Team 46: Generation of Real World Images from Simulation Images

Nischal Maharjan
7058343

Jaykumar Bhagiya
7055903

Hevra Petekkaya
7055462

1. Progress Report

1.1. Task Completed

The task we have completed until now have been briefly explained below.

1.1.1 Literature review of the related Tasks

As mentioned in the proposal report, we intend to explore the integration of Generative Adversarial Networks (GANs) with transformer architectures. Upon conducting a thorough literature review, we identified a relevant study titled ViTGAN: Training GANs with Vision Transformers by Lee et al. (2022) [4]. This paper presents a novel approach where both the generator and discriminator of the GAN are designed using vision transformers.

However, we need to extend their architecture to implement a Conditional Generative Adversarial Network (cGAN) [5]. This extension is necessary for our project, which focuses on generating realistic images based on their corresponding synthetic versions.

1.1.2 Implementation of the baseline model

The general overview of the image translation pipeline is shown in figure 1. The authors of paper [3] have implemented the UNeT architecture for the generator. Hence we have implemented the custom generator resembling autoencoder model with skip connecton similar to the UNet architecture [6]. Unlike Unet we have removed cropping operation in before concatenation, instead we have used proper padding to match the dimension of the skip connection. For discriminator we have used similar architecture of Patch GAN dicussed in the paper.

1.2. Results

The current progress or milestone of the project is shown in the figure 2. It shows few samples of generated images for the given synthetic input images alongside with its ground truth real images. We see that the model has started

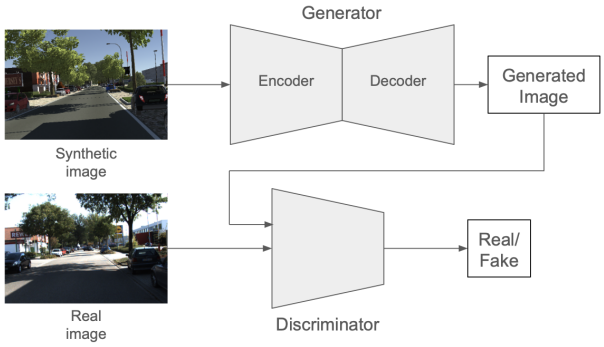
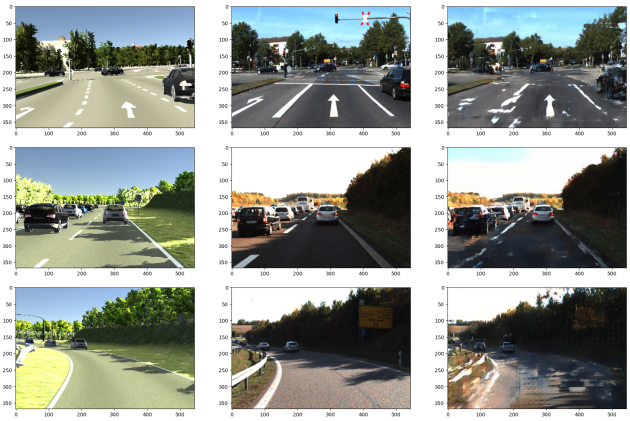


Figure 1. Overview of model architecture

to learn the natural distribution of real world environment, However the images are still not perfect.



Input Images | Ground Truth | Generated Images

Figure 2. Input synthetic images and their corresponding Ground Truth and Generated images

1.3. Evaluation

In the comparison of training and validation results, the RMSE for the training dataset is 0.0995, while the vali-

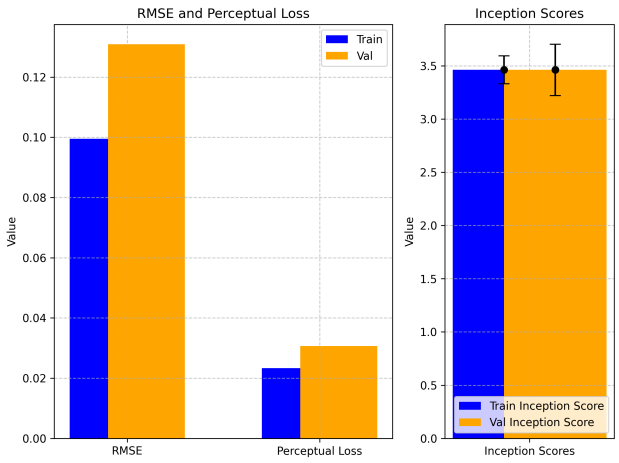


Figure 3. Training & Evaluation metrics comparison for image translation model

dation RMSE is higher at 0.1309. This indicates that the model performs better on the training data, suggesting overfitting, where the model has specialized too much on the training examples. Similarly, the Perceptual Loss is lower in training (0.0233) compared to validation (0.0307), further demonstrating that the model's ability to generate perceptually realistic images diminishes on new data.

The Inception Score follows a similar trend, with the training score at 3.462 and the validation score slightly lower at 3.461, but with a higher standard deviation in the validation set (0.241 vs. 0.131). This increased variability in validation results highlights that the model's performance is less stable and consistent when applied to new, unseen data. Overall, these results indicate that while the model performs well on training data, it struggles to generalize to validation data, suggesting the need for improvements in generalization and robustness.

2. Problems Encountered

- **Limited Data** We are utilizing the KITTI Vision Benchmark Suite dataset [2] and the Virtual KITTI dataset [1]. Despite having **2,126 image pairs** from 5 different environments, this dataset remains limited. The pix2pix method [3] showed promising results with a dataset of around 400 images, but our objective to use generated images as realistic data requires a larger and higher-quality dataset. Therefore, expanding or improving the dataset is crucial.

- **Overfitting of the Model** Analysis of the training logs indicates that while training loss is decreasing, validation loss has plateaued. This suggests that the model may be overfitting the training data, likely due to the small size of the dataset. To address this, we may need to apply techniques such as regularization, data augmentation, or dataset expansion to improve generalization.

- **Artifacts in the Generated Images** The generated images resemble real-life environments but contain artifacts such as inconsistent textures and unrealistic object placements. These artifacts prevent the images from being used effectively in machine learning applications, which is a primary goal of this project. Addressing these imperfections is essential for improving the quality of the generated images.

3. Next Steps

- Implement data augmentation techniques like rotation, scaling, and color adjustment to enhance dataset variety and size.

- Experiment with Vision Transformer architecture for the generator, which may offer improvements over traditional convolutional neural networks by capturing more contextual information.

- Conduct a thorough comparison and analysis of different experimental approaches to identify the most effective strategies for enhancing image quality and reducing overfitting.

References

[1] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtualworlds as proxy for multi-object tracking analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016. 2

[2] 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

[3] 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, 2

[4] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers, 2024. 1

[5] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1