Using GANs to Create Art: Final Report

Adil Zhiyenbayev Rakhat Abdrakhmamov Olzhas Zhangeldinov Aida Eduard

I. Introduction

In this competition, the world of art and data science converge through the power of Generative Adversarial Networks (GANs). Artists' unique styles, characterized by their use of color, brush strokes, and creativity, can now be replicated using algorithms. So, our task is to convert images in their style to Claude Monet style artwork. We will need computer vision and GAN to generate a substantial number of Monet-style images, ranging from 7,000 to 10,000. This involves two key neural networks: a generator and a discriminator. Submission is assessed using a metric known as MiFID (Memorization-informed Fréchet Inception Distance), a variation of the Fréchet Inception Distance (FID). The quality of your generated images improves as the MiFID value decreases. The submission should include 7,000-10,000 Monet-style images that are in jpg format. Their sizes should be 256x256x3 (RGB) [1]. Regarding the research component of the project, art generation using GANs is one of the prospective fields in AI. For now, the purpose of this work will be to test the sota GAN models on the dataset and try to improve the results.

II. Literature Review

The papers in this review were collected similar to our task such as style transfer, especially those models that are able to transfer Monet style to usual picture. Also we aggregated different evaluation metrics used to compare existing models with each other.

The initial GAN is presented in [2], in which it is intended to teach a generator how to accurately represent the distribution of real data. To do this, an adversarial discriminator that is subject to evolution is incorporated in order to differentiate between real data and data that has been artificially manufactured. Soon after, [3] considered potential of conditional adversarial networks as a versatile solution for image-to-image translation problems, discussing their ability to learn mapping functions and loss formulations, and its implications for the community. The authors demonstrated the effectiveness of conditional GANs as a general-purpose solution for image-to-image translation problems, showcasing the versatility of conditional Generative Adversarial Networks (cGANs) by presenting successful applications in synthesizing photos from label maps, reconstructing objects from edge maps, and colorizing images, among other tasks. The proposed method utilized a cGAN combined with a U-Net architecture. The discriminator operates with a 70x70 receptive field, achieving an impressive 66% per pixel accuracy. In specific tasks like colorization, the approach achieved a notable accuracy of 22.5% on the Amazon Mechanical Turk (AMT) dataset, and for maps-to-aerial translation, an accuracy of 18.9% is attained. These results demonstrated the effectiveness and accuracy of the model in diverse image-to-image translation tasks.

[4] presented cycle-consistent adversarial networks (CycleGAN) that incorporated cycled functions and adversarial training to achieve effective translation between domains without paired training data. It translates between domains in both directions using two generators and discriminators. The addition of cycle-consistency loss, which guarantees reversible translations, is the distinctive characteristic. This makes the network adaptable for tasks like style transfer by enabling it to learn mappings between different domains. The approach was validated in diverse tasks such as collection style transfer, object transfiguration, season transfer, and photo enhancement. In assessing realism, aerial photos translated to maps and vice versa were evaluated through AMT, with approximately 27% and 23% of the translations being scored as real. The approach employs fully convolutional networks (FCNs) to generate realistic photos from segmented labeled photos, achieving a per-pixel accuracy of 0.52. Additionally, the method showcases high accuracy (0.74 per pixel) and per-class IOU accuracies for segmentation of photos. Notably, the method effectively preserves style in the generated images, generally disregarding original details, with 5% of scenery correctly classified by CNN and 97.2% of correctly classified artist styles [5].

Hicsonmez et al. propose a new generator network and architectural modifications, achieving a better balance between style and content [5]. They introduced Generative Adversarial Networks for image to illustration (GANILLA) translation that incorporates unique architectural choices, such as the use of residual layers with concatenative connections and upsampling operations, and the ablation studies provide valuable insights into the role of low-level features in the downsampling and upsampling components of the network. These design considerations collectively contribute to Ganilla's approach to achieving high-quality image-to-image translations. They curate an extensive dataset of nearly 9500 illustrations from 24 different artists to facilitate training and evaluation. Their evaluation framework involves training CNNs to classify style (artists) and content (scenery). These CNNs were trained on unused data and then employed to classify the generated images. Then the accuracy of correct guesses of style and content was considered a quantitative measure of style and content preservation. The evaluation framework indicated that the proposed model successfully preserved a moderate amount of style and detail in the generated illustrations. The CNNs correctly classified 20% of images' content and 84.5% of style.

Other authors presented a novel dual Generative Adversarial Networks (DualGANs) mechanism that leveraged two generators for unlabeled images, achieving performance comparable to models trained with labeled data while preserving intricate details [6]. An AMT fake test was conducted to evaluate translation realism, resulting in a 45% score indicating perceived authenticity when changing the material of an object in a photo, e.g., plastic to metal. Furthermore, a realness score assessment involving translations of a face sketch to a realistic photo, a topographic map to an aerial photo, a day setting to a night setting, and a labeled segmented photo to the original facade of a building The translations yielded AMT realness scores ranging from 1.87 to 2.52. The CNN evaluation framework [4] measured 57.5% of content preservation and 46.5% of stylistic accuracy. This model showcased remarkable potential for unsupervised image translation, offering a promising avenue in image processing and computer vision.

Overall, the comparisons of works are presented in tables I and II. Since there is no prior evaluation system to compare the capabilities of the transfer style task or image-to-image translation, different evaluation methods were proposed, like the AMT metric for qualitative measure and FCN for content and style accuracy. From table I, in the context of map-to-aerial translation, cGAN outperforms other methods with a notably low score of 6.1%, indicating successful translation from map to aerial imagery. cGAN also excels in the aerial-to-map translation task, achieving an 18.9% score. DualGAN performs well with a 42.0% score. In label-to-photo task and photo-to-label, cGAN again demonstrates superiority of 66.0% score and 74%, followed by CycleGAN at 52.0%. In the table II, CycleGAN excels in style preservation with a high style percentage of 97.2%, but its content preservation is comparatively lower at 5.2%. DualGAN demonstrates a balanced performance, achieving a significantly higher content percentage of 57.5%, though the style preservation is at 46.5%, while GANILLA strikes has a lower content percentage of 20.0% and higher style preservation at 84.5%. Ganilla shows the best results compared to other GANs. Despite the fact that CycleGAN shows the highest results in style transfer, the content accuracy is still low, resulting in blurred and missed details of the photo. DualGAN shows average results in style and content.

III. Baseline

As a baseline for the project, we chose the work based on the DualGAN, as it shows one of the best performances among the other works on Kaggle. The architecture consists of three convolutional and several residual blocks, two fractionally-strided convolutional layers (with stride 1/2) and one that performs mapping to RGB [4].

Figure 1 shows the examples of Monet-esque generation using input photos. It can be seen that the generation is already pretty accurate.

TABLE I
Comparison Of SOTA Works AMT And FCN

Lit.	Method	AMT,%	FCN,%
[3]	cGAN	MA:6.1 AM:18.9	LP:66.0 PL:74.0
[4]	CycleGAN	MA:23.0 AM:27.0	LP:52.0 PL:58.0
[6]	DualGAN	AM:42.0	-
[2]	GAN	AM:41.0	LP:22.0

The table represents the comparison of the methods on the image-to-image translation tasks like map-to-Aerial photo (MA), Aerial-to-map (AM), label-to-photo (LP) and photo-to-label (PL) evaluated on the AMT and FCN score. The models were trained on these datasets Google maps and city scrapes [7].

TABLE II
Comparison Of SOTA Works On Content And Style

Lit.	Method	Content,%	Style,%
[4]	CycleGAN	5.2	97.2
[6]	DualGAN	57.5	46.5
[5]	GANILLA	20.0	84.5

The table represents the comparison of the methods on the style and content by style and content classifiers. The models were trained on the 10 illustration datasets depicted in [5].

IV. Methodology

Our goal is to improve the baseline by modifying existing GAN models. Possible improvements can be made in DualGAN and CycleGAN models. This section will describe hyperparameter tuning and model modifications that improved performance of the GANs.

A. DualGAN

The initial architecture of the DualGAN consists of two identical GANs. Generator from the first GAN maps photos to Monet styled images, and generator from the second GAN maps Monet styled images to photos. The generators consist of downsampling and upsampling layers. Lastly, there are two discriminators: the first one generates membership scores for generated Monet styled images, and the second one generates membership scores for the generated photos. There are also two losses: Monet loss and photo loss. Monet loss by generating Monet styled image and regenerating the original photo using the second generator. The same is done for photo loss but vice versa.

The first modification that we did is that we made the structure of the DualGAN's generators more complex. That is, we added more downsampling and upsampling layers to the architecture of both generators. The example of the inference can be seen in fig. 1

The second modification was motivated by the structure of the CycleGAN. We defined residual blocks and added them in between downsampling and upsampling layers. Initially, we planned to do this as an addition to the previous modification, however, the model became too complex and required a lot

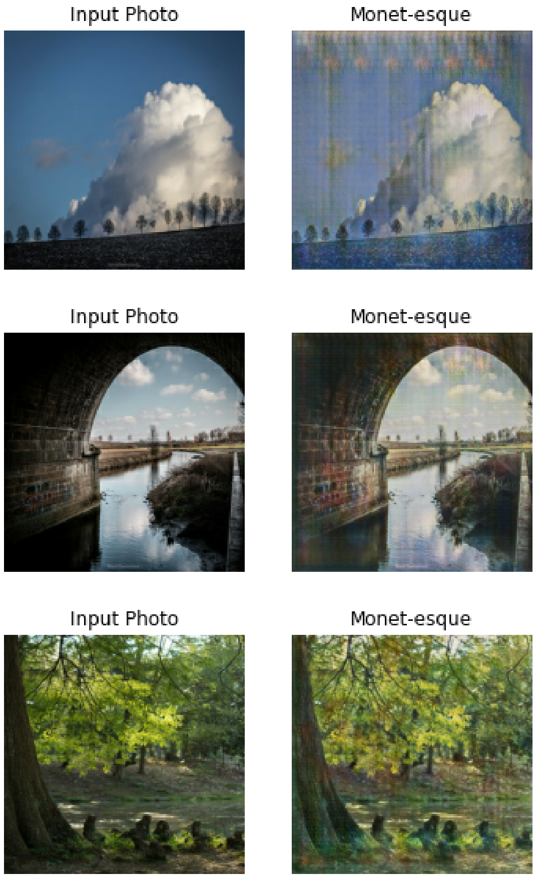

Fig. 1. DualGAN with improved architecture

of computational resources. Eventually, we had two different modified DualGAN models to compare.

B. CycleGAN

CycleGAN is another GAN that performs well with unpaired dataset. The generator model consists of 2 downsampling and upsampling layers connected with 9 RESNet blocks. The discriminator is a CNN with three downsampling layers with 64 filters and 4x4 kernel, finished by an output layer of dimension 32x32 and one channel.

First, we added positive identity loss proposed by [8]. This identity loss is a mean absolute error of a image $x \in A$ with generated image $G_{B \longrightarrow A}(x)$. This loss keeps identity of the original image and prevents color changing in the generator. The weight of the identity loss is 10, while weight of cycle consistency loss is also 10.

Next, we replaced RESNet blocks in the generator with Inception blocks. Inception blocks take less space and computational power which is a viable solution for our limited computing power. We added 9 Inception blocks with 5x5, 3x3, and 1x1 kernels each of size 128. Training CycleGAN with RESNet for 20 epochs took about 10 hours, while with Inception blocks it took only 8 hours.

C. XGAN

One of the hypothesis was to test the performance of XGAN - Cross-GAN model [9]. The idea behind this architecture was to transform images from one domain to a complete different domain while preserving certain features of the original image. The XGAN model contains a generator, which is a dual adversarial auto-encoders with shared layers for determining the semantic context of the images. Discriminator is a 4-layered CNN network, that determines the realness of the generated image.

The XGAN model takes input as 64x64 pixel images, so the input photos were center cropped down to this shape. The model was trained on the competition's dataset for about 400 epochs. The results were unsatisfying - generated images had no details of any of the domains space, only the approximate shapes of the real photos.

V. Evalutaion

We evaluated our models both qualitatively and quantitatively. Quantitative evaluation uses Fréchet inception distance (FID). This metric is computed using pretrained Inception v3. For each set of images, generated and original, it computes Inception's last feature extraction layer values and fits Gaussian distribution over those features. Then, it computes Fréchet distance between those distributions. Lower the distance between distributions means generation of more realistic images. This metric is a good score for evaluating unpaired datasets since we can not compute pixel or patch accuracy on original images.

A. Quantitative Results

Baseline DualGAN provided FID of 140 after training for 1000 epochs. DualGAN with ResNet layers provided same result after training only for 400 epochs. More complex DualGAN results in FID of 130. CycleGAN The best FID score of 109 was obtained with CycleGAN with Inception blocks. Obtained results are summarized with int the Table III

TABLE III
FID comparison of models.

	DualGAN (1000 epochs)	**DualGAN more complex (25 epochs)**	**DualGAN with ResNet (400 epochs)**	**CycleGAN with Identity Loss**	**CycleGAN with Inception blocks**
FID	140	130	140	110	109

B. Qualitative Results

We evaluate generated images based on their resemblance of Monet painting style and general patterns that the generator seems to follow.

Figure 2 illustrates difference between generated images of the best model improvements we have achieved.

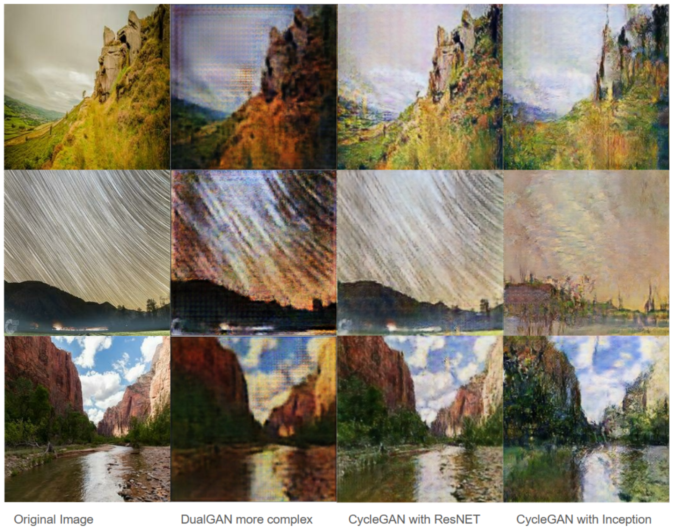

Fig. 2. Comparison of generated images by DualGAN, CycleGAN with ResNet, and CycleGAN with Inception

VI. Conclusion and future work

We compared different GAN models on their ability to transfer Monet painting style onto real-world photos. The baseline for this report is DualGAN model that showed similar or better results compared to other GANs. Then, we tested possible improvements of the models by adding more layers and embedding DualGAN with ResNet blocks. We also trained CycleGAN on the dataset and improved it by adding identity loss to prevent colouring and Inception blocks to simplify the model and speed training up. We used FID to evaluate our models quantitatively, however, different GAN architectures (CycleGAN and DualGAN variants) were trained for a different amount of time on different machines because of limited computational power and time given for the project. Although CycleGAN showed better results, it may be because it was trained for a longer time on a more powerful machine. Nevertheless, we could improve DualGAN generator by adding more convolution layers and inserting ResNet blocks between encoder and decoder layers. CycleGAN model was improved by adjusting identity loss weight and replacing ResNet blocks with Inception blocks.

As a future work, we would like to implement more recent GAN models for unpaired datasets, such as GANILLA [5].

References

[1] "I'm Something of a Painter Myself — kaggle.com," https://www.kaggle.com/competitions/gan-getting-started/overview, 2023, [Accessed 24-11-2023].

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[5] S. Hicsonmez, N. Samet, E. Akbas, and P. Duygulu, "Ganilla: Generative adversarial networks for image to illustration translation," *Image and Vision Computing*, vol. 95, p. 103886, 2020.

[6] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[8] S. Liu, "Study for identity losses in image-to-image domain translation with cycle-consistent generative adversarial network," in *Journal of Physics: Conference Series*, vol. 2400, no. 1. IOP Publishing, 2022, p. 012030.

[9] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy, *XGAN: Unsupervised Image-to-Image Translation for Many-to-Many Mappings*. Cham: Springer International Publishing, 2020, pp. 33–49. [Online]. Available: https://doi.org/10.1007/978-3-030-30671-7$_3$