ToonerGAN: Reinforcing GANs for Obfuscating Automated Facial Indexing

Kartik Thakral, Shashikant Prasad, Stuti Aswani, Mayank Vatsa, and Richa Singh
IIT Jodhpur, India
{thakral.1, prasad.5, aswani.1, mvatsa, richa}@iitj.ac.in

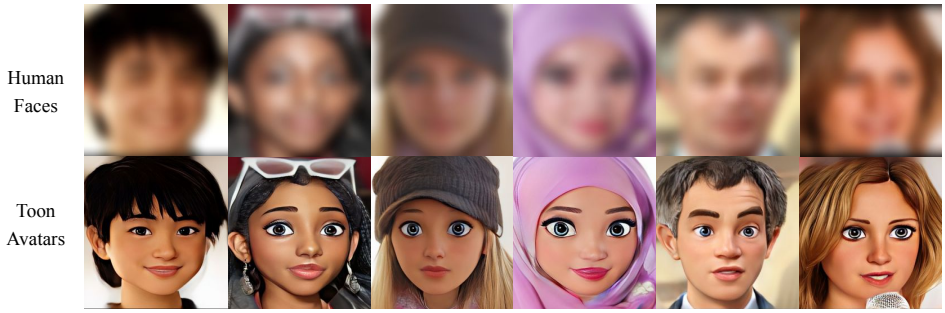

Figure 1. Introducing ToonerGAN, a novel method for generating high-resolution (1024×1024) toon avatars that effectively obfuscate identity details. This method achieves a balance between identity obfuscation and recognizability by human observers. The original facial images used are sourced from the publicly available FFHQ dataset and have been blurred in this display to maintain anonymity.

Abstract

The rapid evolution of automatic facial indexing technologies increases the risk of compromising personal and sensitive information. To mitigate the issue, we propose creating cartoon avatars, or 'toon avatars', designed to effectively obscure identity features. The primary objective is to deceive current AI systems, preventing them from accurately identifying individuals while making minimal modifications to their facial features. Moreover, we aim to ensure that a human observer can still recognize the person depicted in these altered avatar images. To achieve this, we introduce 'ToonerGAN', a novel approach that utilizes Generative Adversarial Networks (GANs) to craft personalized cartoon avatars. The ToonerGAN framework consists of a style and a de-identification module that work together to produce high-resolution, realistic cartoon images. For the efficient training of our network, we have developed 'ToonSet' dataset, consisting of around 23,000 facial images and their cartoon renditions. Through comprehensive experiments and benchmarking against existing datasets, including CelebA-HQ, our method demonstrates superior performance in obfuscating identity while preserving the utility of data. Additionally, a user-centric study exploring the effectiveness of ToonerGAN has yielded compelling observations.

1. Introduction

Face recognition technology, extensively utilized in biometrics, security, and social media applications [4, 16, 22], offers multifold advantages. However, it also has the potential of facial images being misused for unauthorized access [3], leading to threats like sensitive information extraction and susceptibility to harmful exploitation [28]. Therefore, there is a critical need to develop algorithms capable of effectively de-identifying facial identity information. Traditional de-identification methods often rely on generative models to produce altered versions of a person's image [5, 21]. These methods, however, tend to alter facial characteristics so drastically that the person is no longer recognizable even to human observers, thus limiting their usability in many human-centric applications. An optimal technique would retain the facial structure familiar to humans while effectively obfuscating unique features that enable automatic recognition.

In this research, we introduce "ToonerGAN," a novel GAN approach for creating de-identified cartoon face images. ToonerGAN is distinctive in its ability to retain a realistic appearance to human observers while simultaneously evading facial recognition algorithms. Unlike traditional methods, ToonerGAN shifts the distribution of facial images towards a cartoon-like representation, using an encoder-decoder-based generator with a three-step training process: Style Learning, Style Distillation, and De-

identification. We utilize a StyleGAN2-based decoder, training it with a combination of adversarial, reconstruction, and perceptual losses, along with an identity loss to ensure effective de-identification. This paper also introduces the "ToonSet" dataset, which consists of 23,000 face-to-cartoon image pairs. These pairs are derived from the FFHQ dataset and have been stylized into cartoon avatars to train the ToonerGAN. Comprehensive assessment, including comparative analyses with existing algorithms using the CelebA-HQ and LFW datasets, highlights the efficacy of ToonerGAN. Additionally, we conduct a study involving human participants to understand the practicality of the de-identified images from user viewpoint. To the best of our knowledge, this is the first work to employ cartoon images for face de-identification, maintaining a balance between automated indexing and usability.

2. Related Work

In this section, we present literature of both de-identification/anonymization and face portrait style transfer as these forms the basis of the proposed framework.

Face De-identification/Anonymization The face de-identification/anonymization techniques range from attribute anonymization [3, 28] to identity anonymization [21, 34]. Classic face anonymization techniques aim to preserve identity by standard image processing methods such as Gaussian blurring, and pixelation. These techniques successfully preserve face identity but also introduce artifacts, eliminating facial features that human observers and machine learning algorithms can no longer detect. Most of the recent anonymization techniques focus on adversarial training [17, 21, 34] to de-identify face identity while maintaining data utility. CIAGAN [21] utilizes facial landmark and masked-based representation for a conditional GAN to replace the face image with the desired identity. DeIDGAN [17] on the other hand, first starts with anonymizing the input image followed by replacing it with synthesized image generated by a de-identification generator. A^{3}GAN [34] also first targets to anonymize the face identity with a semantic suppression module and then generate a synthesized image by injection attribute information through the AINet module. While all these techniques somehow promote face de-identification/anonymization through face replacements in adversarial training, techniques like LIVE [5] use a simple encoder-decoder architecture conditioned by existing face recognition methods. Recently, RiDDLE [18] proposed to anonymize images with StyleGAN2 [15] along with the encryption-decryption process of identity in latent space from a given password. This method promotes the emerging field of reversibility of de-identified images given a correct password. One recent adversarial training de-identification method, DartBlur [11] promotes blur-based preservation of identity information while suppressing detection artifacts for review convenience.

Portrait Style Transfer Most of the face artistic portrait style transfer techniques involve image-to-image translation-based GAN architectures. Toonify [23] achieved this task by fine-tuning the pre-trained StyleGAN architecture [14] with limited paired data. AnimeGAN2 [2] introduced a novel method for the training of style transfer models by generating synthetic face-toon pairs. This approach simplifies the training process by creating artificial data pairs for distillation, alleviating the dependence on paired data. On the other hand, CariGAN [10] addressed the challenge of weakly supervised tasks by aligning facial and caricature identities without pixel-level correspondences. Recent advancements of feature-disentanglement in the latent space of generative models like StyleGAN [14] have also impacted the domain. For instance, DualStyleGAN [30] exploits this property for style transfer by introducing an extrinsic style path for explicit style control while retaining the intrinsic style path of StyleGAN.

3. Obfuscating Faces via Toonification

This research introduces ToonerGAN, a novel algorithm trained on the **ToonSet** dataset, aimed at generating de-identified cartoon images from real faces to obscure automated facial recognition while maintaining human recognizability. ToonerGAN utilizes paired images from ToonSet for training, with high-resolution versions serving as references, generating high-quality toon images that minimally alter facial attributes and diverge from the original face distribution. Additionally, integration of a face parsing map, generated by BiSeNet [32], enhances the identification of facial attributes during training.

ToonSet is a face-toon paired dataset. The dataset consists of a total of 60,000 original face images and their corresponding toon avatar images. We use the FFHQ dataset [14] and borrow 23,000 face images. These images are then transformed into toon avatars using `toonme.com`. The website takes over 10 seconds to generate a toon avatar. Each face image is of size 512×512; however, the resultant toon image from the website is of size 498×512. For training, all images are resized to 256×256. Each face image was paired with its toon avatar and can be utilized to train models. We report a value of 0.685 and 17.01 of SSIM and PSNR, respectively, calculated for the proposed ToonSet. The dataset can be visualized in Figure 3. The dataset is publicly available to the research community and can be utilized for different tasks computer vision tasks[1].

Face-to-Toon Distribution Mapping: The goal of ToonerGAN is to achieve face de-identification through toonification. To achieve this, we use a paired dataset $U = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where $X =$

[1]The dataset can be accessed here: `https://rb.gy/bl9zz4`

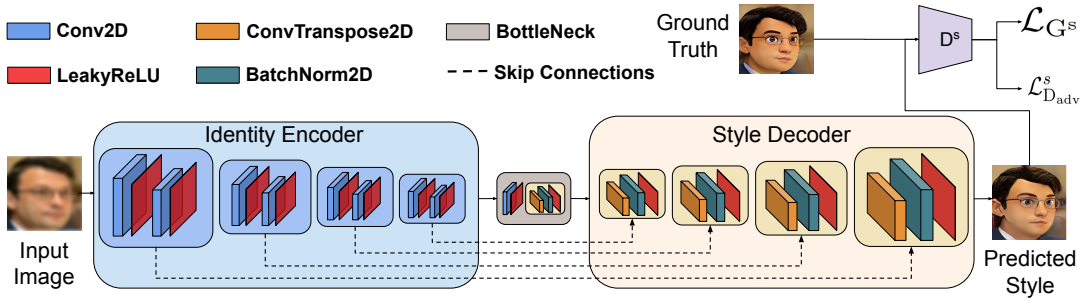

Figure 2. Architecture of the style learning step (i.e. Step 1) in the proposed approach.

Figure 3. ToonSet: Sample face images from FFHQ dataset [14] and their toon avatars. The face images blurred for display purposes here to maintain anonymity.

$\{x_1, x_2, ..., x_n\}$ represents real-face images and $Y = \{y_1, y_2, ..., y_n\}$ represents the set of corresponding toon images. The objective is to generate $\bar{Y}^a = \{\bar{y}_1^a, \bar{y}_2^a, ...\bar{y}_n^a\}$ where $\bar{y}_i^a$ is the de-identified high-resolution (1024×1024) toon image. For this, we introduce ToonerGAN, which consists of a generator $G(.;\omega)$ with weights $\omega = \{\theta_e, \phi_d^s, \phi_d^a\}$. The generator is composed of three components: an Identity Encoder $G(.;\theta_e)$ with encoder weights θ_e, a Style Decoder $G(.;\phi_d^s)$ with decoder weights ϕ_d^s, and a De-identification Decoder $G(.;\phi_d^a)$ with decoder weights ϕ_d^a. For generating outputs, the proposed architecture undergoes a progressive three-step training process. It begins with Style Learning followed by Style Distillation. In the third step, the system applies this style comprehension in the de-identification phase, employing face toonification and de-identification techniques to transform the facial data.

Step 1: Style Learning: In the first step, we focus on generating $\bar{Y}^s = \{\bar{y}_1^s, \bar{y}_2^s, ...\bar{y}_n^s\}$ where $\bar{y}_i^s$ is the output toon image generated using a paired face-toon dataset. For this, we employ the Identity Encoder $G(.;\theta_e)$ and the Style Decoder $G(.;\phi_d^s)$ to comprehend the toon-style. This stage is formulated to ensure that the knowledge acquired during training at this stage is effectively utilized later to generate high-resolution toon images. To extract style-specific features, we perform an image-to-image translation training regime adopted from [9]. However, we introduce modifications in the $G(.;\theta_e)$ of $G(.;\omega)$, which are also illustrated in Figure 2. Skip-connections are established between the $G(.;\theta_e)$ and $G(.;\phi_d^s)$ convolutional blocks to facilitate a smoother information flow. We begin the training process of learning the style and generating the toons by adversarially deceiving the discriminator $D(.;\psi^s)$ while maintaining the right pixel prediction as described below:

$$\mathcal{L}_{\mathrm{G^s}}(G(x;\theta_e,\phi_d^s)) = \lambda_{\mathrm{Gen}}\mathcal{L}^s_{\mathrm{G_{adv}}} + \lambda_{\mathrm{MSE}}\mathbb{E}^{y\in Y}_{\bar{y}^s\in\bar{Y}^s}\|\bar{y}^s - y\|_2^2 \tag{1}$$

where, $\mathcal{L}^s_{\mathrm{G_{adv}}} = \mathbb{E}_{x\in X}[\log(1 - D(G(x;\theta_e,\phi_d^s);\psi^s))]$ is the adversarial loss for $G(.;\theta_e,\phi_d^s)$ and the second term is the $\mathcal{L}_2$ loss on the generated style image with ground truth. For training $D(.;\psi^s)$ to distinguish between real and generated toon images, $\mathcal{L}^s_{\mathrm{D_{adv}}}(D(\psi^s)) = \mathbb{E}_{y\in Y}[\log D(y;\psi^s)] + \mathbb{E}_{x\in X}[\log(1 - D(G(x;\theta_e,\phi_d^s);\psi^s))]$ is the adversarial loss employed for the descriminator. Finally, the network $G(.;\theta_e,\phi_d^s)$ and $D(.;\psi^s)$ is trained from scratch with the objective function:

$$\min_{\theta_e,\phi_d^s}\max_{\psi^s}\mathcal{L}_{\mathrm{G^s}} + \mathcal{L}^s_{\mathrm{D_{adv}}} \tag{2}$$

Step 2: Style Distillation: In this step, for the input set X, our focus revolves around generating $\bar{Y}^a = \{\bar{y}_1^a, \bar{y}_2^a, ...\bar{y}_n^a\}$ where $\bar{y}_i^a$ is the de-identified high-resolution (1024×1024) toon image. To accomplish this task, we utilize a decoder capable of style distillation and generating detailed high-resolution images. For this, we leverage a pre-trained StyleGAN2 architecture [15] as it can be translated to new styles and characterizes both styles, original and transferred. It is employed to serve as the decoder $G(.;\phi_d^a)$ in our proposed network. For a face image x, the toon output $\bar{y}^s = G(x;\theta_e,\phi_d^s)$ generated in step 1 is fed as input to a pre-trained PSP encoder [24] $\epsilon(.;\xi_e^s)$ and its output is given to each high-resolution block of ϕ_d^a to quench for $\mathbb{W}+$ space. We eliminate the noise input to each upsampling block to ensure that the generated images exhibit a consistent appearance without flickering.

Step 3: De-Identification: This step is performed in synchronization with the previous step. Here, for the $\mathbb{Z}^+$ space of $G(.;\phi_d^a)$, we use $z_e^+ = G(x;\theta_e)$ as the input to $G(.;\phi_d^a)$. This is performed to decode the identity information for de-identification. The design of downsampling blocks of $G(.;\theta_e)$ are aligned with the upsampling blocks

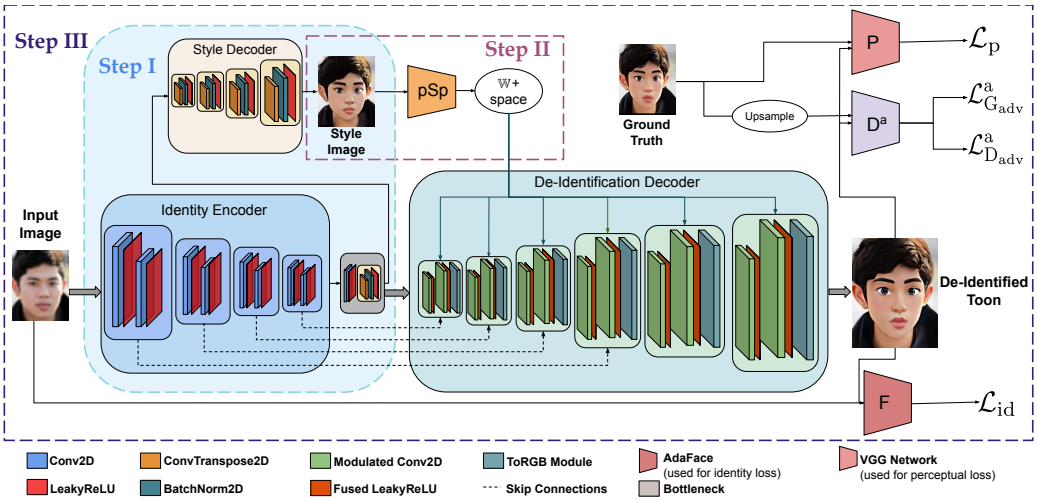

Figure 4. Illustration of the complete training process of the ToonerGAN to generate de-identified toon from a real face image.

of $G(.;\phi_d^a)$ to enable skip-connections for smoother information flow. $G(.;\theta_e,\phi_d^a)$ is then trained along with discriminator $D(.;\psi^a)$.

ToonerGAN: As shown in Figure 4, a real-face image $x \in X$ is given as input to an identity encoder $G(x;\theta_e)$ to obtain the identity embeddings z_e^+. In the first step, they are utilized by the style decoder $G(.;\phi_d^s)$ for style learning by deceiving $D(.;\psi^s)$ and generating $\bar{y}^s$. In the second step, z_e^+ is utilized by the de-identification decoder $G(.;\phi_d^a)$ for $\mathbb{Z}^+$ space. The output from $\epsilon(\bar{y}^s;\xi_e^s)$ is given as input to each block of ϕ_d^a appended with the skip-connections for W^+ space to output the high-resolution, identity de-identified toon image $\bar{y}^a$. To train $G(.;\theta_e,\phi_d^a)$ for style transfer, we apply the reconstruction loss $\mathcal{L}_\nabla$:

$$\mathcal{L}_\nabla = \lambda_{\mathrm{mse}}\mathbb{E}^{h\in H}_{\bar{h}^s\in\bar{H}^s}\|\bar{h}-h\|_2^2 + \lambda_{\mathrm{p}}\mathbb{E}^{h\in H}_{\bar{h}^s\in\bar{H}^s}\mathcal{L}_{\mathrm{p}}(\bar{h},h) + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} \tag{3}$$

This is a combination of three loss terms. The first term is the $\mathcal{L}_2$ loss between the upsampled (1024×1024) ground-truth images $H^s = h_1, h_2, ..., h_n$ and the up-sampled images $\bar{H}^s = \bar{h}_1^s, \bar{h}_2^s, ..., \bar{h}_n^s$ sampled from the set $\bar{Y}^s$. The next term $\mathcal{L}_{\mathrm{p}}$, is the perceptual loss [12] to minimize the semantic difference between the H and $\bar{H}^s$. In eq. 3, the final term is the de-identification loss $\mathcal{L}_{\mathrm{id}}$ to minimize the identity similarity between the real face and the generated toon image defined as follows:

$$\mathcal{L}_{\mathrm{id}} = -\left(1 - \frac{\mathcal{F}_{\mathrm{id}}(x)\cdot\mathcal{F}_{\mathrm{id}}(y)}{\|\mathcal{F}_{\mathrm{id}}(x)\|\cdot\|\mathcal{F}_{\mathrm{id}}(y)\|}\right) \tag{4}$$

$\mathcal{F}_{\mathrm{id}}$ refers to identity feature embeddings extracted from the pre-trained AdaFace model [16]. Eq. 3 is jointly minimized with generator adversarial loss $\mathcal{L}^a_{\mathrm{G_{adv}}} = \mathbb{E}_{x\in X}[\log(1 - D(G(x;\theta_e,\phi_d^a);\psi^a))]$ and discriminator adversarial loss $\mathcal{L}^a_{\mathrm{D_{adv}}} = \mathbb{E}_y[\log D(y)] + \mathbb{E}_x[\log(1-D(G(x)))]$, accounting to a total loss defined below:

$$\min_{\theta_e,\phi_d^a}\max_{\psi^a}\mathcal{L}_\nabla + \mathcal{L}^{\mathrm{a}}_{\mathrm{G_{adv}}} + \mathcal{L}^{\mathrm{a}}_{\mathrm{D_{adv}}} \tag{5}$$

4. Experimental Evaluation and Observations

The performance of ToonerGAN is evaluated through a two-pronged approach: assessing the model's ability in face de-identification and determining the practical utility of the generated images. The evaluation process includes tasks of identification and verification after face de-identification, hereafter referred to as "cartoon re-identification" and "cartoon re-verification". Additionally, the utility of the data is measured by calculating face detection rates on the toon avatars and comparing visual outcomes against established benchmarks. Furthermore, a human-centric study is conducted on both original and generated toon images to assess the algorithm's real-world efficacy from a user perspective.

Datasets: To evaluate the effectiveness of the proposed algorithm for de-identification, we leverage two publicly accessible face datasets: CelebA-HQ and Labeled Faces in the Wild (LFW). **CelebA-HQ dataset** [13] comprises 30,000 high-quality images sourced from CelebA, encompassing a total of 6,216 distinct identities. **LFW dataset** [7] contains 13,233 images, representing 5,749 unique identities. In our experimental setup, we specifically utilize "view 2" of the LFW dataset, which offers a collection of 6,000 face verification pairs evenly divided into positive and negative pairs.

Baseline Algorithms and Evaluation Settings: To evaluate the performance, we have compared the performance with a broad spectrum of de-identification techniques. This includes basic methods such as Blurring and Pixelation, as well as state-of-the-art methodologies, namely AnonymousNet(ANet) [20], NEO [26], LIVE [5], CIAGAN [21], DeIDGAN [17], Identity Transformer [6], IVFG [33], Iden-

Method	Recall@k = 1 ↓	Recall@k = 10 ↓	Detection Rate (%) ↑
Original	76.28%	80.94%	100%
Blurring	0.06%	0.38%	6.58%
Pixelation	0.26%	1.64%	48.54%
NEO [26]	0.33%	6.55%	99.93%
ANet [20]	8.04%	29.38%	99.44%
LIVE [5]	76.38%	80.94%	99.92%
CIAGAN [21]	0.26%	3.74%	99.83%
DeIDGAN [17]	0.02%	0.52%	**99.98%**
Ours	**0.015%**	**0.060%**	99.56%

Table 1. Comparing the performance (%) of ToonerGAN with existing conventional and state-of-the-art techniques for face re-identification task and detection rates (in %) on the CelebA-HQ dataset. Original denotes the baseline results using the original image. Here, lower Recall values signify better performance.

tityMask [29], and RiDDLE [18].

For obfuscating facial indexing experiments, we undertake two specific tasks using different datasets. Firstly, for the cartoon re-identification task, we de-identified a query image and then assessed its similarity against a database of real face images, each belonging to distinct identities. We calculate the *Recall@k* metric to quantify performance, indicating the percentage of correctly matched identity samples with their top-k nearest neighbors. This metric ranges from 0 to 100, where 0 signifies an optimal de-identification rate. Secondly, we perform a cartoon re-verification task, in which real-face and toon image pairs are considered. We de-identified the latter in each pair and subsequently evaluated the true positive rate at a 0.001 false positive rate, referred to as *TPR@0.001 FPR*. To compute this, we employ a pre-trained FaceNet model [25] (pre-trained on CASIA-Webface dataset [31]) to extract features. We also test the identity similarity between the de-identified and real face images and report the results obtained in Table 3. For data utility assessment, we compute the face detection rate of the generated output. For this, we employ MTCNN [35] to detect faces and then calculate the percentage of correctly detected images, where a value of 100 denotes the optimal face detection performance. Finally, we also analyze qualitative results and conduct a human study to analyze the visual quality of de-identified images.

4.1. Implementation Details

The generator-discriminator network $G(.;\theta_e,\phi_d^s)$ and $D(.;\psi^s)$ in step 1 is trained with Adam optimizer with learning rate 0.0005 to generate 256×256 toon images. The final de-identification model $G(.;\theta_e,\phi_d^a)$ and $D(.;\psi^a)$ is trained with Adam optimizer with a learning rate of 0.001 to generate final 1024×1024 de-identified images. We set λ_{mse} as 1000, λ_{p} as 100 and λ_{id} as 100 while training. All the experiments are performed on DGX A100 with an Intel Xeon CPU, 2 TB RAM, and eight 80 GB Nvidia A100 GPU cards. The code was written in PyTorch 1.12 with the Horovod framework for multi-GPU training. We discuss the architectural details of each component of the ToonerGAN in supplementary.

Method	TPR ↓
Original	0.965
LIVE [5]	0.035
CIAGAN [21]	0.019
Identity Transformer [6]	0.027
IVFG [33]	0.019
IdentityMask [29]	0.017
RiDDLE [18]	0.016
Ours	**0.009**

Table 2. Comparing ToonerGAN with conventional and state-of-the-art methods for face de-identification. A pretrained FaceNet model was tested on the LFW dataset at an FPR of 0.001. The values are between 0 and 1. Lower values signify better performance.

Method	ID Similarity ↓
CIAGAN [21]	0.068
FIT [6]	0.147
Personal [1]	0.187
RiDDLE [18]	0.056
Ours	0.104

Table 3. Comparing Identity Similarity between the original and de-identified images of CelebA-HQ dataset. The values are between 0 and 1. Here, lower values imply better performance.

4.2. Results on Obfuscating Facial Indexing

Experiments on cartoon re-identification and re-verification are performed and following observations are made:

Cartoon Re-Identification: The efficacy of the proposed algorithm for face de-identification is first tested on the cartoon re-identification task, and the performance is reported in Table 1. For this, features of the real face image from the CelebA-HQ dataset and its corresponding toon face image are extracted using a pre-trained CurricularFace [8] model. Then the percentage (%) of samples whose top-k nearest neighbors are from the same identity (Recall@k) is calculated. For the original model, we report an original recall rate of 76.28% for k=1 and 80.94% for k=10. Additionally, conventional anonymization methods such as Blurring and Pixelation exhibit optimal performance. Table 1 also illustrates that all other techniques attain promising results except LIVE [5], whose performance is closer to the original dataset's performance. However, the proposed approach surpasses all techniques, including Blurring and Pixelation, expressed as recall@k values, achieving state-of-the-art performance of 0.015 and 0.060 for both k=1 and k=10, respectively.

Cartoon Re-Verification: Table 2 presents the cartoon re-verification performance, measured as True Positive Rate at

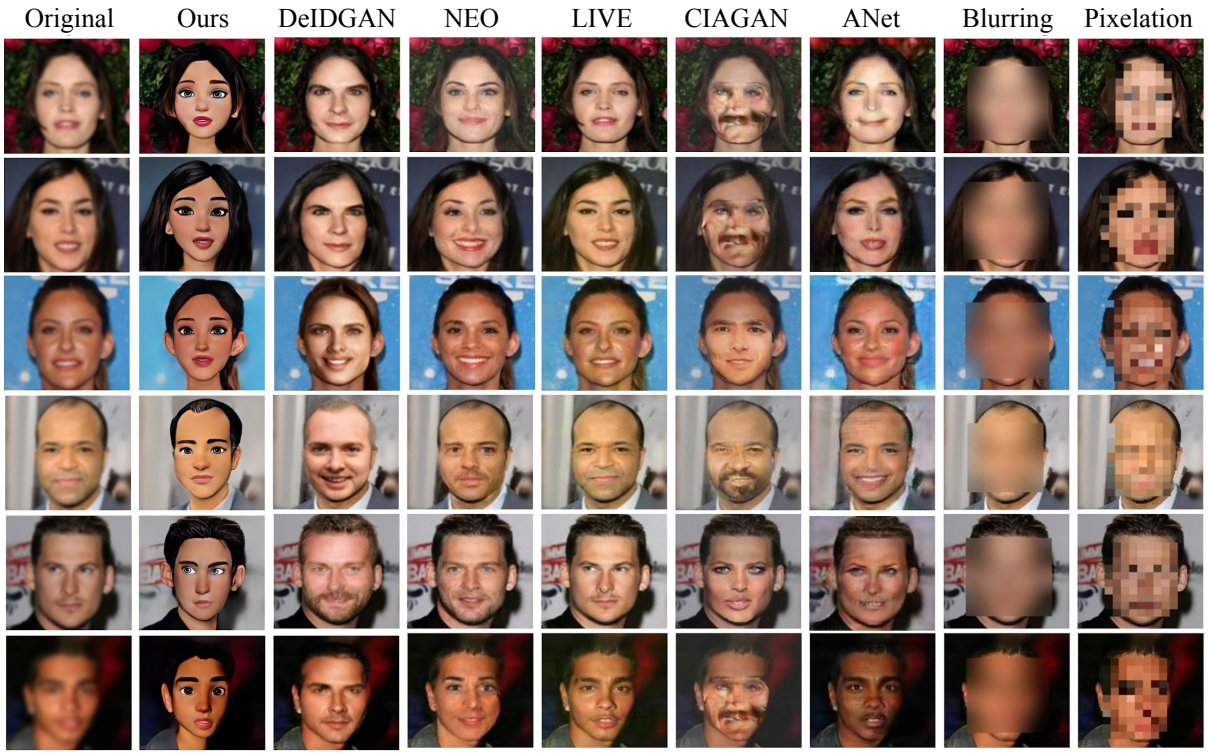

Figure 5. Visual comparison of the de-identified images generated using various existing methods and the proposed algorithm on the CelebA-HQ dataset. The original faces are blurred here for display to maintain anonymity.

0.001 False Positive Rate (TPR @ 0.001 FPR) on the LFW dataset. Notably, a FaceNet model [25] pre-trained on the CASIA-WebFace dataset [31] attains an initial TPR value of 0.965 on the verification set of the LFW dataset. From Table 2, it can be observed that while LIVE does not exhibit favorable re-identification results, it demonstrates competitive re-verification performance on the LFW dataset. We further note that ToonerGAN achieves state-of-the-art performance with a TPR of 0.009, followed by [18].

To further assess the efficacy of the proposed algorithm in obfuscating individual identities, we measured the average identity resemblance between original facial images and their corresponding cartoon avatars generated by ToonerGAN. For the evaluation, we used CurricularFace [8] as the benchmarking tool, focusing on the mean cosine similarity for images from the CelebA-HQ dataset. The findings, detailed in Table 3, are compared against results from existing methods. This comparative analysis shows that ToonerGAN achieves similar identity similarity scores, effectively demonstrating its capability in maintaining security and obscuring specific identity information.

4.3. Data Utility and Visual Assessment

With the significant improvement in obfuscating facial features achieved by our proposed ToonerGAN, surpassing existing methods, we further evaluate its performance in terms of visual quality, image aesthetics, and data reusability.

Qualitative Analysis: Figure 5 presents a comparative analysis of the visual de-identification results between the proposed and existing techniques. It is evident that approaches like Pixelation and Blurring significantly reduce facial details. In contrast, methods such as ANet and CIAGAN, while retaining more details, exhibit noticeable visual imperfections. NEO and DeIDGAN, on the other hand, slightly modify facial features but still retain enough detail for human recognition. LIVE shows minimal alteration to facial appearances. ToonerGAN has produced more realistic and distinctly altered cartoon faces compared to the original images. The cartoon faces effectively obscure facial features for automated systems while maintaining image utility for human viewers.

In Figure 6 (a), we demonstrate the temporal consistency of the proposed ToonerGAN algorithm. It is compared with the results of Maximov et al. [21] and Gafni et al. [5]. It can be observed that the pose is successfully maintained throughout, ensuring excellent temporal stability. Concurrently, Figure 6 (b) shows the performance of ToonerGAN on low-resolution data, demonstrating that it works effectively with low-resolution face-toon pairs without any dedicated training for low-resolution data. This demonstrates

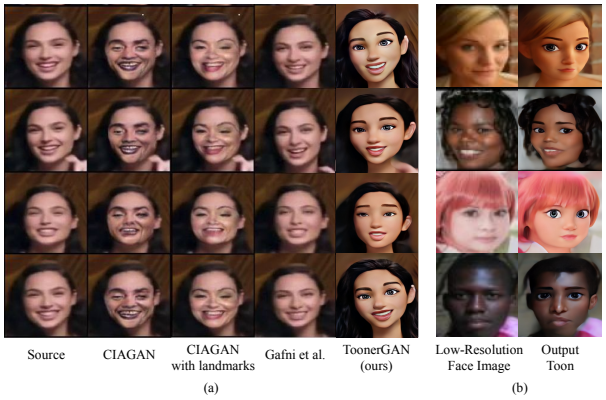

Figure 6. (a) Comparison of pose variation and Temporal Consistency of ToonerGAN with existing methods [5, 21], (b) Performance of ToonerGAN on low-resolution (32x32) data.

the applicability of the ToonerGAN algorithm in real-world applications such as maintaining temporal consistency and handling low-resolution data.

Quantitative Analysis: To further evaluate the performance of the proposed algorithm in terms of privacy preservation, we carried out an experiment reporting the mean identity similarity between the original face images and corresponding generated toons produced by ToonerGAN. For a fair comparison, CurricularFace [8] is used for this experiment and mean cosine similarity is calculated for CelebA-HQ dataset. In Table 3, we compare these results to the existing methods and observe that our approach achieves comparable identity similarity scores showing the ability of ToonerGAN to maintain security well.

For estimating utility of the data after de-identification, we calculate and report the face detection performance in Table 1. MTCNN model is utilized as the face detector on de-identified images of the CelebA-HQ dataset generated using the referred approaches. While Blurring and Pixelation exhibit limited face detection performance with detection rates of 6.58% and 48.54%, respectively, LIVE, CIAGAN, and DeIDGAN demonstrate nearly perfect face detection rates. We achieve a 99.56% detection rate, which is comparable to existing methods. We also test with DSFD [19] face detection algorithm and achieved a near-perfect performance of 99.98%.

Evaluating Human Performance for Data Utility: In many real-world scenarios, such as on social media platforms or in the metaverse, it is crucial for de-identified images to deceive deep models while remaining recognizable to humans. To assess the effectiveness of the proposed ToonerGAN algorithm from a human perspective, we conducted three distinct studies involving human subjects aged between 18 and 65 years.

The first study involved 15 human subjects (seven males and eight females) and focused on 50 random pairs of original and cartoon images generated using ToonerGAN. The subjects were tasked with identifying the original face from the corresponding toon images, a closed-set identification task. The second study involved 50 subjects and aimed to test human verification performance. Each subject was presented with 1000 random pairs of original and cartoon images generated using ToonerGAN, and they were asked to verify whether the original and corresponding toon images belonged to the same identity. In the third study, human evaluators were asked to label six facial attributes of 30 toon images generated by ToonerGAN.

Figure 7. Illustration of the inability of existing state-of-the-art toon style transfer method [30] to preserve semantics and context for effective de-identification.

Our observations from the first experiment indicate that ToonerGAN successfully preserves data utility for humans while achieving identity obfuscation from deep models. In the first task, all users were able to correctly identify identities with a 100% success rate. For the second task, the users achieved a toon-verification accuracy of 92.35%. Some users reported using ancillary information, such as background, to aid recognition, suggesting that the subjects were not easily recognizable by facial features alone. The third study resulted in an overall accuracy of 88% (gender: 97%, race: 80.10%, age: 68.67%, smiling: 87.30%, glasses: 97.16%, hat: 98.16%), underscoring the preservation of facial features. These results demonstrate the effectiveness of our algorithm in creating cartoon images that obfuscate identity information while maintaining recognizability.

4.4. Comparison with Existing Toonifications

Existing style transfer methods, such as Resolution Dependent GAN [23], DualStyleGAN [30], and Toonify [23], though capable of altering face images into various styles, often compromise the integrity of facial structures and characteristics due to overfitting. Similarly, techniques like AnimeGAN [2] and white-box cartoon representations [27], while transforming images, tend to lack diversity in styles and may inadvertently embed artifacts within the facial images. These drawbacks highlight the inadequacy of current algorithms for effective de-identification, as illustrated in Figure 7. To overcome these challenges, we introduce a

Figure 8. Ablation of the proposed ToonerGAN with its different components plugged in.

Metric	Step 1	Step 1 + (Step 2 + Step 3)
Recall@ (k=1, k=10) ↓	(8.08, 12.79)	(0.015,0.06)
TPR @ FPR = 0.001 ↓	0.19	0.009
ID Similarity ↓	0.355	0.104
Detection Rate (%) ↑	99.23	99.56

Table 4. Ablation performance over different metrics with different components of ToonerGAN algorithm.

Set	ToonSet	ToonerGAN (Step 1)	ToonerGAN (Steps 1+2+3)
ID Similarity ↓	0.344	0.322	0.1421

Table 5. ID-Similarity with Raw Toons from ToonSet, Step 1 of ToonerGAN, and final ToonerGAN (lower is better).

face-toon paired dataset, named ToonSet. By training our ToonerGAN in a supervised manner, we ensure that it not only maintains the original facial structure but also seamlessly transfers the toon style without introducing artifacts, thereby successfully de-identifying identities in the output.

4.5. Ablation Study

To thoroughly evaluate the impact of each component within ToonerGAN, we incrementally integrated components and generated the corresponding toon outputs. The results of this ablation study are presented graphically in Figure 8. We commence with the original face image x. The subsequent image illustrates the toon output generated by a model variant lacking the style distillation network. We notice that the model does not learn much without the style information. Following this, we examine the output from the model trained exclusively in stage 1, i.e., the output of $G(x; \theta_e, \phi_d^s)$. While this output bears resemblance to a toon, it falls short in capturing intricate details such as hair texture, ear contours, and teeth nuances, and lacks structural integrity. The subsequent image showcases the toon generated using the StyleGAN2 decoder, but without the incorporation of the perceptual loss $\mathcal{L}_{\mathrm{p}}$. Though this is a high-resolution avatar, it still does not comprehend the fine hair details, and the skin is not rendered as smooth. After enabling the style distillation through $\epsilon(G(x; \theta_e, \phi_d^s); \xi^s)$ and training it with $\mathcal{L}_{\mathrm{p}}$, the final toon image is high resolution and also consists of fine details in different facial regions while maintaining the overall structural integrity.

The contributions of each component of ToonerGAN are quantitatively analyzed in Tables 4 and 5. Table 4 reports the Recall (@k=1 and @k=10), TPR@FPR=0.001, ID similarity, and Detection Rate. Except for the detection rate, there is a significant difference in the outputs of all other metrics between step 1 and steps 2+3. Table 5 presents the ID-similarity after de-identification through Toonification. We observe a decline of over half from step 1 to steps 2+3, demonstrating their contribution to maintaining privacy. Therefore, it can be inferred that while step 1 focuses on the de-identification (privacy) objective, steps 2+3 significantly aid in both privacy and data-utility.

4.6. Inference Time

To demonstrate the real-time processing efficiency of ToonerGAN, we selected a varied sample of 1000 facial images from the ToonSet test dataset. The average inference time for these images was clocked at 68.69 milliseconds, highlighting the rapid processing capability. These time measurements were recorded using a single Nvidia A100 GPU card, leveraging the PyTorch framework.

5. Conclusion

This research introduces ToonerGAN, a novel approach to face de-identification through the process of toonification. The algorithm is trained in three distinct phases: style learning, style distillation, and de-identification. An integral component of our approach is the creation of ToonSet, a comprehensive paired-toon dataset comprising over 23,000 image pairs that feature faces and their corresponding toon representations. ToonerGAN's effectiveness is rigorously evaluated in terms of obfuscating facial features and preserving data utility, with comparative analysis against existing methods. Our findings showcase that ToonerGAN yields high-fidelity, realistic, and diverse toon images that effectively obscure facial features while retaining data utility. The empirical evidence highlights a significant enhancement in performance over current techniques. Furthermore, a human study affirms the visual appeal and recognizability of the generated toon images by human observers. Conclusively, this paper establishes ToonerGAN as an efficacious solution for face de-identification through toonification, with broad applicability in various computer vision domains.

6. Acknowledgement

This study is supported by the Aria Research Grant by META. The authors thank O. Parkhi for insightful discussions, and B. Dutta and the volunteers for their help in the user study. Thakral received partial support from the PMRF Fellowship and Vatsa is partially supported by the Swarnajayanti Fellowship.

References

[1] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3334–3342, 2021.

[2] Jie Chen, Gang Liu, and Xin Chen. Animegan: a novel lightweight gan for photo animation. In *International symposium on intelligence computation and applications*, pages 242–256. Springer, 2020.

[3] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k-facial attributes via adversarial perturbations. In *International Joint Conference on Artificial Intelligence*, page 656–662. AAAI Press, 2018.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[5] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *IEEE International Conference on Computer Vision*, pages 9378–9387, 2019.

[6] X. Gu, W. Luo, M. S. Ryoo, and Y. J. Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European Conference on Computer Vision*, pages 727–743. Springer, 2020.

[7] 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*, pages 07–49, 2008.

[8] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[10] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. Stylecarigan: caricature generation via stylegan feature map modulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–16, 2021.

[11] Baowei Jiang, Bing Bai, Haozhe Lin, Yu Wang, Yuchen Guo, and Lu Fang. Dartblur: Privacy preservation with detection artifact suppression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16479–16488, 2023.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[14] 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*, pages 4401–4410, 2019.

[15] 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*, pages 8110–8119, 2020.

[16] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[17] Zhenzhong Kuang, Huigui Liu, Jun Yu, Aikui Tian, Lei Wang, Jianping Fan, and Noboru Babaguchi. Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3182–3191, 2021.

[18] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8093–8102, 2023.

[19] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[20] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–65, 2019.

[21] Maxim Maximov, Ismail Elezi, and Laura Leal Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2020.

[22] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[23] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020.

[24] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[26] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfus-

cation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018.

[27] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8090–8099, 2020.

[28] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

[29] Y. Wen, B. Liu, J. Cao, R. Xie, L. Song, and Z. Li. Identitymask: Deep motion flow guided reversible face video de-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[30] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022.

[31] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[32] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[33] Zhuowen Yuan, Zhengxin You, Sheng Li, Zhenxing Qian, Xinpeng Zhang, and Alex Kot. On generating identifiable virtual faces. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1465–1473, 2022.

[34] Liming Zhai, Qing Guo, Xiaofei Xie, Lei Ma, Yi Estelle Wang, and Yang Liu. A3gan: Attribute-aware anonymization networks for face de-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5303–5313, 2022.

[35] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23 (10):1499–1503, 2016.