Spacewarp: aesthetic space style transformation

Jacopo Ferro, Jonathan Haymoz, supervised by Martin Nicolas Everaert
CS-413 Computational Photography - EPFL

***Abstract*—In the project "Spacewarp: Aesthetic Space Style Transformation," we aimed to generate a synthetic dataset and train an image-to-image translation model to transform photographs into space-themed images. We deviated from the suggested pipeline of using the Prompt-to-Prompt method for data generation and the U-Net and Pix2Pix architectures for image-to-image translation training. Instead, we employed the Instruct-Pix2Pix method for data generation, used Language Models (LLM) for prompt engineering, and utilized the Unsplash-25K dataset for real images. We then trained from scratch the U-Net and Pix2Pix architectures for image-to-image translation training on a filtered dataset to complete the practical implementation.**

Although our pipeline, if expanded, could achieve impressive results, we faced limitations due to training the U-Net model from scratch with only 1K images. However, the primary purpose of the project was to gain a deeper understanding of diffusion models, style transfer, and the methodologies presented in the referenced papers, and to gain practical end-to-end experience. We thus propose a more streamlined pipelined that allows to take advantage of the state-of-the-art advancements in diffusion models. We could not fully implement these propositions due to time constraints but we believe this to be the basis for state-of-the-art custom style transfer which generalizes to an extremely wide range of visual content. In synthesis, the proposed pipeline is the following: create or gather a few examples of your custom style transfer, subject, or concept. Fine-tune a text-to-image diffusion model with a technique like LoRA to leverage the knowledge of the model on top of your examples. Create a small dataset of pairs (image, image with custom style) with a technique like prompt-to-prompt by freezing the cross-attention layer. Use the obtained dataset to fine-tune an image-to-image diffusion model like instruct-Pix2Pix instead of training a GAN specific to the custom style in order to leverage the previous training.

I. Introduction

In the "Spacewarp: Aesthetic Space Style Transformation" project, we explored the use of advanced diffusion models and image-to-image translation models to transform photographs into images depicting potential space-themed states of the scene. We generated a synthetic dataset using the Instruct-Pix2Pix method, which offered several advantages over the Prompt-to-Prompt method.

To guide the image generation process, we used prompt engineering with GPT-3.5 and GPT-4, creating prompts that instructed the model to set the subject of the image against a deep space background while maintaining the subject's original colors and shapes. We then trained an image-to-image translation model using the synthetic dataset, with the model architecture based on the U-Net model proposed by pix2pix.

The project highlighted the potential of these advanced techniques for image transformation and provided valuable insights into the challenges and considerations involved in generating high-quality synthetic datasets and training effective image-to-image translation models. The deliverables of the project included the code, a written report explaining the literature and steps taken for the project, and the data and datasets that were used. We presented the project on the 2nd of June 2023 at EPFL in the context of the CS-413 Computational Photography course.

II. Dataset Generation

The objective of our data generation process was to create a synthetic dataset consisting of pairs of images: an original image and its corresponding space-themed transformation. This dataset would serve as the training set for our image-to-image translation model.

A. Choice of Instruct-Pix2Pix over Prompt-to-Prompt

Our initial consideration was to employ the Prompt-to-Prompt method for generating our synthetic dataset. However, upon rigorous experimentation and analysis, we opted for the Instruct-Pix2Pix method. This decision was driven by several factors:

Efficiency: Instruct-Pix2Pix demonstrated faster performance in generating synthetic images compared to Prompt-to-Prompt. **Advanced Implementation:** Instruct-Pix2Pix is built on top of the Prompt-to-Prompt method, incorporating its strengths and addressing its limitations. This resulted in better implementations and more desirable synthetic images. **Real Image Input:** Unlike Prompt-to-Prompt, Instruct-Pix2Pix can take a real image as input and transform it, eliminating the need to generate both images synthetically. This feature was particularly beneficial for our project, as it allowed us to maintain the authenticity of the original image while creating a convincing space-themed transformation.

B. Prompt Engineering with GPT-3.5 and GPT-4

To guide the Instruct-Pix2Pix model in generating the synthetic images, we employed prompt engineering with Language Models (LLM), specifically GPT-3.5 and GPT-4. This process involved transforming an image description into an instruction for a space-themed transformation specific to each image.

However, we observed certain limitations with the Instruct-Pix2Pix model. It seemed to focus more on specific trigger words rather than comprehending the intricate details of the fine-tuned prompt. We also noted that the choice of prompt words and the seed had a disproportionate influence on the quality of the generated images.

To mitigate these limitations, we generated approximately 15k images using more generic prompts such as ”Space”, ”Cosmic Space”, and ”Place in deep space”. We generated 6 to 8 versions for every image to ensure diversity in our dataset.

C. Initial Dataset and Filtering

Our initial dataset was a subset of the Unsplash-25k dataset, which consists of high-quality photos of real-world subjects and scenery. We downsized these images to 512x512 to make them suitable for our model.

We then applied a filtering process to this initial dataset based on aesthetic considerations. We aimed to select images that would lend themselves well to a space-themed transformation. This filtering process resulted in two datasets: a larger dataset of around 1k image pairs and a smaller, more refined dataset of 150 image pairs.

In conclusion, our dataset generation process involved a combination of advanced techniques and careful selection to create a high-quality synthetic dataset for our image-to-image translation model. This dataset was instrumental in the success of our project, as it allowed our model to learn the desired style transformation effectively.

III. Image-to-Image Model

Now that we have our datasets, we can focus on the training of our image-to-image translation model. For our model, we choose to follow the work of our predecessors and use the Pix2Pix conditional GAN model. Following their paper, we can choose the best configuration for each of our parameters. It helped us a lot as with the limitation of our computational power, it may more sense to ”sit” on the shoulder of our predecessor which tried multiple architectures and parameters. It follows the structure of a traditional GAN model, with the adding the source image to the discriminator input. It can then have the before and after for both generated and real data. Here you can see the structure with an example of 1

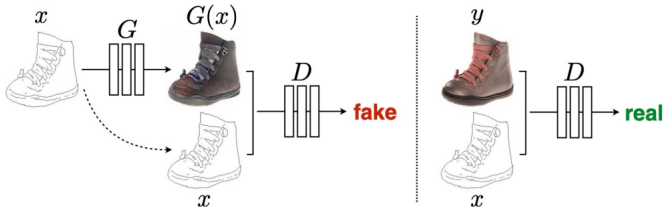

Fig. 1: Pix2Pix conditional GAN structure

A. Generator

The Generator has a classical U-Net structure. It has a Encoder-Decoder structure with the particularity of keeping a copy of the image on the down-sampling path and concatenate them with the results on the up-sampling path. This is not an heavy U-net, meaning that there is less convolutions than usual, such that the computational time does not explode. 2

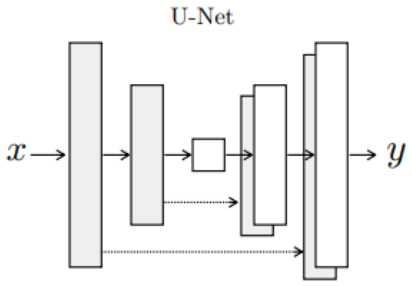

Fig. 2: The Unet

In the Appendix you can find a detailed table about the Generator.

B. Discriminator

The discriminator is a convolutional network called a Patch-Gan. The model tests on a 70x70 patch of the image and decide if the patch is a valid image transfer or not. It also takes the source image in order to have a greater context about the transformation. A weighting 1/2 is applied such that the model’s effect are reduced. The optimization is done with a binary cross-entropy loss.

Once again, the structure of the convolutional network is presented in the appendix.

IV. Results

A. Quantity vs Quality

First, we wanted to understand the impact of the number of pairs used in the training. To do so we filtered the lists and created 2 datasets:

- **153 Images:** The good quality images which were the ”perfect” results.
- **1000 Images:** Which is composed of all the images that were from ok to perfec, including the 153 perfect from the other dataset.

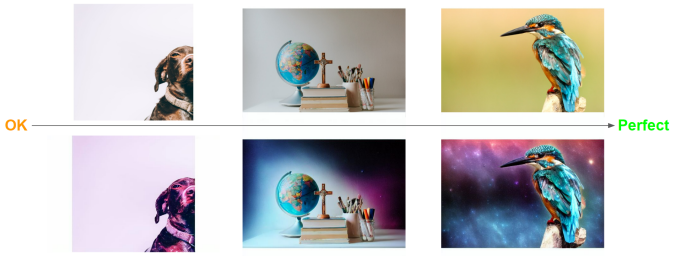

Fig. 3: the filtering process

With these datasets we trained the model for:

- **153 Images:** 100 epochs
- **1000 Images:** 20epochs

Which gives use the results in the fig.4

With multiple different images, we arrive at the same conclusion: with around the same number of steps (15’000

Fig. 4: results for the prediction of the 2 datasets. Original (left), with the trained model on 153 images (middle) and the one with 1000 images (right)

for the 153 images, 20'000 for the 1000 ones), we get the same kind of results. The color is quite nice and does look a little bit more like we wanted. However, there is no details; we do no see the creation of details such as stars. This may come from our definition of our problem as we explain in the next section.

B. Insights

As we see the similar results in the 153vs1000 images models, we want to have more detail. The model seems to be confused. We then emit the hypothesis that our "spacewarp" definition is not clear enough. If you look at the pictures from fig 5, you see that we have aesthetically pleasant pictures that could well go into the "spacewarp" definition.

Fig. 5: 3 different images

However, we got here 3 different proposition to the cGAN:respectively, put the background into space and put a nice effect on the whole image. To reduce this confusion, we decide to create a new dataset with only the first case: a clear front with a spacewarped background. We then do not get enough data to really train our model, so we do a data augmentation by mirroring the pictures horizontally. In the end we get 270 images, trained it for 60 epochs (16'000 steps) and get the results in fig 6

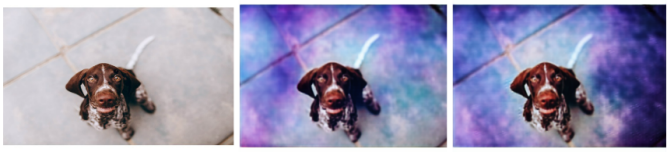

Fig. 6: orignal(left), only the chosen ones (middle) and the mirrored chosen (right) after training

The hypothesis seemed wrong as we have slightly better results overall but not a big enough difference.

V. Conclusion

In conclusion, we did not reach a "perfect" model that gives us good enough results. The last model with a more defined problem is the one with the most promising results but is limited to its definition. We suppose that the similarity between the results comes from the fact that the smaller datasets are actually subsets of the bigger ones. The pairs that we rate the highest have the biggest change usually and they are the one creating the most change in the model, hence the similarities. It would be interesting with more computational power to generate more correct data from the start and/or to tune the parameters of the models. As an example, our model do not manage to create details such a stars, therefore it could interesting to reduce the patch size from the discriminator.
We however learned a lot about both tasks of the project and even though we are not deligthed by the results, we are still happy enough about the journey.

References

[1] T. Brooks, A. Holynski, and A. A. Efros, "Instruct-pix2pix: Learning to follow image editing instructions," arXiv preprint arXiv:2211.09800, 2022.

[2] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," arXiv preprint arXiv:2208.01626, 2022.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241. Springer, 2015.

[5] Placeholder for LoRA paper.

[6] Placeholder for GPT paper.

[7] Jason Brownlee, "How to Develop a Pix2Pix GAN for Image-to-Image Translation" url = "https://machinelearningmastery.com/how-to-develop-a-pix2pix-gan-for-image-to-image-translation/", las accessed 01.06.2023

[8] Phillip Isola, Jun-Yan Zhu Tinghui, Zhou Alexei, A. Efros, " Image-to-Image Translation with Conditional Adversarial Networks", jornal: arXiv:1611.0700v3, 2018

Acknowledgements

We thank Martin Nicolas Everaert for guidance and supervision, Prof. Sabine Süsstrunk for the valuable lectures, and the teaching team.

Appendix

A. LoRA and considerations

During the project, we discovered more about LoRA fine-tuning applied to Stable Diffusion. Ideally, we demonstrated how using Prompt-to-Prompt on a fine-tuned text-to-image diffusion model with added LoRA, which is knowledgeable in generating planets, space, and galaxies, can then fine-tune the Instruct-Pix2Pix model with LoRA to generalize the specific Spacewarp style.

This approach eliminates the need to train a different GAN from scratch for every style transfer. Instead, we can add to larger models, fine-tune on small datasets, and leverage the knowledge of the model. A GAN that has never seen good galaxies and space will never be able to create them, emphasizing that GANs are only as good as the data we feed them.

A few years ago, the state-of-the-art datasets and styles were present in the real world, such as artists, paintings, drawings, mangas, and cartoons. Now, we can invent extremely specific styles. We should build on top of this instead of training from scratch, as we can only approximate the synthetic generation exactly how a GAN could only approximate a Studio Ghibli or Van Gogh style a few years ago.

Next Steps

In the future, we plan to expand our pipeline to include more images and leverage the power of larger models. We believe that by fine-tuning on small datasets and leveraging the knowledge of the model, we can create more specific styles and improve the quality of our image transformations.

We also plan to explore the use of LoRA fine-tuning applied to Stable Diffusion in more depth. Although we did not fully implement this in our current project we believe it holds great potential for enhancing the performance of our image-to-image translation models.

B. Generator

The activation function for the encoder blocks is a LeakyReLu and the one for the decoder is the ReLu.

Block	Conv	Batchnorm	Dropout
encoder 1	C64	/	/
encoder 2	C128	yes	/
encoder 3	C256	yes	/
encoder 4	C512	yes	/
encoder 5	C512	yes	/
encoder 6	C512	yes	/
encoder 7	C512	yes	/
Bottleneck	C512	/	/
decoder 7	C512	yes	yes
decoder 6	C512	yes	yes
decoder 5	C512	yes	yes
decoder 4	C512	yes	/
decoder 3	C256	yes	/
decoder 2	C128	yes	/

C. Discriminator

All the layers are then activated with a LeakyReLu and the last one with a sigmoid function.

Block	Conv	Batchnorm	Dropout
1	C64	/	/
2	C128	yes	/
3	C256	yes	/
4	C512	yes	/
5	C512	yes	/
6	C512	/	/