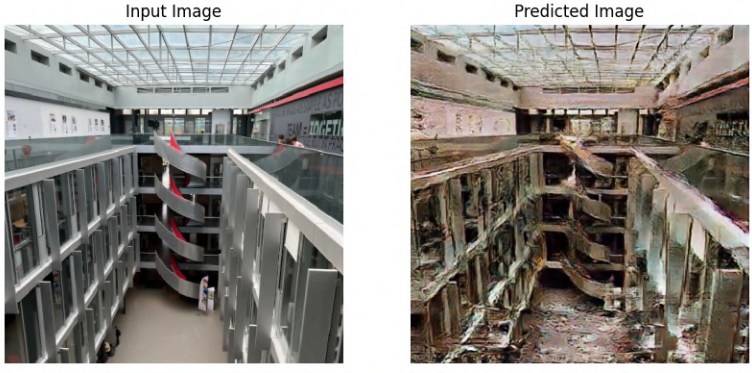
TimeWarp: How would this scene look in 100 years?

Marino Müller, Juliette Parchet, Camille Montemagni

EPFL - CS413 - Computational Photography
Supervised by Martin Everaert
and Sabine Süsstrunk

October 2, 2025



Abstract. We present a model designed to transform a photograph of a scene into a new photograph representing what this scene might look like after it was abandoned for 100 years. First, we discuss a procedure for generating a synthetic dataset using state-of-the-art diffusion models. Then we describe the process of training and optimizing an image-to-image translation model.
Please have a look at our GitHub repository[1] to find our code, notebooks, datasets.

Keywords: image-to-image, prompt-to-prompt, synthetic data, data generation, prompt engineering, prompt generation

1 Introduction

The goal of this project is to train a model so that it can transform a photograph of a scene into a new photograph of what this scene might look like after being abandoned for a hundred years. This goal notably implies the generation of a relevant synthetic training set, as well as the training of an image-to-image

[1] https://github.com/Jucifer06/CS413-Computational-Photography.git

translation model. A lot of computer vision and image processing problems, such as image synthesis, segmentation, style transfer, restoration, and pose estimation, come down to an image translation problem [4]. Synthetic data generation is also an important part of image processing, and brings its own set of problems and requirements as a basis for the implemented solutions. This projects thus draws its interest from these two fundamental problems of image manipulation.

Existing solutions for the image to image translation problem include notably generalist image edition models like InstructPix2Pix[1], Midjourney and Dall-e. Below is a review of generations outputed by InstructPix2Pix and Dall-E aimed at solving this paper's problem. The results produced by InstructPix2Pix clearly lack an actual understanding of the represented scene. Fig. 1 and Fig. 2 both only show changes in the colour tone of the initial photograph, without modifying any of the semantic properties of the image. Results produces by Dall-E are even less conclusive as there are no visible changes in both Fig. 3 and Fig. 4. Ideally, the model should modify vegetation so that it appears drier or more luxurious, the cars should be rusty and the surroundings should be negatively affected by time. In summary, the ideal model should be able to modify elements represented by the image, and not only high-level properties of it like colour tone.

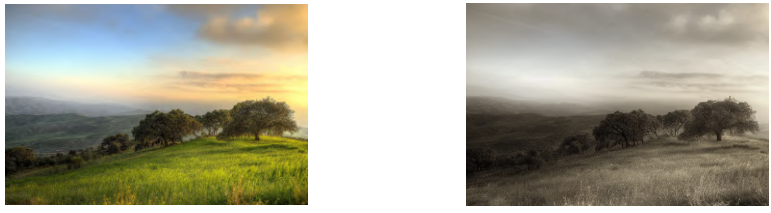
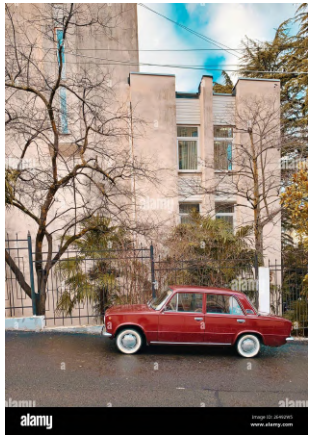


Fig. 1. InstructPix2Pix edition of a photograph of a scenery when given the prompt "the scene has been abandoned for 100 years"

Existing solution for the synthetic data generation problem include the use of "Prompt-to-prompt image editing with cross attention control"[2]. This model actually produces very good results as long as the entered prompts are relevant. We present examples in Fig. 5 and Fig. 6. The limitations of the model include suppressing crowds and people from a photograph, as shown in Fig. 7: a lot of artifacts are produced and the scale of the picture is deformed.

2 Literature Review

[4] presents an overview of the image to image translation works developed in recent years and analyses their key techniques. The two most representative and commonly adopted generative models appear to be variational autoencoders



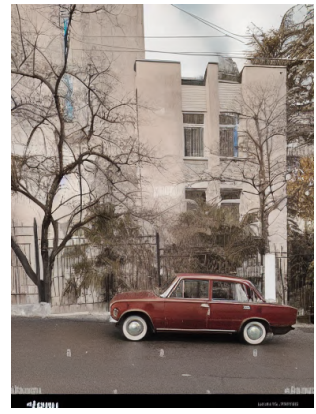


Fig. 2. InstructPix2Pix edition of a photograph of a car when given the prompt "the scene has been abandoned for 100 years"

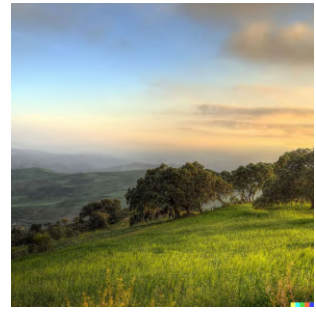
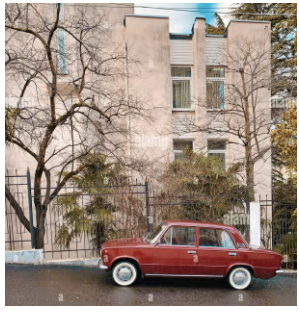


Fig. 3. Dall-e edition of a photograph of a scenery when given the prompt "tGenerate this image after the place has been abandoned for 100 years"



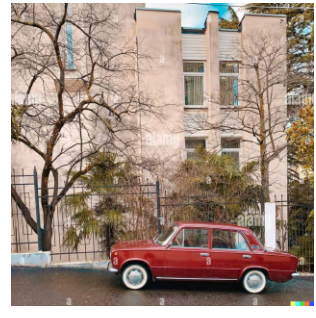


Fig. 4. Dall-e edition of a photograph of a car when given the prompt "Generate this scene after the place have been abandoned for 100 years"

Fig. 5. Prompt to Prompt generations when given the initial and modified prompts "A car, with modern city in the background." and "but the scene looks like it has been abandoned for 100 years"

Fig. 6. Prompt to Prompt generations when given the initial and modified prompts "A house, with modern city in the background." and "but the scene looks like it has been abandoned for 100 years"

Fig. 7. Prompt to Prompt generations when given the initial and modified prompts "A street crowded with people" and "but the scene looks like it has been abandoned for 100 years"

(VAEs) and generative adversarial networks (GANs). Both of these models construct a function $x = g(z)$ for generating the desired x from the latent variable z. A VAE models data distribution by maximizing the lower bound of the data log-likelihood, and a GAN tries to find the Nash equilibrium between a generator and discriminator. The paper also categorises the image to image translation problems into two sets of tasks: two-domain image to image translation and multi-domain image to image translation. For the problem this paper is trying to solve, only two-domain translation is relevant. [4] distinguishes supervised and unsupervised two-domain translation, noting that unsupervised is a good way to avoid having to acquire large, paired training data. As this paper aims to generate a synthetic dataset as well as a model, only supervised learning is relevant here. [4] cites pix2pix[3] as a strong baseline image translation framework that inspires many improved works based on it. A mentioned flaw of pix2pix is its inability to capture the complex scene structural relationships through a single translation network when the two domains have drastically different views and severe deformations. The proposed solution is to combine the multichannel attention selection module with GAN into a Selection GAN[7]. [4] also reviews a few application of image to image translation, notably image semantic manipulation in [8], which presents GAN-based model that uses the space of deep features learned by a pre-trained classification model. Looking into the methods cited by [4], pix2pix[3] presents conditional adversarial networks as a general-purpose solution to image-to-image translation problems. It achieves very conclusive results and plays its role as a general-purpose model.

3 Implementation

The TimeWarp project can be divided into two main tasks:

- **A synthetic dataset generation task**, which contains prompt engineering, and generation model optimisation.
- **An Image-to-Image translation training task**, which contains model selection and model training optimisation

3.1 Synthetic dataset generation

A first milestone of the TimeWarp project was to have a relevant synthetic datasets of images to train a model on. The *desiderata* of such a dataset is to have pairs of images, each consisting of an image representing a random and general scene, and another image representing the exact same scene as if it had been abandoned for a hundred years.

As suggested by the handout, the dataset generation was made using Prompt-to-Prompt image editing with cross attention control[2]. Prompt-to-Prompt takes as input a prompt and a modification of this prompt in order to output two images: one corresponding to the input prompt and another one corresponding to an edited version of the first image following the modification of the initial prompt. This idea is shown in Fig. 8.

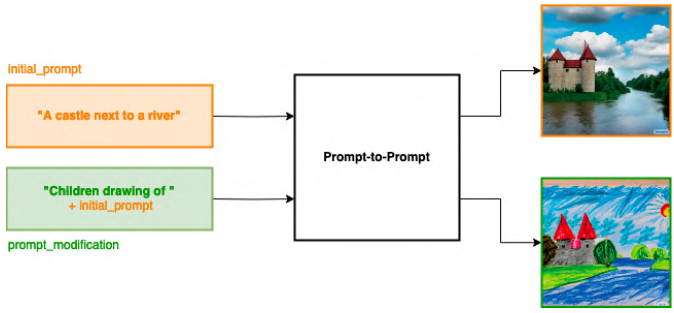


Fig. 8. Basic Prompt-to-Prompt operation. Images from [2]

Using Prompt-to-Prompt for data generation requires to first establish a dataset of initial prompts, and to engineer the relevant modification of these prompts in order to get the best output results.

Initial prompts A first idea for the initial prompts dataset was to use an existing one. DiffusionDB[9] is "the first large-scale text-to-image prompt dataset. It contains 14 million images generated by Stable Diffusion using prompts and hyperparameters specified by real users" and is publicly available. It rapidly came that using the prompts available in DiffusionDB would not be possible, as they are way too specific or complex. For example, "a cat wearing a fake mustache, rainbows, bright colors, psychedelic, fish eye lens, vector art, fantasy poster by helen huang and frank frazetta and salvador dali and norman rockwell, " is a randomly picked prompt from DiffusionDB, and contains a lot of elements that are not relevant for this project, namely the fish eye lens, the poster and psychedelic effect and the depiction of a cat as the main point of interest. Indeed, it does not make sense to render an unrealistic cat with complex light effects as if it had been abandoned for a hundred years. Lots of prompts from DiffusionDB have the same problems, and there is no easy way to sort the database in order to get the prompts that are relevant for this project. The idea of using DiffusionDB was thus abandoned. The second idea for the initial prompt dataset was then to create it from scratch. It has been decided to use a very general sentence structure and to modify certain words of it in order to get a relevant and diverse prompts dataset. The focus was mainly put on very general scenes with a wide angle, that did not include any animals and that included human figures only as a crowd. An example of such a prompt construction is shown in Fig. 9.

Modified prompts For the modified prompt engineering, the idea was to have a base that was general enough such that it can be applied to most of the images. Starting with the prompt "it has been abandoned for centuries", a first common feature of scenes that represent an abandoned place is the main colour palette in it. We thus added to our prompt modification "the scene is a bleached, washed-out colored post-apocalyptic photograph". A second feature that was general

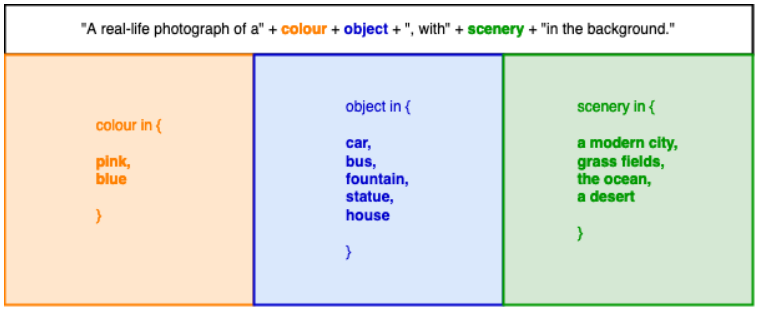


Fig. 9. Visualisation of an example prompts structures and keywords used to create the initial dataset.

enough to add was "The objects are rusty and trashed" or "the interior and furniture are dirty and decaying" depending on the content of the image. Another common feature of all scenes that represent abandoned places are statements about the vegetation, we thus added when relevant either "the vegetation is dry and dead" or "the vegetation is overgrown". Lastly, a common feature that we want on all our images is that they do not contain any human figure. We thus added to prompts that depicted a crowded area "The room is empty and there is nobody" or "movement has stopped and there is nobody". An example combination of all these ideas gives the following prompt: "The scene is a bleached washed-out colored post-apocalyptic photograph, it has been abandoned for centuries. The room is empty and there is nobody. The interior and furniture are dirty and decaying. The objects are rusty and trashed. The vegetation is dry and dead.".

Cross attention control Cross-attention allows for more precise control over the editing process of the input image. Attention weights between a query and key-value pairs are calculated, they determine the influence each key-value pair has on the query, enabling specific regions or features in the prompts to be selectively modified or transferred. Cross-attention enhances the model's ability to capture desired editing intent and generate more accurate outputs aligned with the provided prompts. This concept is shown in Fig. 10. Using cross-attention control allowed us to put more emphasis on specific words depending on the query, for example, queries involving people had more weight put on the word "nobody".

3.2 Image-to-Image translation training

The image-to-image translation task for the TimeWarp project requires a model which can take as input an image and output a version of this image as if the scene in it had been abandoned for a hundred years.

Fig. 10. By reducing or increasing the cross-attention of the specified words (marked with an arrow), we can control the extent to which it influences the generated image, taken from [2]

Choice of model Two models were considered for the image-to-image translation task: a Pix2Pix[3] based conditional GAN using U-Net[6] as the generator and a CycleGAN[10].

The conditional GAN architecture, shown in Fig. 11, contains a U-Net-based generator and a convolutional PatchGAN classifier (proposed in the pix2pix paper[3], shown in Fig. 12) as the discriminator. U-net from [6], shown on Fig. 13, "relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization" and the paper shows "that such a network can be trained end-to-end from very few images and outperforms the prior best method".

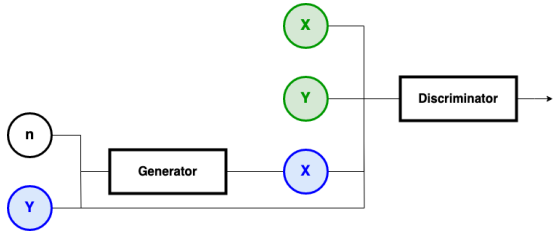


Fig. 11. Conditional GAN architecture. X and Y represent respectively the input and output images. The blue nodes are images generated by the generator and the green nodes are images from the dataset

CycleGAN is a type of Generative Adversarial Network (GAN) used for unpaired image-to-image translation. It learns the mapping between two different image domains without requiring corresponding images in the two domains to

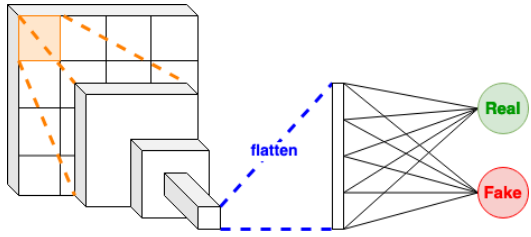


Fig. 12. Convolutional PatchGAN classifier process

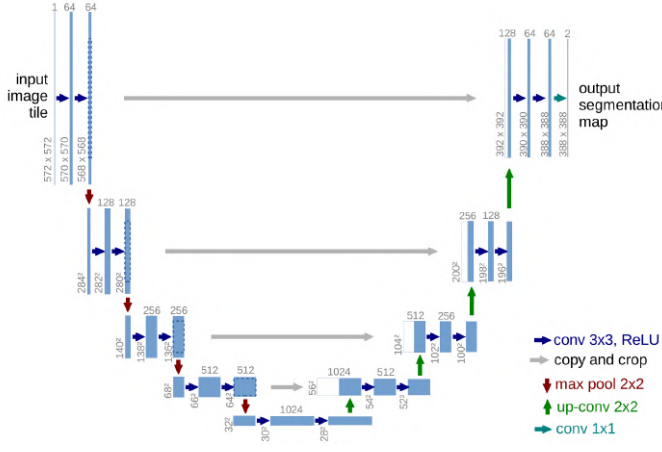


Fig. 13. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. Taken from [6]

be paired. CycleGAN pairs two GANs together, one for mapping images from domain A to domain B and the other for mapping images from domain B to domain A. By simultaneously training these two GANs the network learns to translate images from one domain to another in both directions, while also enforcing a "cycle consistency" constraint that helps ensure the translated images are plausible. We tried training a CycleGAN with the intent of ultimately using only one of these two GAN.

After testing and training both models with our dataset, we concluded that Pix2Pix with U-Net as the generator was the better option for our project. We managed to get visually more appealing predictions from it compared to the CycleGAN right from the start. We then decided to concentrate exclusively on the Pix2Pix model.

Dataset used for training In the beginning, we included in the training set, all the data that we generated[2] (cleaned a bit by hand to remove extremely bad generation). We had about 10'000 image pairs. There were various scenes, indoors, outdoors, nature-like, city-like, with/without humans, with text,... We tried to be quite exhaustive on real-like sceneries, in the hope to have in end a model that would perform well on a large scope of image-to-image translation.

Fig. 14. The images are both generated by Prompt-to-Prompt, on the top, the original image, and on the bottom the translated abandoned image

Prompt-to-Prompt, used to generate the dataset, did not work really well on several types of images, as we can observe in Figure 14. The model trained on this huge amount of mediocre data was not good: the images looked a bit older, but the model was not aging elements in the image, but the image itself. We can see in Figure 15 some kind of brown filter, but no real image translation.

[2] Have a look at the Data-Construction folder of the GitHub repository https://github.com/Jucifer06/CS413-Computational-Photography.git to know more about the dataset construction process

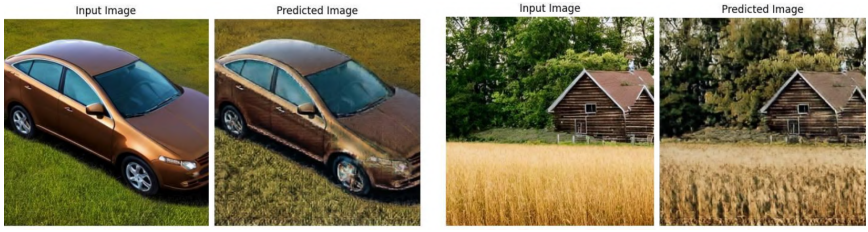


Fig. 15. Some results of the first version of our Pix2Pix model. On the left, are the original images, and on the left are the translated abandoned images

So what we did, was to reduce the scope of the training dataset, to include only the best generation Prompt-to-Prompt had to offer: the scenes with cars, and streets. We generated around 6'000 image pairs for the newly selected training dataset.

Fig. 16. Some generation of Prompt-to-Prompt, of cars and urban scenes

We hoped that by removing all the mediocre image pairs, the model would not only generates better-translated images for car and street scenes but also for completely different scenes. As you can see in Figure 17, the generations of streets and cars reached a very satisfying translation state, with local changes for example the ground getting dirty, the car getting rusty, and the vegetation getting dry,... But furthermore, the model translates also very well images not present in the new training dataset, like interior scenes, mountain scenes,...

Training evaluation Evaluating a conditional GAN is quite hard, but for reference, we show our achieved generator L1 loss and the generator total loss after 100k steps in Fig 18 and Fig 19 respectively.

The pix2pix paper[3] defines the generator loss as a sigmoid cross-entropy loss of the generated images and an array of ones. The authors also define the L1 loss as a mean absolute error between the generated image and the target image. The total generator loss is then defined as a linear combination of the generator loss and the L1 loss. Even though the Generator Losses seem to rise

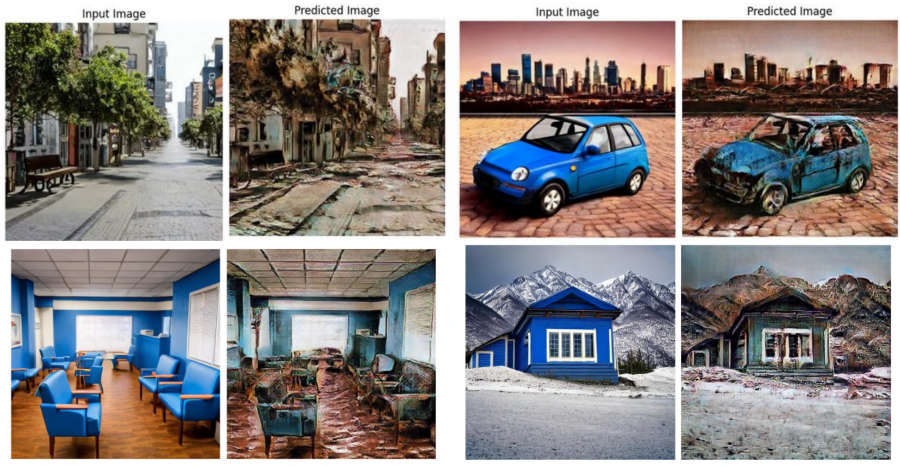


Fig. 17. Some results of the last version of our Pix2Pix model. On the left, are the original images, and on the left are the translated abandoned images

after 65k steps in both metrics, the predictions look visually better after 100k steps.

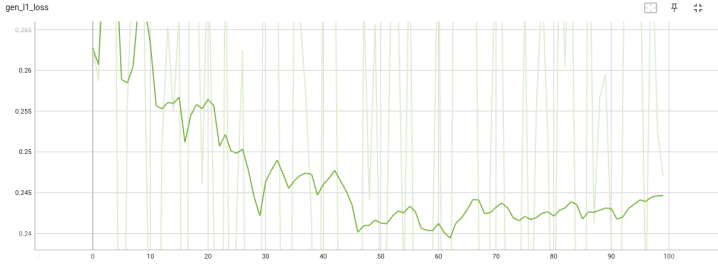


Fig. 18. Generator L1 Loss

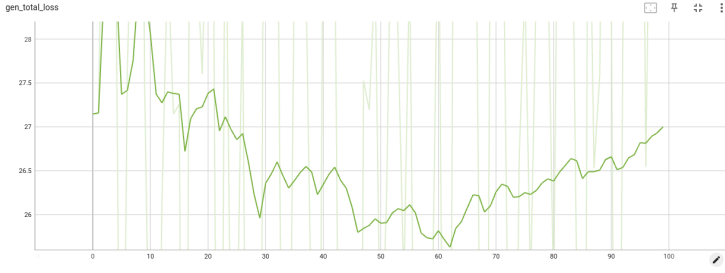


Fig. 19. Generator Total Loss

CLIP Because of the problems with the loss comprehension, we decided to have a more understandable evaluation method. That is why we are using the CLIP model [3]. CLIP is a model that takes as input a textual description and an image and outputs a similarity score between the two.
To evaluate the results given by our model, we prepared a separate evaluation dataset of 25 image triplets, containing the original input image, the ground truth generated by Prompt-to-Prompt, and the prediction of our model. For each of these images, we calculated via CLIP the similarity to two prompts: a modern and tidy one[4], and an abandoned one[5].
We then came up with the `diff` score, which is the difference of the similarity scores between the Prompt-to-Prompt and our model generation on the abandoned description.
This score, mostly close to 0 or even positive, told us that we were almost always nearly as good, and sometimes even better, than the Prompt-to-Prompt abandoned generation.
Please have a look at the `model-evaluation-with-CLIP.ipynb` GitHub file[6] for the technical details of the evaluation method.

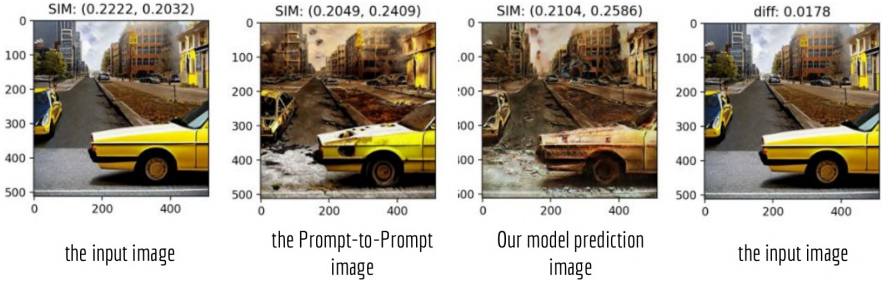


Fig. 20. The similarities with the "modern" and "abandoned" descriptions: (modern, abandoned) on the three images on the left; The difference in the 2 images of the "abandoned" similarity score on the right

4 Results

We show below in Table 1 some pictures generated by our model, compared to the input image as well as the ground truth image (Pix2Pix generated).

[3] CLIP: Learning Transferable Visual Models From Natural Language Supervision by Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever
[4] "The scene looks modern and well-maintained, tidy, preserved, and unspoiled"
[5] "The scene has been abandoned for one hundred years and is dirty, rusty, trashed, and washed out"
[6] https://github.com/Jucifer06/CS413-Computational-Photography.git

In general, the model performs visually quite well on pictures of vehicles and urban sceneries. It however does not perform that well on more "natural" sceneries, as we can see in Table 2 showing a picture of a water fountain: the model does not manage to handle the water properly. Compared to the prompt-to-prompt images, our model seems to perform better at preserving texts in images, prompt-to-prompt usually transforms it to gibberish.

We also tried using real-life photographs that we took as inputs, which gave satisfying results, as shown in Fig 21.

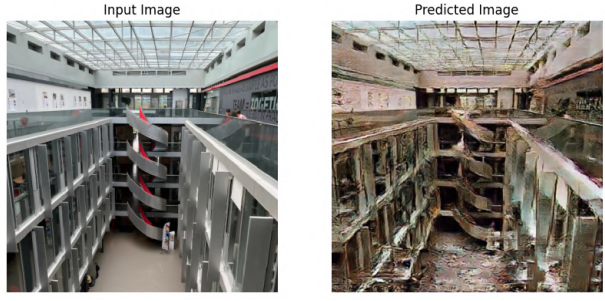


Fig. 21. Predicted image (on the right) of a real life photograph (on the left)

4.1 Image generation time

In Table 3 you can see the time it takes Prompt2Prompt and our model to generate a timewarp version of an image (Note: the exact prompt for the Prompt2Prompt model did not change the duration). The time measurement was done on a *Tesla T4 GPU* in *Google Colab*.

5 Conclusion

We created a synthetic dataset of pairs of images using prompt-to-prompt image editing with cross attention control[2]. Each pair consists of an image of a general scenery, and an image of the same scenery as if it had been abandoned for a hundred years. We focused on prompt engineering and cross-attention control in order to get the best possible results from prompt-to-prompt for our dataset. We then trained a Pix2Pix[3] based conditional GAN using U-Net[6] as the generator on this dataset. We thus provide a model capable of taking as input an image and outputting an image of the same scene as if it had been abandoned for a hundred years. We managed to get satisfying results, when compared to the ground truth of our dataset, with the limitation being put on humanoids and living creatures, as well as scenes that strongly include nature. We however

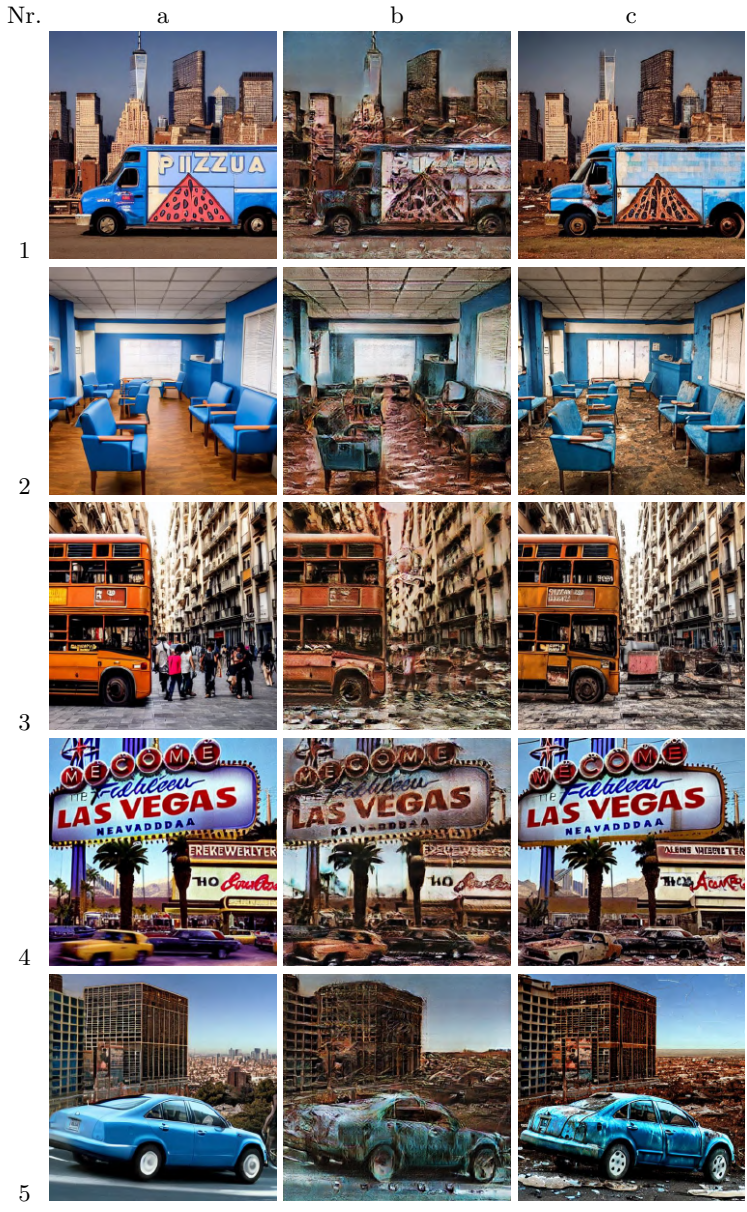


Table 1. Column a shows the input image, column b shows the output image of our model, and column c shows the ground truth, that is the image generated by prompt-to-prompt as a pair with the image in column a.

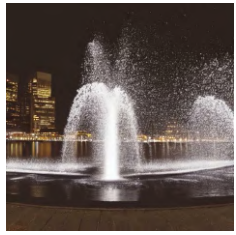
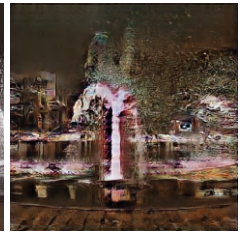
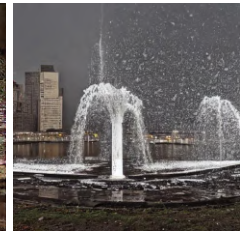
Nr.	a	b	c
1			

Table 2. Limitations of our model: the water is not properly handled. Column a shows the input image, column b shows the output image of our model, and column c shows the ground truth, that is the image generated by prompt-to-prompt as a pair with the image in column a.

Table 3. Time it takes (in sec) to generate a timewarped image

Prompt2Prompt	trained Pix2Pix
16.2 s	0.2 s

managed to get good results with pictures including crowds and text as well as real-life photographs.

Future work could focus on getting better performance with images containing living things and nature objects, like water. This could be achieved by constructing a dataset with Prompt2Prompt directly for this use case. Another option for future work could include trying to train an image to image model backwords, from a time warped image back to an image how it would look like 100 years ago when people also inhabited it.

All our work can be found in our github repository[5].

References

1. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to Follow Image Editing Instructions (2022), http://arxiv.org/abs/2211.09800
2. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-Prompt Image Editing with Cross Attention Control pp. 1–19 (2022), http://arxiv.org/abs/2208.01626
3. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 **2017-Janua**, 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
4. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-Image Translation: Methods and Applications. IEEE Transactions on Multimedia **24**, 3859–3881 (2022). https://doi.org/10.1109/TMM.2021.3109419
5. Parchet, J., Montemagni, C., Müller, M.: CS413-Computational-Photography Project. https://github.com/Jucifer06/CS413-Computational-Photography (2023)
6. Ronneberger, O., Fischer, P., Brox, T.: INet: Convolutional Networks for Biomedical Image Segmentation. IEEE Access **9**, 16591–16603 (2021). https://doi.org/10.1109/ACCESS.2021.3053408
7. Shen, Y., Luo, M., Chen, Y., Shao, X., Wang, Z., Hao, X., Hou, Y.L.: Cross-view image translation based on local and global information guidance. IEEE Access **9**, 12955–12967 (2021). https://doi.org/10.1109/ACCESS.2021.3052241
8. Shocher, A., Gandelsman, Y., Mosseri, I., Yarom, M., Irani, M., Freeman, W.T., Dekel, T.: Semantic pyramid for image generation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 7455–7464 (2020). https://doi.org/10.1109/CVPR42600.2020.00748
9. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Large-scale prompt gallery dataset for text-to-image generative models. arXiv:2210.14896 [cs] (2022), https://arxiv.org/abs/2210.14896
10. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision **2017-Octob**, 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244