Photorealism Style Transfer

Group 8: Hongzhuo Chen 46926636, Kuo-Tien Wu, 28256157

December 2023

Abstract

The emergence of photorealism style transfer has brought simulation-to-reality conversion a lot of attention in graphics and computer vision. This technology can transform artificial images into realistic ones, having great potential for a variety of applications, including trajectory prediction in self-driving systems. Our study focuses on unpaired photorealism style transfer and investigates the performances of three well-known models - Instruct Pix2Pix, Enhancing photorealism enhancement (EPE), and Contrastive Unpaired Image Translation (CUT). We explicitly investigate the generated realistic images of the GTA V dataset, which contains synthetic images, through depth and canny information, to test each method's performance.

1 Introduction

Photorealism refers to the process of improving the realism of images, which means making them appear more realistic. The goal of this project is to conduct a comprehensive comparison of the performance of three state-of-the-art image-to-image translation . As in Fig. 1, photorealism can make visually unrealistic images (as the left image) more realistic (as the right image). In most cases, photorealism is treated as an image-to-image translation task, where the model learns to translate images of one style to another style.

Figure 1: Example of photorealism: The left image is from GTA[9] dataset, the right image is enhanced by CUT[8]

.

For image translation there has been a series of deep learning-based models. CUT (Contrastive learning for unpaired image-to-image translation)[8] trained a contrastive learning model to maximize the mutual information between input and output images. Staple Diffusion[10] , a model based on diffusion model[11] , can be used for image-to-image translation, and other image synthesis works including image generation, text-to-image generation, and super-resolution.

In this work, we conducted a comparison of state-of-the-art image-to-image translation and color transfer methods to enhance the realism of synthetic images. We tested the models on transferring images from GTA V [9] to realistic scenes such as Cityscapes [1].

2 Motivation

The objective of this photorealism is to bridge the gap between simulated and real-world images. To enhance the stability and robustness of tasks such as trajectory prediction, it is necessary to reduce the discrepancy between simulated and real-world images. By using realistic data to train deep learning models, we can achieve accurate and reliable results, which not only enhances the generalization capabilities of the models, but also can be applied to autonomous driving systems which emphasizes safety and efficiency.

Our goal, after conducting extensive research on the subject, was to push the boundaries of existing style transfer technologies, exploring their capacity to bridge the gap between real-world images and their simulated equivalents. It is possible to improve the realism of computer-generated images by identifying the best applicable technique, which could lead to more exact and dependable analysis, prediction, and decision-making in a variety of sectors. The ultimate goal of this research is to seamlessly integrate simulated surroundings with the physical world, ushering in a new wave of applications that thrive on computer-generated images.

3 Related Works

3.1 Photorealism

The ideal way to synthesis a photorealistic image is to simulate every physical process, which is computationally expensive, making it infeasible. Therefore, common photorealism enhancement approaches train models to generate images similar to realistic scenes. Generated images may not be physically correct, but these models can leverage feasibility and accuracy. Since photorealism is transferring a less realistic image to a more realistic image, it is often seen as an image-to-image task.

In deep learning, photorealism's goal is to generate or enhance images that appear like real-world photographs. Common photorealism models adopt convolutional neural networks (CNNs)[6] or generative adversarial networks (GANs)[3]. In most photorealism tasks, models are trained on real-world image datasets, learning data distributions and feature in these images. In this manner the model learns generating new images with similar attributes (e.g. textures, lighting and structural elements).

Conditional image synthesis (e.g. [4]) learns the entire image formation process from data, synthesizing images from semantic label maps, which in most cases results in underconstrained images. There are some methods (e.g. [2]) using image-based rendering, using real imaginery of a scene for rendering novel views. This requires photos of various scenes, making the tasks infeasible. Another way adopted different data-driven approaches. [5] improve realism of photos by learning nearest neighbor patches from photos with different structure. But variation between rendered and reference image can degrade the quality of enhanced images.

3.2 Style Transfer

Unsupervised learning allows models to be trained without labels which require intense labor to make. Contrastive learning and generative learning, different forms of unsupervised learning, has been widely applied to image-to-image tasks.

Contrastive learning is an unsupervised learning framework that learns representation by comparing data and its augmented samples. CUT (Contrastive Unpaired Translation) is a contrastive learning-based framework. It trains a generator to generate images from one style to another, then minimize the contrastive loss function, InfoNCE loss[7]. In this way, the model learns the similarity between the two domains.

Diffusion[11] models is an unsupervised generative model inspired by nonequilibrium thermodynamics. It learns representation by gradually adding noise to the input data to destroy the structure in data distribution and train a reverse process to restore the data. Diffusion model can be applied to different image-to-image tasks, as well as text-to-image tasks. Stable diffusion [10] is a diffusion model that can generate images from text, unconditionally generate images, as well as translate images. It introduces a cross-attention mechanism to enable the a more general-purpose model, which can be used to different tasks.

4 Methodology

In this paper, we compared the performance of CUT, Instruct Pix2Pix and EPE on unpaired photo-realism style transfer and evaluated the performance through depth and canny images.

4.1 CUT

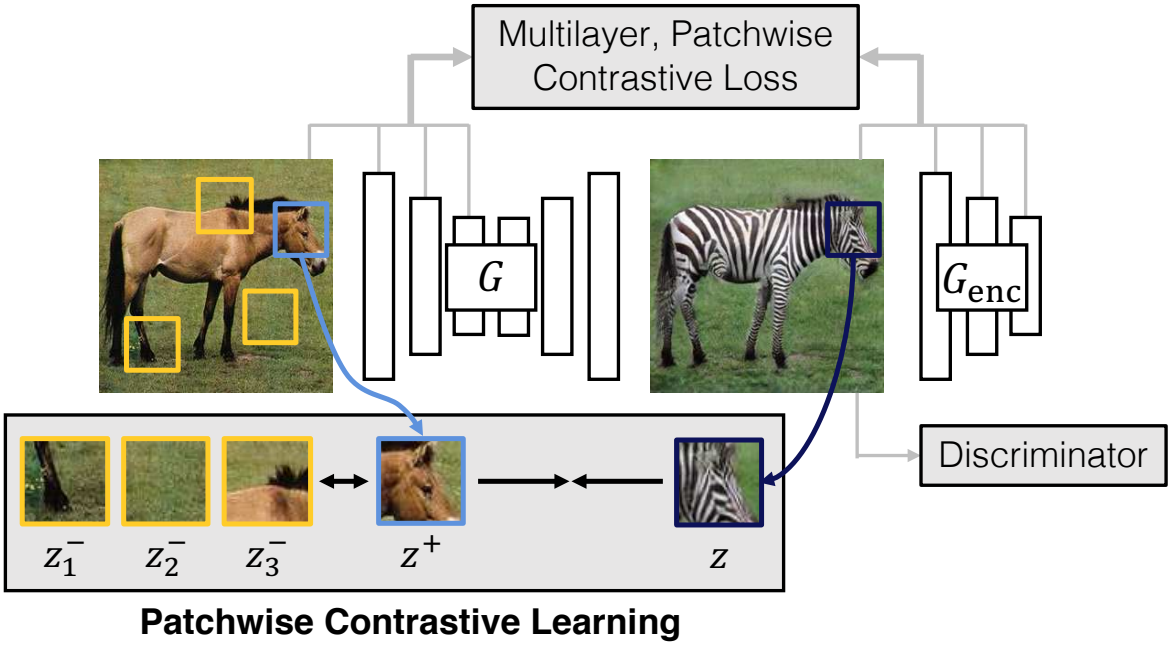

Figure 2: Architecture of CUT[8]

.

The goal of CUT is to translate images from one domain to be more like images from another domain. CUT learns images from a patch level. As in Fig. 2, in an embedding space, CUT brings an input patch (z) and its corresponding output patch (z^+) closer and other ramdom patches z_n^- farther. It adopted adversarial loss from GAN[3] and maximize the mutual information using a noise contrastive estimation framework[7].

4.2 EPE

EPE adopts an image enhancement network enhancing an input rendered image. Also an adversarial objective training provides supervision at different perceptual perspectives. It also proposed a novel strategy for sampling image patches during training to avoid artifacts occured in previous works.

Also, EPE employed G-buffers providing physical information (geometry, materials, lighting, etc.). The network uses a perceptual discriminator to calculate realism score between real image and enhanced image and LPIPS loss function [12] to compare the structural differences between the input and output. Using G-buffers and LPIPS, EPE can accurately enhance images while preserving features of the original image.

4.3 Instruct Pix2Pix

We're using the Instruct pix2pix model for performing style transfer. The process begins with encoding the input image into a latent space. Instruct pix2pix's diffusion process uses two separate elements that impact it: the text prompt and input image, which condition and guide it.

Employing the Instruct pix2pix model, we can generate stylized images by manipulating the text CFG and image CFG parameters. This diffusion process is guided by the Classifier Free Guidance (CFG) mechanism, which is similar to the vanilla Stable Diffusion. The stylized images we produce effectively transfer the intended style from the text prompt to the input image. To determine these images' visual quality, realism, and adherence to the desired style is crucial.

5 Experiment

In our experimentation, we opted for CUT, EPE, and Instruct Pix2Pix techniques as they have shown to be effective in style transfer tasks. Our endeavor was to deploy these methods to unpaired photorealism style transfer on the well-known GTA dataset, which is a commonplace benchmark utilized in the segmentation field. Our objective was to assess the performance of these approaches and recommend the most appropriate technique to achieve top-notch style transfer in the sim2real context.

5.1 Datasets

In the experiment, we translate images from GTA V[9] dataset to realism images such as Cityscapes [1] dataset.

- **GTA V Dataset**: This dataset contains 24996 images extracted from the video game Grand Theft Auto V. This game features a highly realistic city environment. Compared to datasets extracted from real life, this dataset is easier to extract and label. The images and videos in the GTA V dataset cover various scenes and scenarios found within the game, including street scenes, driving sequences, pedestrian behaviors, and interactions between different entities. This dataset can be used for various computer vision tasks, such as object recognition, scene understanding, and autonomous driving.

- **Cityscapes**: Cityscapes is a widely-used benchmark for various computer vision tasks. The dataset contains high-resolution images of streets from 50 cities captured from a vehicle-mounted camera, covering various urban scenes with different weather conditions and lighting conditions. Apart from images, the Cityscapes dataset also has detailed annotations for evaluation. It includes a separate validation set and a test set without publicly available ground truth labels, allowing objective evaluation of models and comparing performance with other state-of-the-art approaches. The dataset also provides fine-grained instance-level annotations for specific object classes, enabling instance segmentation and other advanced tasks.

5.2 Hardware and Software Environment

Our experiments were conducted on a cluster of four NVIDIA RTX V100 GPUs with 32GB of memory each. We used CUDA version 11.7 for GPU acceleration. We executed our experiments in a terminal environment and edited our code using Visual Studio Code.

5.3 Result and Analysis

From the GTA V dataset, we scrutinized the depth and canny information of the images created by each approach to evaluate their efficacy thoroughly.

When taking a closer look at the canny images shown in Figure 4, one can notice that Instruct Pix2pix's generated image (Figure (b)) stands out from the other approaches because it maintains the most details. Instruct Pix2pix excels at safeguarding the intricate attributes of the source image, including edges, contours, and delicate textures. This suggests that Instruct Pix2pix has a remarkable ability to uphold the visual precision of the original image throughout the style transfer procedure.

In contrast, when looking at figures (c) and (d) produced by EPE and CUT, the overall detail appears to have been diminished as compared to the original image. This can be seen by the edges being less prominent and the fine textures being blurred. Despite the reduction in detail, EPE and CUT were still able to effectively transform the texture and lighting conditions of the original image.

From this, it can be concluded that these methods perform well by capturing the general visual characteristics and aligning the appearance of the input image with the target domain - Cityscapes dataset.

From analyzing the clever pictures to inspecting the depth data of the crafted images, we discovered that Instruct Pix2pix has a remarkable ability to conserve the same depth data seen in GTA V images, as shown in Figure 5. By maintaining the relative distances and perception of a three-dimensional realm, our findings reveal that Instruct Pix2pix can adeptly preserve the depth cues contained in the authentic image.

The changes in depth information with CUT stand out compared to the other methods. The generated image by CUT (Figure (d)) portrays noticeable changes in depth perception. These changes can be attributed to the transformation of lighting conditions and texture, and may play a role in the spatial relationships between objects within the scene.

Examining the canny images and depth data allows a complete evaluation of the photorealism enhancement capabilities of each technique. The Pix2pix instruction excels in maintaining the minutiae and depth information, attaining a remarkable resemblance between the output and the source. Conversely, CUT and EPE's efficiency in altering the texture and illumination intensifies the similarities between the generated and the target domains, albeit at the expense of precision reductions. Consider that the selection of a method heavily relies on the peculiar qualities and preferences for a particular scenario, as each method maintains its own unique advantages and shortcomings. Remember that when deciding between methods, it is important to account for individual requirements and priorities.

We also use SSIM (structural similarity index) to evaluate the generated images. SSIM is a metric used to quantify the similarity between two images. It's widely used in the field of image processing to assess the quality of images, particularly when evaluating the performance of image generation or image enhancement algorithms. The SSIM index ranges from -1 to 1, where 1 indicates perfect similarity between images. The scoring of Pix2Pix, EPE, and CUT are 0.81, 0.69, and 0.51 respectively. Pix2Pix has the highest score indicating has the most similarity between the original image.

(a) Image from GTA V dataset

(b) Image generated by Instruct Pix2pix

(c) Image generated by EPE

(d) Image generated by CUT

Figure 3: Images from different style transfer models

(a) Canny image of the original GTA V image

(b) Canny image from Instruct Pix2pix

(c) Canny image from EPE

(d) Canny image from CUT

Figure 4: Canny images

(a) Depth image of the original GTA V image

(b) Depth image from Instruct Pix2pix

(c) Depth image from EPE

(d) Depth image from CUT

Figure 5: Depth images

6 Conclusion

In an effort to tackle the inconsistency between computer-generated and real-life visuals, a study was conducted to investigate the efficacy of different style transfer approaches in achieving photorealism. The primary goal was to bolster the consistency and resilience of activities such as predicting trajectories, resulting in greater precision and dependability. To assess the capabilities of three prominent techniques - CUT, EPE, and Instruct Pix2Pix - experimentation was conducted. These strategies were chosen precisely for their established effectiveness in style transfer undertakings. A benchmark in the segmentation arena known as the GTA dataset was used to execute the unpaired photorealism style transfer. To determine the optimal method for sim2real style transfer of superior quality, performance evaluations of various approaches were conducted.
Remarkable preservation of fine details and depth information was demonstrated by the Instruct Pix2Pix technique, producing highly realistic outputs that closely resembled the original GTA V dataset. Its retention of intricate features such as edges, contours, and textures was noteworthy. However, both EPE and CUT were able to transform the texture and lighting conditions of original images, bringing them closer to the Cityscapes dataset. This transformation caused a decline in overall detail compared to the original images. These distinct characteristics and performance were highlighted by the experiment's results. Some finer details were sacrificed by EPE and CUT while altering the visual appearance of the images.

Pix2Pix Instruct has shown promise in the field of sim2real context, following a comprehensive assessment. Its superior capacity in conserving minute details, depth information, and visual authenticity, demonstrate its aptitude in accurate and realistic image creation, making it an excellent fit for applications that call for high-quality style transfer.When picking a style transfer technique, it's vital to prioritize the application's particular needs. Balancing the preservation of delicate components with visual transformations must be assessed thoughtfully to guarantee the intended result.

In bridging the gap between simulated and real-world images, this study's exploration of style transfer techniques adds to our knowledge. By advancing these findings, future research can develop methods that lead to photorealism. This will permit simulated environments and the real world to integrate effortlessly.

7 Future Work

Future work can aim to explore and enhance various aspects of three prevalent photorealism style transfer techniques when applied to unpaired image transfer. Although this study offers noteworthy perspectives, further investigations are possible in the following potential directions:

- Evaluating the authenticity and projection ability of style transfers can be developed through a variety of metrics. Specific measures, like perceptual similarity indices or human perception studies, can provide a more in-depth assessment of these images. By including these additional evaluation techniques, we can better understand the quality of various style transfer methods. Additionally, testing the predictive accuracy of these images with real-world driving datasets can help us measure the effectiveness of photorealistic style transfers on prediction algorithms. Thoroughly evaluating these trajectory predictions can provide valuable insight into the robustness and precision of these algorithms.
- Autonomous driving systems are equipped to handle many different driving scenarios, but in order to really put them to the test, we need to evaluate their performance in a wide range of environments. This means testing them in varied lighting conditions, different types of weather, and on various types of roads. By evaluating their ability to generate realistic representations in these scenarios, we can determine how effective they will be in a broader range of real-world driving situations. Potential future work can focus on this type of testing to improve these systems even further.
- Photorealism style transfer methods require representative training data to perform well. Thus, future inquiry can investigate fine-tuning or adaptation of existing models to enhance their effec-

tiveness in matching the unique features present in autonomous driving situations. Such methods could entail using fresh datasets obtained from actual driving experiences or even constructing specialized loss functions that prioritize critical visual markers needed for trajectory prediction.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] Micah K Johnson, Kevin Dale, Shai Avidan, Hanspeter Pfister, William T Freeman, and Wojciech Matusik. Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Transactions on Visualization and Computer Graphics*, 17(9):1273–1285, 2010.

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[8] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[9] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.