or altering identifiable features within an image to protect the privacy and identity of the individuals depicted. This process aims to make it difficult or impossible to recognize or identify individuals while retaining other non-sensitive information present in the image. Various techniques can be employed for image anonymization, including blurring, pixelation, masking, or using advanced methods like Generative Adversarial Networks (GANs). We examine how foundation models can generate anonymized images directly from textual descriptions, thereby circumventing the complexities associated with data preprocessing, model training, and fine-tuning. Through empirical evaluation and analysis, we demonstrate the effectiveness and efficiency of this approach in achieving anonymization goals while highlighting its implications for simplifying DL workflows.

The rest of this paper is structured as follows: in Section 2, we offer a comprehensive exploration of foundation models and their capabilities, with a specific focus on text-to-image generation, as well as an examination of the current state of the art in anonymization techniques. Section 3 delves into the proposed methodology for tackling the anonymization task without the need for training. Here, we elucidate the steps involved in our approach and provide insight into the underlying rationale. In Section 4, we present our experimental results and perform a detailed analysis, highlighting the performance of our approach in anonymization tasks and its comparative efficacy. Lastly, Section 5 discusses the broader implications of our findings, potential applications across various domains, and outlines avenues for future research.

II. Related Works

A. Foundation models

In recent years, the exploration of foundation models has fascinated researchers in various domains, ranging from natural language processing to image generation. This burgeoning field has seen significant advancements, with scholars delving into the capabilities, limitations, and ethical implications of these models. Pioneering research by [1] laid the groundwork for understanding the scaling of language models, showcasing that larger models yield superior performance in text generation tasks. Their work underscored the crucial role of scale in achieving state-of-the-art results. Bommasani et al.[2] offered insights into the factors driving the scalability of foundation models, highlighting advances in computer hardware, model architectures (such as the Transformer), and the abundance of training data. Their analysis shed light on the technical intricacies facilitating the success of large-scale AI models. Meanwhile, [3] delved into novel AI model architectures,

particularly diffusion models, which have revolutionized image generation tasks. Diffusion models (e.g., DALL-E [4] and Stable Diffusion [5]) showcased a remarkable ability to produce high-quality, diverse images based on textual prompts, thereby opening avenues for creative AI applications.

B. Image anonymization

In the realm of image anonymization, traditional techniques such as pixelation, blurring, and masking have long been utilized [6], [7], [8]. However, these methods often introduce distortions that may compromise the utility of anonymized images for downstream applications, particularly those reliant on facial recognition or attribute detection. Moreover, they are susceptible to attacks aimed at reversing the anonymization process [9], [10].

The advent of Generative Adversarial Networks (GANs) has catalyzed a paradigm shift in image anonymization, particularly through the lens of conditional generative models [11], [12], [13], [14], [6], [8], [7], [15], [16], [17], [18]. These approaches strive to anonymize images while preserving salient features, rendering the anonymized data suitable for a broader spectrum of applications.

For instance, DeepPrivacy [11] utilizes a GAN conditioned on pose and background but faces challenges with irregular poses and complex backgrounds, potentially leading to distortions. DeepPrivacy2 [12] extends this to full-body anonymization, but still encounters limitations in identity generation and complete anonymization assurance.

CIAGAN [13] tackles anonymization using an inpainting GAN conditioned on various inputs, including facial landmarks and desired identities. However, its efficacy depends on the accuracy of landmark detection, and the use of real identities poses privacy concerns.

A further advance is represented by FALCO [14], which aims to preserve facial features during anonymization by optimizing image representations in the latent space of a pretrained StyleGAN2 [19] model.

With the use of diffusion models, CAMOUFLaGE [20] overcomes the limitations seen in earlier techniques. It presents two different versions: one uses a caption generator, pretrained ControlNets [21], and a new negative controller network, and the other trains two adapters [22], [23] to avoid captions. Our approach offers a simplified version of CAMOUFLaGE, bypassing the need for training. It generates facial reconstructions that retain facial attributes while sacrificing background details, akin to FALCO.

III. Methodology

In our pursuit to tackle the anonymization task while leveraging the inherent capabilities and knowledge of foundational models (Figure 1), we avoid complex pipelines, opting instead to explore a more streamlined approach. To compare our methodology with FALCO, our focus remains on anonymizing facial features without necessarily preserving background details while striving to retain as many facial attributes as possible. For each input image, we extract age, ethnicity, and

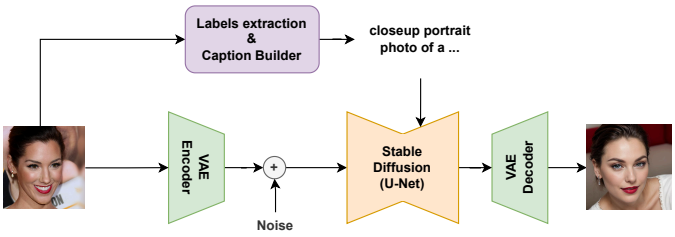

Fig. 1. Our architecture. The caption is a combination of the extracted labels.

the 40 labels defined in CelebA-HQ [24], covering various facial attributes. These labels serve as pivotal anchors in our anonymization process. Using this rich information, we craft captions to guide the model (Algorithm 1), guiding it toward synthesizing portraits that adhere closely to the attributes provided. By exploiting facial feature information, our approach strives to generate images with fidelity to the original while obfuscating individual identities effectively.

Algorithm 1: Prompt generation

Input: I; `# Input image`
Output: $P, \neg P$; `# Positive and Negative prompt`
Function `generatePrompt` $(image)$ **:**
  $face \leftarrow getBiggerFace(image)$;
  $attrs \leftarrow getAttrs(image, face)$;
  $ethnic \leftarrow attrs.pop(\text{“}Ethnic\text{”})$;
  $age \leftarrow attrs.pop(\text{“}Age\text{”})$;
  $isMale \leftarrow attrs.pop(\text{“}Male\text{”}) > 0.5$;
  $P \leftarrow$ “closeup portrait photo of a ”;
  $P \leftarrow P + ethnic +$ “ ”;
  $P \leftarrow P + age +$ “ ”;
  if $isMale$ **then**
    $P \leftarrow P+$ “man, ”;
  else
    $P \leftarrow P+$ “woman, ”;
  end
  $\neg P \leftarrow$ “”;
  forall k *in* $attrs$ **do**
    $x \leftarrow attrs[k]$;
    if $x > 0.5$ **then**
      $P \leftarrow P + k +$ “, ”;
    else
      $\neg P \leftarrow \neg P + k +$ “, ”;
    end
  end
  return $P, \neg P$

The architecture is illustrated in Figure 1. Two models were employed for information extraction: FACER [25], used to identify the 40 CelebA-HQ attributes, and DeepFace [26], used to determine ethnicity and age. The composed prompts are then used to guide the generation process. Classifier-free guidance [27] was emplyoed to push the image content

in the direction of the positive prompt P and far from the negative prompt $\neg P$. The generative model used is Realistic Vision v5.1, which is derived from Stable Diffusion v1.5 [28]. The Variational Autoencoder (VAE) [29] is the VAE ft-MSE [30], which helps against artifacts on smaller faces. The scheduler adopted is the DPM++ SDE Karras [31]. Images were generated at resolution of 512^2 pixels using 15 steps.

IV. Results

Following the evaluation methodology outlined by [20], we evaluated the results through anonymization, downstream task performance, and objective visual quality of the results. All analyzes were conducted on a CelebA-HQ subset consisting of 1,000 randomly sampled examples.

The risk of re-identification is evaluated using face-level and image-level protocols. Face-level mimics Facial Recognition system, extracting face crops with MTCNN [32] and generating embeddings using FaceNet on VGGFace2 [33] or CASIA WebFace [34]. Image-level protocol, resembling web-based image retrieval, uses CLIP visual encoder for embeddings. We evaluate re-identification rate at different ranks (Re-ID@K)for K=1, 5, and 10 mean Average Precision (mAP). The results show the effectiveness of the method, reaching a performance comparable to state-of-the-art methods when FaceNet models are used, and surpassing them when using CLIP Table I.

We then measured the impact of anonymization on downstream tasks by comparing the performance of specific tasks on real and anonymized images. Specifically, we focus on six different downstream tasks: classification of 21 inner features of the face, 17 outer features of the face, classification of emotions, ethnicity, gender, and Valence-Arousal. The inner and outer features were defined as in FALCO [14]. The results are interesting (Table II): while performance is similar in most tasks, our method outperforms all other methods in ethnicity classification with an improvement of $\sim 25\%$.

An essential aspect when evaluating synthetic image generation is the quality of the resulting images. Figure 2 showcases examples of images generated by various methods. Inpainting-based methods, such as DP2, are more prone to artifacts, especially in the presence of unusual poses, hairstyles or accessories. On the othe rhand, FALCO changes the overall appearance with washed-out colors.

Visual DNA, a novel technique introduced by Ramtoula et al. [35], offers a sophisticated approach to comparing individual images and datasets. It analyzes the distributions of neuron activations across layers of a pre-trained feature extractor, specifically the Mugs-ViT-B/16 model in our case. By leveraging the Earth Mover's Distance (EMD), Visual DNA computes the semantic distance between original and synthetic images. At the image level, we calculate the distance for each anonymized-real image pair and then derive the mean and standard deviation for the entire anonymized dataset. Additionally, we employ the Fréchet Inception Distance (FID) [36] to gauge the overall quality of synthetic datasets. Notably, Visual DNA has demonstrated superior performance in providing reliable per-dataset measures, even with fewer samples compared to conventional metrics like FID. The proposed method, as shown in Table III, achieves slightly higher values with respect to competitors despite neither relevant artifacts nor other kinds of disturbance are introduced in the synthetic image. We argue that the increase of FID is partly due to the higher visual quality of the synthetic images which appear sharpened with corrections introduced by the model especially in the background (e.g., blurring reduction). In fact our model, which is based on Stable Diffusion, is not specifically trained on CelebA-HQ. Hence, it may generate images that are somewhat visually different from the target dataset justifying the distance among the synthetic and the real dataset. At the same time, excessively beautifying the original images may pose issues in terms of representativeness of the general population.

The higher values for Visual DNA metrics, which quantify the semantic distance between real and synthetic images, result from the most extensive mutations introduced by the method: conditioning the generation process using only text allows for a higher diversity while preserving the aspects specified in the prompt, as confirmed by the results on downstream tasks.

V. Conclusion

In conclusion, this paper introduces a novel approach to anonymization leveraging foundation models, specifically focusing on the task of anonymizing facial features in images. By harnessing the latent knowledge embedded within these models and crafting descriptive prompts, we demonstrate the feasibility of directly generating anonymized images from textual descriptions, circumventing the need for explicit training pipelines. We employ a streamlined approach that relies on information extraction from input images using state-of-the-art models such as FACER and DeepFace. By crafting captions that guide the model towards synthesizing portraits adhering closely to provided attributes, we aim to generate anonymized images that retain fidelity to the original while effectively obfuscating individual identities. Evaluation of our approach demonstrates promising results across multiple metrics. We achieve comparable or superior performance to state-of-the-art methods in anonymization effectiveness, downstream task performance, and visual quality of the generated images. In particular, our method outperforms others in ethnicity classification, showcasing its efficacy in preserving crucial demographic information while anonymizing images. While our approach exhibits slight increases in metrics such as FID and Visual DNA, we attribute these to the higher visual quality and semantic distance introduced intentionally during the generation process. The diversity of generated images, coupled with the preservation of specified attributes, underscores the versatility and adaptability of our method across various applications. Our study highlights the potential of leveraging foundation models for low-resource anonymization tasks, offering a simplified yet effective alternative to traditional DL workflows. By embracing this novel paradigm, we pave the way for more streamlined and efficient approaches to addressing privacy concerns in image data while preserving essential attributes for downstream tasks. Our work concentrates on anonymization as

TABLE I
RE-IDENTIFICATION RATE AND M-AP FOR CELEBA-HQ [24]

Encoding	CLIP ViT-B/32				FaceNet-vggface2				FaceNet-casia-webface			
Method	**Re-ID@1**	**Re-ID@5**	**Re-ID@10**	**mAP@50**	**Re-ID@1**	**Re-ID@5**	**Re-ID@10**	**mAP@50**	**Re-ID@1**	**Re-ID@5**	**Re-ID@10**	**mAP@50**
DeepPrivacy2[12]	0.268	0.404	0.469	0.113	0.004	0.018	0.028	0.006	**0.003**	0.020	0.036	0.005
FALCO[14]	0.101	0.222	0.292	0.054	**0.002**	**0.014**	**0.020**	**0.003**	0.005	**0.013**	**0.022**	**0.003**
Ours	**0.007**	**0.021**	**0.037**	**0.006**	0.010	0.028	0.037	0.006	0.007	0.014	0.029	0.005

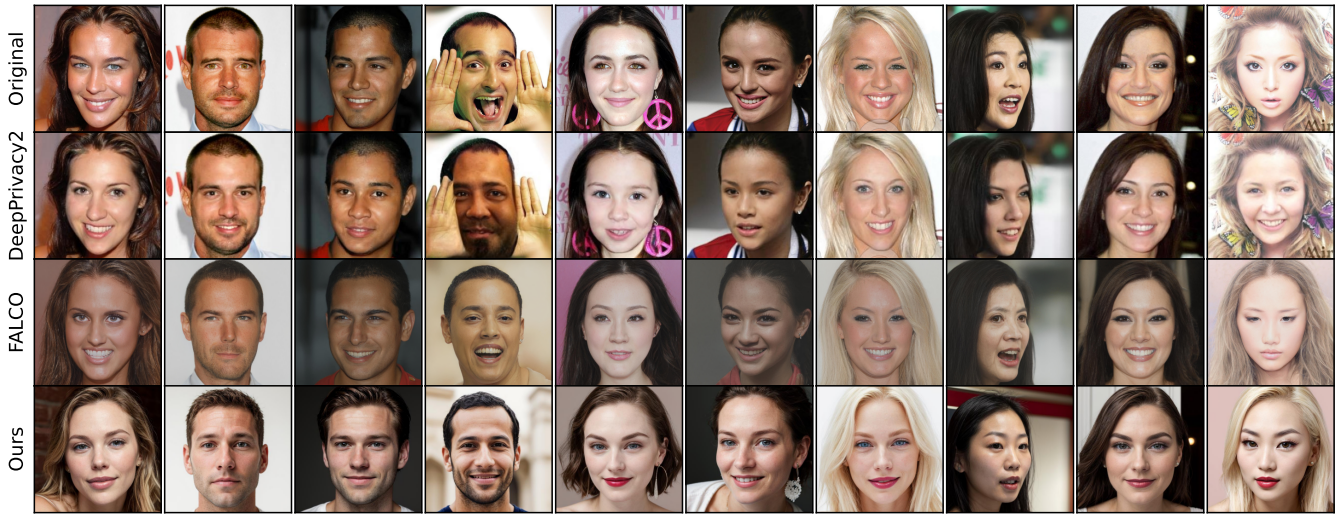

Fig. 2. Comparison of the proposed method to DeepPrivacy2[12] and FALCO[14] in terms of quality on CelebA-HQ. (Randomly sampled images)

Downstream task	**Real → Anonymized**					
Method	**Inner face attrs** (AUC ↑)	**Outer face attrs** (AUC ↑)	**Emotions** (AUC ↑)	**Valence-Arousal** (MSE ↓)	**Ethnicity** (Accuracy ↑)	**Gender** (Accuracy ↑)
DeepPrivacy2[12]	0.823	**0.976**	0.664	0.18	0.674	0.925
FALCO[14]	**0.898**	0.944	**0.777**	**0.08**	0.654	0.950
Ours	0.878	0.954	0.730	0.09	**0.928**	**0.965**

TABLE II
PERFORMANCE ON DOWNSTREAM TASKS EVALUATED ON CELEBA-HQ.

Method	**FID ↓**	**Visual DNA ↓** (dataset level)	**Visual DNA ↓** (image level)
DeepPrivacy2[12]	49.4	**5.0**	**10.4±1.3**
FALCO[14]	**41.2**	6.3	15.2±2.1
Ours	53.3	7.1	18.3±2.1

TABLE III
IMAGE QUALITY RESULTS ON CELEBA-HQ.

a case study, however, the our methodology holds potential for application in diverse domains beyond anonymization.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme (Grant Agreement No 951511 - AI4Media).

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022.

[4] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[6] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 1–10.

[7] S. Tansuriyavong and S.-i. Hanaki, "Privacy protection by concealing persons in circumstantial video image," in *Proceedings of the 2001 workshop on Perceptive user interfaces*, 2001, pp. 1–4.

[8] D. Chen, Y. Chang, R. Yan, and J. Yang, "Tools for protecting the privacy of specific individuals in video," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, 2007.

[9] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-based face de-identification," in *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*. IEEE, 2006, pp. 161–161.

[10] C. Neustaedter, S. Greenberg, and M. Boyle, "Blur filtration fails to preserve privacy for home-based video conferencing," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 13, no. 1, pp. 1–36, 2006.

[11] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *International symposium on visual computing*. Springer, 2019, pp. 565–578.

[12] H. Hukkelås and F. Lindseth, "Deepprivacy2: Towards realistic full-body anonymization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1329–1338.

[13] M. Maximov, I. Elezi, and L. Leal-Taixé, "Ciagan: Conditional identity anonymization generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5447–5456.

[14] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8001–8010.

[15] Y. Wen, B. Liu, M. Ding, R. Xie, and L. Song, "Identitydp: Differential private identification protection for face images," *Neurocomputing*, vol. 501, pp. 197–211, 2022.

[16] Y. Wu, F. Yang, and H. Ling, "Privacy-protective-gan for face de-identification," *arXiv preprint arXiv:1806.08906*, 2018.

[17] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1589–1604.

[18] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[20] L. M. Luca Piano Pietro Basci, Fabrizio Lamberti, "Latent diffusion models for attribute-preserving image anonymization," *Come to appear in arxiv*, 2024.

[21] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[22] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023.

[23] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[25] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 697–18 709.

[26] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICEET53442.2021.9659697

[27] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[28] "Stable diffusion v1.5," https://huggingface.co/runwayml/stable-diffusion-v1-5, accessed: 2023-17-11.

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[30] "Stable diffusion vae finetuned mse," https://huggingface.co/stabilityai/sd-vae-ft-mse, accessed: 2023-17-11.

[31] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 565–26 577, 2022.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, pp. 1499–1503, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:10585115

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206592766

[34] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," *ArXiv*, vol. abs/1411.7923, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:17188384

[35] B. Ramtoula, M. Gadd, P. Newman, and D. De Martini, "Visual dna: Representing and comparing images using distributions of neuron activations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 113–11 123.

[36] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017.