ToonStyleGAN
Bilkent University CS485 Project

Utku Kurtulmus, Onurcan Atac, and Kutay Senyigit

***Abstract*—In this paper, we present a novel approach for transforming real images into high-quality, stylized cartoon representations using an integrated pipeline of DualStyleGAN and SemanticStyleGAN. DualStyleGAN initially converts real facial images into cartoon-style portrait images by utilizing global and local styles, ensuring high artistic fidelity while preserving realistic proportions. Subsequently, SemanticStyleGAN refines these outputs, enabling precise edits to specific facial features, such as adjusting the curve of a smile or skin texture details. Our evaluation using Kernel Inception Distance (KID) scores shows the effectiveness of this dual-model approach. The KID scores of 0.173659 for DualStyleGAN outputs and 0.173416 for SemanticStyleGAN outputs indicate strong stylistic fidelity and consistency with the reference cartoon dataset. Compared to existing state-of-the-art models like StyleGAN2 and CartoonGAN, our method provides superior control and detail in the generated images. Despite the challenges posed by the relatively small training dataset, our results verify that our approach has merit. Future work will focus on expanding the dataset and optimizing the model architecture to further enhance performance and efficiency.**

***Index Terms*—StyleGAN, Generative Adverserial Networks, SemanticStyleGAN, DualStyleGAN**

I. Introduction

In computer vision and graphics, synthesizing realistic and expressive images through deep learning has shown revolutionary advancements. Our project, focusing on the generation of cartoon character faces, seeks to explore and expand those technologies, particularly the capabilities of cutting-edge generative adversarial networks (GANs), through the lenses of DualStyleGAN [1] and Semantic StyleGAN [2]. These technologies made creating detailed and varied representations of cartoon characters possible and enabled the manipulation of concrete facial features with unprecedented control and flexibility.

Our primary investigation is the application of DualStyleGAN and Semantic StyleGAN to a specially selected 'Toonify Dataset' [3]. This dataset provides an environment for our models to learn and adapt the distinct stylistic features typical to cartoon images. DualStyleGAN, known for its proficiency in exemplar-based high-resolution portrait style transfer, allows for the detailed synthesis of faces that blend realistic textures with cartoonish styles. On the other hand, Semantic StyleGAN brings a compositional understanding, enabling the manipulation of image synthesis in a more detailed manner such as changing the curve of a smile or the flick of a hair.

II. Related Work

The field of generative adversarial networks (GANs) has grown significantly with each innovation, beginning with the work of Goodfellow et al. [4], which builds the foundation for adversarial training methodologies in generative modeling. This concept, where two models (a generator and a discriminator) are trained in competition, leading to the synthesis of increasingly realistic images.

Deep Convolutional GANs (DCGAN) by Radford et al. [5] built upon this foundation by incorporating convolutional neural networks into the GAN architecture, which was a pivotal improvement. DCGAN not only stabilized the training process but also enhanced the resolution and quality of the generated images, making it a benchmark for future developments and applications in image generation tasks.

Progressive advancements in the methodology become more visible with Progressive Growing of GANs by Karras et al. [6], which introduced a method of gradually increasing the resolution of generated images during training. This technique addressed training stability and quality issues that existed in previous models by incrementally growing the network's architecture, which allowed for the generation of high-resolution images without compromising the stability of the learning process.

Building on these improvements, Karras et al. introduced StyleGAN [7], which implemented a style-based approach to the generator architecture. This model significantly diverged from traditional GANs by using a mapping from latent codes to generate styles that directly influence each convolutional layer through adaptive instance normalization (AdaIN). This allowed for precise control over the "style" aspects of image synthesis, such as texture and color details, enhancing the ability to produce diverse and realistic images. StyleGAN's architecture also helped to separate and identify different high-level features and random variations in images, which improved both the interpolation properties and the visual quality of the generated outputs.

Following the introduction of StyleGAN, Image2StyleGAN by Abdal et al. provided an innovative exploration into the embedding of arbitrary images into the latent space of a pre-trained StyleGAN. This method allowed for detailed manipulation of existing photographs through operations like morphing, style, and expression transfer, expanding the

utility of StyleGAN for practical image editing applications. Image2StyleGAN demonstrated how the StyleGAN latent space could be used to not only generate new images but also transform existing images with high levels of precision and flexibility.

The advancements offered by StyleGAN and Image2StyleGAN [8] enabled further specialized adaptations such as DualStyleGAN [1], and Semantic StyleGAN. DualStyleGAN extended the StyleGAN framework by introducing a dual-pathway architecture that allowed for the manipulation of intrinsic and extrinsic styles, providing enhanced control over the artistic and stylistic elements of the images. This model was particularly effective for high-resolution portrait style transfer, enabling precise control over the detailed synthesis of image attributes.

Semantic StyleGAN [2] further increased the precision of style control by splitting the image generation process into localized semantic components, each represented by independent latent codes. This approach significantly advanced the capacity for detailed, localized editing of specific image attributes, such as facial features in portraits, offering a powerful tool for tasks that require fine-grained image manipulation.

The integration of these technologies—DualStyleGAN, Semantic StyleGAN, and Image2StyleGAN—exhibits the dynamic environment of GAN research. Each contribution adds to image quality and realism but also expands the practical applications of GANs in digital art, animation, and other creative industries. Our project utilizes these advanced models to explore a novel application in generating and manipulating cartoon character faces, aiming to create a valuable tool that can implement both of those goals adequately.

III. Experiments

A. Dataset Description

For our project on cartoon character face generation using DualStyleGAN and Semantic StyleGAN, we have utilized the 'Toonify Dataset' [3]. This dataset is specially preserved to support generative adversarial network models in learning to synthesize and manipulate facial features of cartoon characters with high realism and artistic fidelity.

The Toonify Dataset is utilized in a staged training process, where DualStyleGAN first learns to adapt the styles from real faces to cartoon representations. This involves detailed synthesis where the model must manipulate concrete features such as hair texture, eye shape, and facial contours to align more closely with cartoon aesthetics while retaining underlying realistic structure and proportions.

Subsequently, Semantic StyleGAN utilizes the dataset to gather control over the manipulation of individual semantic components. By training on the Mixed images, Semantic StyleGAN learns to apply fine-grained edits, such as adjusting the curve of a smile or an eyebrow, in a stylistically coherent manner with cartoon art.

To optimize the performance of our GANs, the images in the Toonify Dataset go under several preprocessing steps including alignment, normalization, and augmentation. Alignment ensures that facial features are consistently positioned in the images, assisting the network in learning feature localization. Normalization adjusts the pixel values for uniformity, enhancing model sensitivity to stylistic variations rather than color scale differences. Augmentation techniques such as random cropping, flipping, and color jittering are applied to increase the robustness of the model against overfitting and to improve its ability to generalize across different cartoon styles.

This dataset not only supports our primary objective of generating high-quality cartoon character faces but also enhances our models' capabilities in handling complex style transfers, making them powerful tools for creative industries. By utilizing the Toonify Dataset, we aim to get impressive and accurate results in the synthesis and artistic rendering of digital faces.

B. Method Description

In our approach, we have integrated the capabilities of DualStyleGAN and SemanticStyleGAN by combining their architectures in series. This integration utilizes the strengths of both models to produce a pipeline for image synthesis and editing.

The output of DualStyleGAN, known for its advanced style transfer capabilities, is fed directly into SemanticStyleGAN. This sequential processing allows us to first generate a high-resolution, stylistically enhanced image with DualStyleGAN, which effectively handles complex style adaptations and the transfer of artistic qualities from one domain to another.

Once this image is synthesized, the encoder of SemanticStyleGAN takes over. The encoder in SemanticStyleGAN is specifically designed to encode the output from DualStyleGAN into its own system. This step is crucial as it translates the stylized output into a format that SemanticStyleGAN's generator can comprehend and manipulate further.

The encoder in SemanticStyleGAN is capable of handling images that contain mixed elements of realism and artistic flair, making it particularly suitable for the outputs produced by DualStyleGAN. Both models are trained on the Toonify dataset, which consists of a mix of cartoon and real facial images. This common training ground ensures that the encoder can effectively perform an "inversion" of the style-transferred images back into a latent space that SemanticStyleGAN can manipulate. This inversion is of high quality due to the shared data characteristics and training, enabling a smooth transition between the two model outputs.

This combined approach allows us to leverage DualStyleGAN for its ability to generate high-quality artistic portraits with flexible style control, and then use SemanticStyleGAN to fine-tune the structure and texture of specific local parts of the image. The result is an exceptionally refined image where artistic styles are integrated with local control over detailed aspects, such as detailed facial features.

By integrating these models in series, we enhance our capability to produce images that are not only visually pleasing but also highly customizable. This integration is beneficial in scenarios where detailed, controlled modifications are required after the initial style transfer, providing users with a powerful tool for image synthesis and editing. This method holds significant potential for applications where precise control over both global styles and local details is paramount.

C. Details of the Trained Models and Configuration

Fig. 1: General Structure of the ToonStyleGAN

Our architecture integrates DualStyleGAN and SemanticStyleGAN in a sequential pipeline to transform real images into detailed and customizable cartoon representations. Initially, DualStyleGAN, which incorporates both a global and local style network, generates high-resolution cartoon images by adapting styles from real faces while retaining realistic proportions. The model utilizes a multi-scale discriminator and a combination of adversarial, perceptual, and style losses to achieve high stylistic fidelity. SemanticStyleGAN, leveraging the StyleGAN2 framework with additional layers for semantic control, enables fine-grained edits to specific facial features, ensuring that modifications such as smile curve or eyebrow arch remain consistent with those in the cartoon faces. This integration allows for the production of images that are both compelling and also finely tuned in detail.

Figure 2 illustrates the architecture of DualStyleGAN. It incorporates two style paths: intrinsic and extrinsic. The intrinsic style path and generator network form a standard StyleGAN, which remains fixed during the fine-tuning process. This path accepts an intrinsic style code (z) of unit Gaussian noise, embedded artistic portraits (z_i^+), or real faces (z^+) as input. On the other hand, the extrinsic style path utilizes the extrinsic style code (z_e^+) of artistic portraits, which captures semantic cues such as hair colors and facial shapes. Extrinsic style codes can also be sampled through a sampling network (N) by mapping unit Gaussian noises to the extrinsic style distribution.

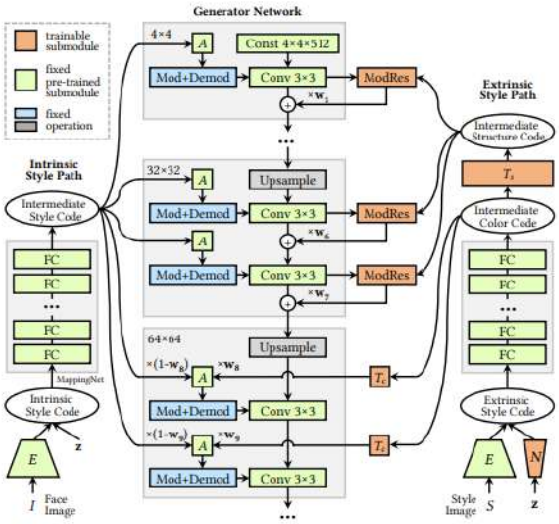

Fig. 2: Architecture of DualStyleGAN [1]

DualStyleGAN offers hierarchical style control at different resolutions. In the fine-resolution layers (8-18), the extrinsic style path utilizes the same strategy as StyleGAN, by using color transform blocks (T_s) and affine transform blocks (A) to characterize domain-specific colors. For the coarse-resolution layers (1-7), modulative residual blocks (ModRes) are introduced to adjust structural styles, and a structure transform block (T_s) is added to characterize domain-specific structural styles.

Experimental results demonstrate that using residual blocks with AdaIN in the convolution layers of the residual path can effectively approximate the changes in convolution weight matrices during fine-tuning. This approach proves to be more effective than modulations in channel or spatial dimensions alone.

DualStyleGAN enables hierarchical modeling of complex styles, allowing for the control of both color and structural styles at different resolutions. Moreover, it supports flexible style manipulation between two domains using the weight parameter w. Furthermore, by keeping the pre-trained StyleGAN during fine-tuning, DualStyleGAN reduces the issue of mode collapse, preserving the original, diverse facial features. Finally, the additive property of the modulative residual block helps with robust structure preservation.

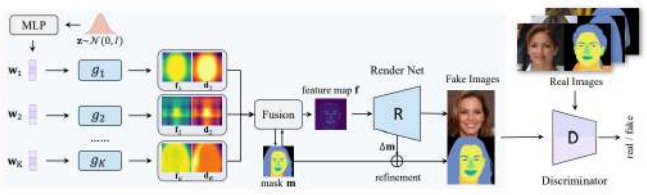

Fig. 3: Architecture of SemanticStyleGAN [2]

Figure 3 presents an overview of the training framework for

SemanticStyleGAN. SemanticStyleGAN learns to disentangle the latent space for different semantic areas in an image. The generator takes randomly sampled codes and maps them into a W space using an MLP. The resulting w code is then used to modulate the weights of local generators, each responsible for generating a specific semantic part of the image.

Each local generator outputs a feature map (f_k) and a pseudo-depth map (d_k). The pseudo-depth maps are used to generate a coarse segmentation mask m through a softmax function, which mimics the z-buffering process in composition. The feature maps from all local generators are then aggregated using the coarse segmentation mask to produce a global feature map f.

The render network R, conditioned only on the global feature map f, refines the upsampled coarse segmentation mask m into a high-resolution segmentation mask by learning a residual Δm. It also generates the final fake image based on the refined segmentation mask and the global feature map. The render network is similar to the original StyleGAN2 generator, with modifications such as the removal of modulated convolution layers and the input of feature maps at both low and high resolutions.

To ensure that the generated images and their corresponding segmentation masks follow the joint distribution of real images and masks, a dual-branch discriminator is used. The discriminator takes both the RGB image and the segmentation mask as input, processing them through separate convolution branches before summing up the outputs for fully connected layers. This design allows for separate regularization of the gradient norm in the segmentation branch using an additional R1 regularization loss.

The training framework of SemanticStyleGAN is similar to that of StyleGAN2, with the addition of a mask regularization loss to ensure that the final refined segmentation mask does not diverge too much from the coarse mask. The overall loss function is a combination of the StyleGAN2 loss, the mask regularization loss, and the R1 regularization loss for the segmentation branch in the discriminator.

IV. Results

A. Comparing and Discussing the Evaluation Results of the Models

1) Evaluation of Model Generation Performance: In ToonStyleGAN, we are able to generate multiple images with different structural and color weights. Structural weight represents the recreated version of the original image, ”the real face”, and the color weight represents the selected toon style image. By experimenting with those weights, it is possible to determine the better samples out of the generated images. Figure 4 shows the spectrum of the outputs with different structural and color weights.

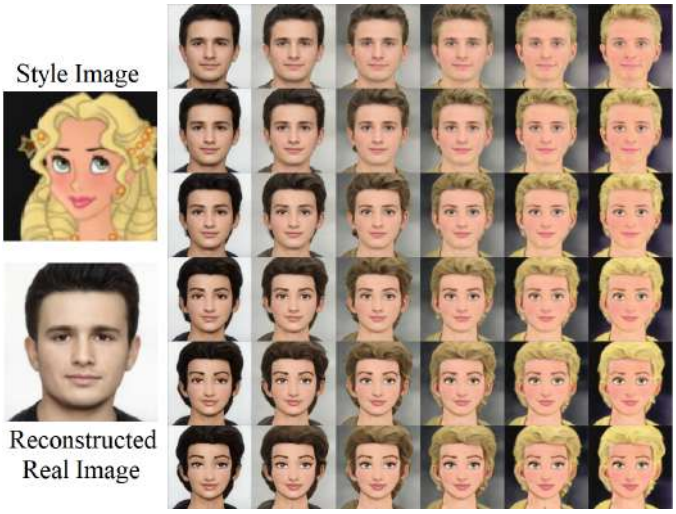

Fig. 4: Style Mixing Results

To assess how well our combined DualStyleGAN and SemanticStyleGAN design performs we used the Kernel Inception Distance (KID) metric. This metric helps to measure the similarity, between the images generated by the model and a predefined set of reference images. The KID score proves to be a choice for our assessment because it offers a measure of both the quality and variety of the generated images when compared to the reference dataset. Additionally, it is less sensitive to the sample size in comparison to metrics, like the Inception Score (IS).

We chose KID for several reasons:

- **Robustness**: KID is known for its robustness to the number of samples, making it a reliable metric for evaluating models even when the dataset size is relatively small.
- **Bias-Free**: Unlike the Fréchet Inception Distance (FID), KID does not assume Gaussianity, providing a more unbiased measure of distance between distributions.
- **Precision**: KID allows for an unbiased estimation of the similarity between two distributions. It provides both a mean score and a confidence interval.

We calculated two KID scores to comprehensively evaluate our model:

- **KID for DualStyleGAN Output**: This score was calculated between the style dataset (317 cartoon images) and the output dataset of DualStyleGAN (100 images). This metric helps us understand how well DualStyleGAN alone can translate real images into cartoon styles.
- **KID for SemanticStyleGAN Output**: This score was calculated between the style dataset (317 cartoon images) and the output dataset of SemanticStyleGAN (generated from the 100 DualStyleGAN output images). This metric evaluates the combined performance of DualStyleGAN and SemanticStyleGAN, indicating how effectively the final output aligns with the cartoon-style dataset.

TABLE I: KID Scores for DualStyleGAN and SemanticStyleGAN Outputs

Model Output	KID Score	KID Score (± Std. Dev.)
DualStyleGAN Output	0.173659	0.173659 ± 0.005863
SemanticStyleGAN Output	0.173416	0.173416 ± 0.006320

These scores indicate that both models achieve a comparable level of quality and stylistic fidelity relative to the reference cartoon dataset. Also, SemanticStyleGAN showed a slight improvement in the alignment with the style dataset.

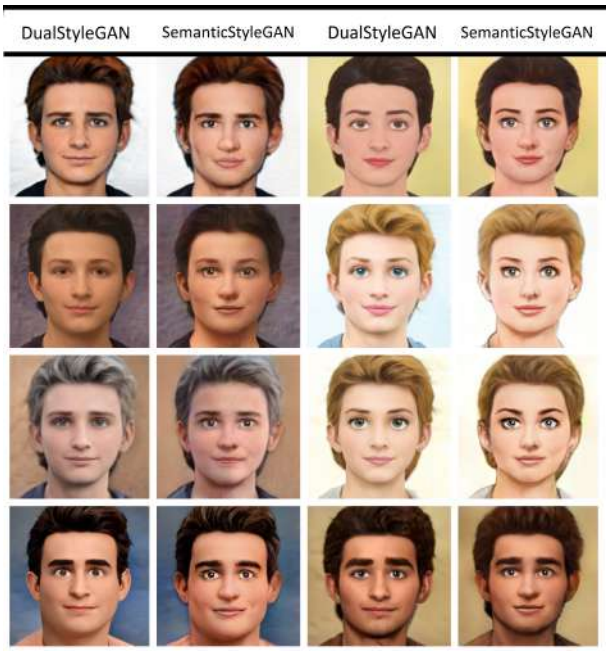

Fig. 5: Comparison of DualStyleGAN output and SemanticStyleGAN reconstruction

As we can see from Figure 5, DualStyleGAN outputs and their reconstructed versions in SemanticStyleGAN latent space are very similar, verifying our KID findings.

2) Evaluation of Style Editing Performance: Figure 6 showcases the editing capabilities of ToonStyleGAN through latent interpolation results. Each row represents a specific aspect of the synthesized facial images, such as hair, coarse details, ears, clothing, face, and background. The first row's three column demonstrate the model's ability to smoothly interpolate between different hairstyles, levels of coarseness, and ear shapes. The clothing column illustrates how the model can alter the clothes of the individuals, including accessories like glasses. The face column shows the model's proficiency in capturing and manipulating the general structure of facial features and textures. Lastly, the background row demonstrates the model's capacity to generate and blend various backgrounds behind the facial images.

The success of our style editing results comes from the learned semantic mappings in the latent space of the SemanticStyleGAN architecture. Different from StyleGAN where the latent codes are ordered from coarse to fine, latent codes

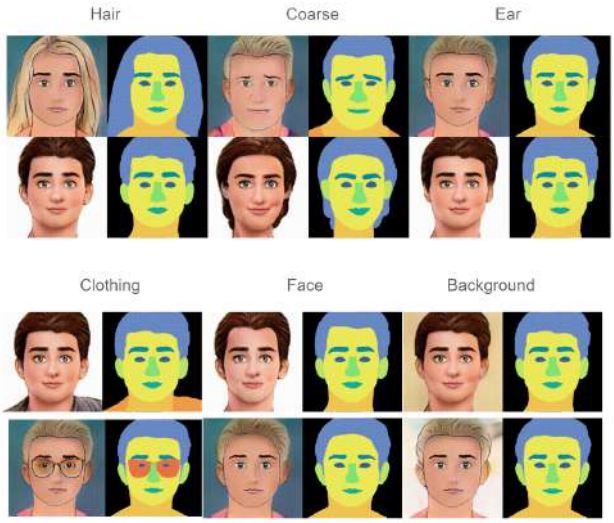

Fig. 6: Latent Interpolation Results

in SemanticStyleGAN are learned to represent the semantic information of the different parts of the images, allowing more precise control over different parts of the images. By interpolating in those latent codes, specific features in the image can be changed without changing the overall structure.

B. Comparing and Discussing the Evaluation Results of the Models

The inclusion of DualStyleGAN and SemanticStyleGAN in our pipeline shows significant advancements in transforming real images into high-quality, stylized cartoon representations while allowing precise semantic edits. The KID scores for our models—0.173659 for the DualStyleGAN output and 0.173416 for the SemanticStyleGAN output—reflect their ability to produce images that closely match the artistic style of the reference cartoon dataset. These results indicate a strong stylistic fidelity and consistent performance across both stages of the pipeline.

When compared to other state-of-the-art models such as StyleGAN2 [10], CartoonGAN [9], and other GAN variants focused on style transfer, our approach shows various distinct advantages. StyleGAN2, while it is great at generating high-fidelity images, it lacks control over specific features that SemanticStyleGAN provides. CartoonGAN, designed specifically for cartoon stylization, often falls short in maintaining the fine details and realism in the generated images, which our combined pipeline handles more effectively.

Despite these strengths, our models have limitations. The reliance on a relatively small dataset (317 images for training) could potentially limit the diversity and robustness of the generated images. Additionally, the sequential nature of our pipeline, while beneficial for control and quality, introduces complexity and longer processing times compared to single-step models.

Our evaluation using KID highlights the strength of our dual-model approach in achieving high stylistic fidelity and detailed control. The slight improvement in KID score from DualStyleGAN to SemanticStyleGAN underlines the effectiveness of the second model in refining and enhancing the initial outputs. However, future work could focus on expanding the training dataset and optimizing the model architecture to further improve performance and reduce processing time.

REFERENCES

[1] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, *Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7693-7702.

[2] Y. Shi, X. Yang, Y. Wan, and X. Shen, *SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11254-11264.

[3] J. Pinkney, *Toonify*, GitHub. [Online]. Available: https://github.com/justinpinkney/toonify.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative Adversarial Nets*, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[5] A. Radford, L. Metz, and S. Chintala, *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[7] T. Karras, S. Laine, and T. Aila, *A Style-Based Generator Architecture for Generative Adversarial Networks*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] R. Abdal, Y. Qin, and P. Wonka, *Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?*, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[9] Y. Chen, Y.-K. Lai, and Y.-J. Liu, *Cartoongan: Generative Adversarial Networks for Photo Cartoonization*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9465-9474.

[10] 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 (CVPR)*, 2020.