Team 46: Generation of Real World Images from Simulation Images

Nischal Maharjan 7058343
Jaykumar Bhagiya 7055903
Hevra Petekkaya 7055462

1. Task and Motivation

The primary motivation for this project is to address the current demand for high-quality data, particularly in the context of self-driving cars. Ensuring the reliability of autonomous driving systems in diverse environments is crucial, given the variability in weather conditions and the presence of numerous dynamic elements such as pedestrians, cyclists, other vehicles, animals, and unexpected obstacles. To achieve this robustness, it is essential to expose the models to as diverse a dataset as possible during training, minimizing the effects of distribution shifts between the training data and real-world environments.

Collecting data for these systems is especially challenging for several reasons: the cost of precise equipment, the need for accurate labels, and unpredictable weather conditions. To mitigate these issues, researchers have increasingly turned to simulated data, using various simulators such as **CARLA (Car Learning to Act), LGSVL (LG Silicon Valley Lab) Simulator, AirSim (Aerial Informatics and Robotics Simulation), Apollo Simulation**, and others. These tools provide the easy method for generation of datasets for self-driving cars.

Despite this, models trained solely on simulated data often lack robustness when applied to real-world scenarios. As highlighted by the authors of "Measuring Robustness to Natural Distribution Shifts in Image Classification" [5], there is little to no transfer of robustness from synthetic to natural distribution shifts. This project aims to balance the trade-off between the ease of dataset generation and the robustness of models.

The task involves generating realistic images from synthetic images produced by simulation data. Specifically, we intend to extend the methodology proposed in "Image-to-Image Translation with Conditional Adversarial Nets" by Isola et al. [4] to the domain of autonomous driving. Our objective is to transform synthetic images from the Virtual KITTI dataset into photo-realistic images that can be used for training autonomous driving models.

2. Goals

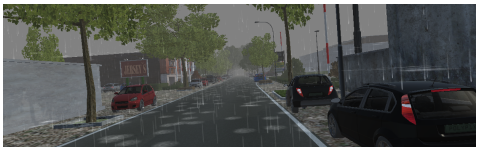
- The final goal is to create a model that is capable of generating as photo realistic images as possible from the given synthetic dataset.
- By mid term we aim to achieve the milestone of implementing the framework, based on pix2pix image translation. Develop scripts for preprocessing data and preparing it for the training process.
- After above milestone we try to improve and analyze the performance of the image generation and its evaluation.

3. Datasets

Real Image

Synthetic image

Fog Image



Rain Image

Figure 1. Sample images from Kitti and Virtual Kitti Datasets

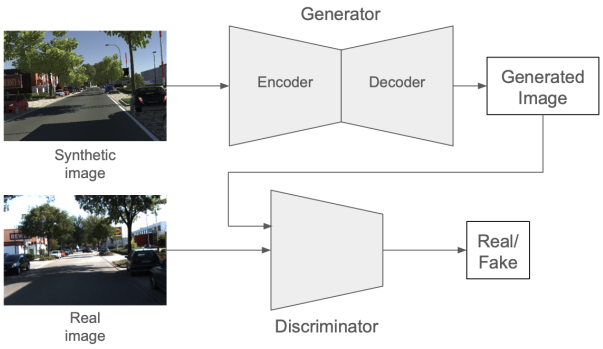


Figure 2. Proposed method

For this project, we have leveraged the The KITTI Vision Benchmark Suite Dataset introduced in [3] and Virtual KITTI dataset introduced in [2]. Virtual Kitti. Virtual KITTI is a photo-realistic synthetic video dataset designed to learn and evaluate computer vision models for several video understanding. Unity game engine and a novel real-to-virtual cloning method was used to generate synthetic data of the real world data present in the Kitti dataset. For given environment, Virtual Kitti has data for different weather condition such as fog, morning, rain as well as shown in Figure 1

4. Methods

Our proposed method is to train an image translation model using the real-synthetic image pairs discussed earlier. The high-level architecture is shown in Figure 2. We plan to train a generator resembling an autoencoder model such as U-Net which is able to reconstruct the real images from the synthetic image. In general, our generator tries to learn the mapping between synthetic and real images. However training just an autoencoder with l2 loss, does not generate realistic images hence we introduce the adversarial loss. The discriminator model tries to differentiate between the real image and generated images. The training is carried out until there is stability between generator loss and discriminator loss.

We also plan to play around with the architecture to improve the performances. Since the amount of data is limited we intend to experiment with the data augmentation methods to improve the performances.

The authors of the paper mention that for some image translation tasks, there is a significant amount of shared information that is low level. Low level detail usually corresponds to high frequencies such as edges. Using a simple encoder - decoder, would require all the information to pass through all the layers. However, it is known that as one gets closer to the bottleneck usually the high frequencies (low level detail) gets destroyed. Consequently, being able to extract the low-level details would be a challenging task for the decoder. Hence, they motivate using a U-Net for the generator so that the low-level information can flow easily without the need to survive all the way down to the bottleneck.

Furthermore, ViTs (Vision Transformers) [1] were introduced after this paper was released. Hence, we are curious about how well they can improve performance compared to CNNs, which are used in U-Nets. Realistically, even though ViTs have a larger capacity compared to CNNs due to their global attention mechanisms, they might not perform as expected with a small number of training samples. This is because, unlike CNNs, ViTs lack an inductive bias and are therefore more reliant on large amounts of data.

In the paper, the authors are using PatchGAN for the discriminator, which means that they rely on high frequency data such as texture to test whether the image is real or fake. On one hand, as the authors mentioned this is more efficient because it requires fewer number of training parameters. On the other hand, it could potentially fail in some scenarios where global coherence is needed. In other words, the generator won't have any incentive of necessarily generating globally coherent images because the discriminator does not have global awareness. Hence, we also plan on testing with an image discriminator rather than testing each patch on its own.

5. Evaluation

Evaluating the Generative Adversarial Networks has always been a challenging task. The trivial metric would be the root mean square loss between the pixel values of real image and the generated image. However RMSE may not capture the perceptual quality or the high-level features that image generation network are typically evaluated on.

Usually metrics such as inception scores, Fréchet Inception Distance and perceptual loss are better metrics to evaluate the GAN networks.

Another evaluation criteria is to see the performance of some baseline models for any task on the generated images. For example if yolo detection model pretrained was able to perform well on the generated images that signifies good quality of the generated images.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[2] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtualworlds as proxy for multi-object tracking analy-

sis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016. 2

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2

[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. 1

[5] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1