CS485 Homework
Literature Survey and Critical Analysis

Onurcan Ataç, 22002194

***Abstract*—The main aim of this paper is to conduct a literature review of various papers that use Generative Adversarial Networks (GANs). These papers are all retrieved from 2022-2023 conferences and capture recent developments. The propositions and limitations of these papers are explored. In the end, the MI-GAN algorithm is investigated in detail by input and output images, and cases in the algorithm that do not perform well are found.**

I. Introduction

In this paper, Generative Adversarial Networks (GANs) will be further studied by exploring 6 papers on the topic. Then, one of the papers that have their code released will be selected, results will be tested by running the code the authors have provided, and different input images will be fed to the code to see how the results of the algorithm change. This topic is chosen since GANs played a crucial part in the development of image generation using machine learning, they cover a significant part of our course, and further study of this topic would be especially beneficial for me since our course project is also about GANs. Moreover, there are many papers to choose from since GANs are a popular research area.

II. Investigation of Chosen Papers

A. Paper 1: StyleGAN knows Normal, Depth, Albedo, and More [1]

Intrinsic images are defined by Barrow and Tenenbaum as "characteristics – such as range, orientation, reflectance, and incident illumination – of the surface element visible at each point of the image" [2]. Intrinsic images are heavily used in computer vision and computer graphics tasks, intrinsic images have to be created in order to be used in certain tasks. The main intrinsic images gathered from images in the experiment were surface normals, depth, albedo, shading, and segment.

The paper proposes and demonstrates that StyleGAN can be used to access the intrinsic images of an image accurately and in a straightforward way. The process is maintained by searching offsets for the latent variables that produce the desired type of intrinsic image. Assuming G(w) is produced by the generator using latent code w, for each type of intrinsic image, there is a constant dt such that G(w+dt) is that type of intrinsic version of the image G(w). The authors also emphasize that as a result of their control experiments, they have seen that StyleGAN does not behave as a simple image regressor. It does not directly regress from input images to intrinsic properties, but it manipulates its latent space accordingly so that it can produce intrinsic images with desired characteristics.

The authors compared their results with SOTA image regression techniques. The recovered intrinsic images compare very well to the ones that are produced by other methods. The authors have found that even though the intrinsic images should ideally be invariant to the lighting changes, SOTA methods were sensitive to those changes. The images that StyleGAN produces are more robust to changes in lighting conditions compared to SOTA methods. This might be especially useful in computer vision implementations. This gives the StyleGAN method an advantage. The authors used StyLitGAN to generate different lighting versions of the images in order to use them in their experiment. The authors used a pre-trained StyleGAN model, and this method to obtain intrinsic images does not need to be shown many image intrinsic image pairs. Therefore this method does not require further complications and effort and is straightforward. In the experiments, 2000 unique images generated by the StyleGAN were used, but 200 was good enough for a sensible prediction. Note that a d(c) intrinsic image type vector learned works for any image.

Even though the experiment and paper are indeed valuable, they come with limitations. First and most important, the code is not accessible, so the results are not easily practically verifiable by outside contributors. Another important limitation of the experiment is that there are no "ground truth" intrinsic images. The reference intrinsic images were obtained by SOTA methods, and these were regarded as ground truth. The authors also found out that StyleGAN cannot complete non-intrinsic tasks, such as changing the halves of the image by an offset.

B. Paper 2: Fine-Grained Face Swapping via Regional GAN Inversion [3]

Face swapping is a process that aims to transfer the identity information of a source face to the given target face while retaining the identity irrelevant information of the target image such as expressions, head position, and image background. The authors mention existing methods for face swapping such as using pre-trained 2D face recognition networks, but they emphasize that these models may miss important identity information since they are mainly designed for classification, not for generation. Therefore, the results obtained from those methods have the "in-between effect", where the swapped face resembles the target face along with

the source face, which is not wanted since it is more a face mix than a swap.

The paper mainly suggests a new method for high-fidelity face swapping. A new approach "editing for swapping" (E4S) is proposed. E4S does not treat face swapping as a simple transfer of facial features, it treats face swapping like a detailed editing task considering individual components of the face and manipulates these components separately. E4S also disentangles the shape and texture information of facial components. Shape refers to the geometric structure of facial components, and texture refers to color variations, wrinkles, etc. This approach enables a more consistent and realistic face swap operation. The Novel Regional GAN Inversion (RGI) method that constitutes the core of the system allows for shape and texture disentanglement. The framework proposed also allows for controlling swap, and partial face swapping.

The E4S network is made of two main parts, reenactment and swapping, and generation. In the reenactment part, the face regions of the images are cropped, and using StyleGAN and a pre-trained face reenactment model FaceVid2Vid an output image is produced. In the second part, two mask-guided multi-scale encoders, shape swapping, and texture swapping are done separately, and a pre-trained StyleGAN generator is used to generate the final image with a mask-guided injection module to synthesize the swapped face. The results are then compared to other face-swapping methods. This method outperforms others in source face ID retrieval metrics and competes with target image feature retention even though it is slightly outperformed in those metrics.

There are limitations of the paper. Even though the method obtains close results in the target image feature retention compared to other methods, it is still outperformed. This model seems to be excessively optimized for source face identity preservation and therefore sometimes fails to retain target image features. Also, some images appear to be unnatural compared to other methods.

C. Paper 3: StyleDomain: Efficient and Lightweight Parameterizations of StyleGAN for One-shot and Few-shot Domain Adaptation [4]

Domain adaptation of GANs is the problem of fine-tuning the GAN models that are pre-trained on large datasets in order to fit them in a domain. This is a crucial contemporary problem since the training of these models takes excessive time and computational resources.

This paper provides a systematic and in-depth analysis of domain adaptation problems in GANs. The most important parts of a StyleGAN in adapting to a new domain are explored, and novel efficient, lightweight parameterizations are proposed for domain adaptation. The authors suggest that most of the existing methods have an assumption that StyleGAN can be adapted to a new domain only if we fine-tune almost all its weights, even for similar domains. The authors think this is not the correct approach and see a lack of analysis in this decision. In the case of similar A and B domains, it is indicated that fine-tuning the affine layers is sufficient for domain adaptation. More parameters should be optimized for distant domains, but still not the whole network.

New parameterizations of StyleGAN are proposed in the paper. StyleSpace and StyleSparse parameterizations enable efficient domain adaptation, particularly for one-shot domain adaptation. StyleSpace works by optimizing the direction in the style space in fine-tuning, by optimizing the the style vector. StyleSpaceSparse builds on StyleSpace, by pruning technique, it zeroes out most of the coordinates in the style domain vector. It focuses on reducing the number of parameters significantly without significant information loss. StyleSpace and StyleSpaceSparse work well, especially in one-shot scenarios. Affine+ built on Affine, and AffineLight+ built on Affine+ work well with few-shot scenarios by outperforming existing methods, reducing parameters significantly with insignificant degradation in quality.

A limitation of the paper is that it primarily focuses on specific types of image datasets such as faces, cats, and dogs, and therefore the results might be non-generalizable.

D. Paper 4: MI-GAN: A Simple Baseline for Image Inpainting on Mobile Devices [5]

Image inpainting is the task of completing the missing regions in an image in a realistic way. It is mostly used for restoring old or disrupted images, and object removal from images.

The authors indicate that even though a lot of image inpainting methods have been developed and provided impressive results in recent years, there is no image inpainting model designed to work on mobile devices. They therefore developed MI-GAN, a significantly computationally cheaper and smaller model than existing inpainting models. Their main concern is computational cost and easy deployment on mobile devices, while still matching the image quality performance of SOTA models.

MI-GAN introduces the first mobile generative image inpainting network, with a customized knowledge distillation method, and model re-parametrization strategy. Furthermore, it compares strongly to the existing inpainting models (in metrics such as FID, LPIPS, and human evaluators), while being lightweight. When compared to other image inpainting models, MI-GAN performs significantly better in FLOPS metric and number of parameters. This is as expected since MI-GAN is designed to be a lightweight network. FLOPS is a measurement of the required computing power in order to run the algorithm. The architecture of the model includes depthwise convolution with re-parametrization, bilinear resizing, and pointwise convolution while also including random noise. A combination of adversarial and knowledge distillation loss functions is used.

Even though the model performs similarly to the other image inpainting methods, it still struggles with inpainting tasks of images that contain complex 3D structures. This prevents the algorithm from producing visually plausible results in those kind of images.

E. Paper 5: GLeaD: Improving GANs with A Generator-Leading Task [6]

GAN is essentially a two-player game played between a generator (G) and a discriminator (D), where G tries to fool D with the images it generates and D tries to beat G by knowing the image G generated was fake. Since D is the ruler of the game, and G only gets better by the feedback of D, the authors argue that GAN is an unfair game in which D most likely dominates. Moreover, the capability of the discriminator usually determines the generator's performance.

In order to create fairer competition, and therefore better results, the authors decide to make G assign D a task as well. They therefore incorporate Generator-Leading Task (GLeaD) to GANs. In this method, D is assigned to extract representative features that G can later decode and use to generate better images. D is forced to extract as much information from the image as possible, preventing its laziness and forcing D to focus on the entire image like G. The performance of D is measured by Reconstruction loss, which gives feedback from the difference between the input image and the reconstructed image, and derives gradients.

This method is determined to improve the FID of StyleGAN2 in three different datasets significantly (FFHQ, LSUN Bedroom, LSUN Church). It also improves the precision and recall in two of them (mainly large datasets). The "realness scores" are the probability that D gives G for the generated image to be real. The curves of realness scores provided in the paper indicate that the discriminator gives more balanced realness scores with the implementation of GleaD.

An obvious limitation of the implementation of GleaD is that along with improving the metrics most of the time, it increases computational cost because of the implementation of an additional loss function and the feature extraction process of the discriminator. Therefore this method introduces a new trade-off.

F. Paper 6: DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis [7]

The authors indicate that current text-to-speech GANs have obvious problems. Their stacked architecture causes entanglements between different generators that have different image scales. The existing studies take the route of applying and fixing extra networks in the learning process for text-image consistency. And this limits the supervision possibilities. The authors also suggest that those models use a cross-modal attention-based text-image fusion, and that is very computationally expensive.

The paper proposes Deep Fusion GANs, which they argue are more effective and simpler. DF-GANs have a one-stage text-to-image backbone that directly synthesizes high-res images, aiming to solve the problem of entanglements between generators. A target-aware discriminator is proposed which enhances the text-image semantic consistency without introducing extra networks. For this discriminator, a new loss function the matching-aware zero-centered gradient penalty is introduced, which is used by a regularization on the discriminator that can enable better convergence. A new deep text-image fusion block is also proposed. It is used for completely and effectively fusing the text and visual features.

The authors test their model in CUB and COCO datasets against SOTA methods like StackGAN and DAE-GAN. The results show that DF-GAN has great FID scores and competitive IS scores while having significantly fewer parameters (NoP) than other models. The authors also note that DF-GAN is observed to capture more fine-grained details compared to other models in qualitative evaluation.

An important limitation of the model is that it only considers sentence-level text information, preventing more detailed textual descriptions as input for the model. This impacts the usability of the model in areas where the detailed description is crucial. The authors also emphasize that additional knowledge from pre-trained LLM's might increase the performance of the model, the model is not published as the final version.

III. Example Images from MI-GAN Algorithm

Even though all of the papers I have selected include GANs, their results are mostly not comparable since they propose different things to be used in different areas. Paper 1 introduces that StyleGAN can be used in gathering intrinsic images, paper 2 does face swapping, paper 4 introduces a lightweight image inpainting model, etc. Therefore, the model with the "best score" cannot be determined. Among the papers that have their code available, I have chosen to demonstrate the MI-GAN algorithm since the paper and the code are well-structured and documented, and displaying an image inpainting algorithm is exciting. The example images provided by the developers contain various examples from the FFHQ dataset, and the Places2 dataset (256 and 512). I have provided one of the examples and its output in Figure 1.

The examples seem to be different kinds of images, but in most of them, isolated images of scenery are removed. Therefore the model performs very well, as expected.

Fig. 1: Example image and output of the MI-GAN.

IV. Cases Where The Algorithm Fails

In order to test MI-GAN with my own image inputs, as suggested by the developers, I installed IOPaint from pip which is an open-source inpainting tool that has a command line option for running MI-GAN with an easy-to-use interface. This enabled me to conveniently try many cases for the MI-GAN algorithm, setting masks for the images in an intuitive way.

Even though the algorithm performs very well with the deletion of isolated images, the algorithm has a hard time when an object that has connections with another object is deleted. The example below in Figure 2 demonstrates that. When the woman figure skater is removed by masking in the picture, the algorithm cannot complete a realistic image.

Fig. 2: Image with closely connected objects and MI-GAN's output.

The case in which I have seen the algorithm most miserably fail was the complex pattern/texture images. This is expected since most of the inpainting algorithms struggle to capture the intricate patterns of those kinds of images. Here is an example image that I have tried to test for this case, in Figure 3, where a disc is masked from the middle part of the image.

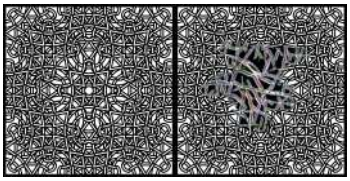

Fig. 3: Image with complex pattern/texture and MI-GAN's output.

As indicated in the paper, the algorithm also fails to generate realistic results with complex 3D inputs. An example image that I used is shown below, in Figure 4.

Fig. 4: Image with complex 3D objects and MI-GAN's output.

These are the cases where most of the other image inpainting models also fail, since capturing the patterns between extremely closely connected, intertwined objects and capturing complex patterns in textures is an extremely hard task even for GANs. However, there are also reasons specific to MI-GAN. For example, the depthwise separable convolutions used in the model have a hard time capturing complex spatial dependencies. Model re-parametrization also reduces the number of parameters in the model, even though it aims to keep most of the information. Therefore, the trade-off between making the model lightweight and capturing complex patterns is also observed.

V. Conclusion

In conclusion, this paper critically examines 6 other papers published in well-known conferences while conducting a literature survey on Generative Adversarial Networks. This paper contributes to a deeper understanding of the possible applications and developments of GANs in diverse areas, shedding light on the current landscape of GAN research. The listed propositions and limitations indicate the possible developments and improvement areas. The detailed analysis of MI-GAN also provides more information about both GANs and image inpainting tasks, indicating the cases where the algorithm fails, and addressing possible ways of improvement.

References

[1] A. Bhattad, D. McKee, D. Hoiem, and D. Forsyth, "Stylegan knows normal, depth, albedo, and more," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] H. Barrow and J. Tenenbaum, "Recovering intrinsic scene characteristics," *Comput. Vis. Syst.*, vol. 2, no. 3-26, pp. 2, 1978.

[3] Z. Liu, M. Li, Y. Zhang, C. Wang, Q. Zhang, J. Wang, and Y. Nie, "Fine-grained face swapping via regional GAN inversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8578-8587.

[4] A. Alanov, V. Titov, M. Nakhodnov, and D. Vetrov, "Styledomain: Efficient and lightweight parameterizations of StyleGAN for one-shot and few-shot domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2184-2194.

[5] A. Sargsyan, S. Navasardyan, X. Xu, and H. Shi, "Mi-GAN: A simple baseline for image inpainting on mobile devices," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7335-7345.

[6] Q. Bai, C. Yang, Y. Xu, X. Liu, Y. Yang, and Y. Shen, "Glead: Improving GANs with a generator-leading task," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12094-12104.

[7] M. Tao, H. Tang, F. Wu, X. Y. Jing, B. K. Bao, and C. Xu, "DF-GAN: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16515-16525.