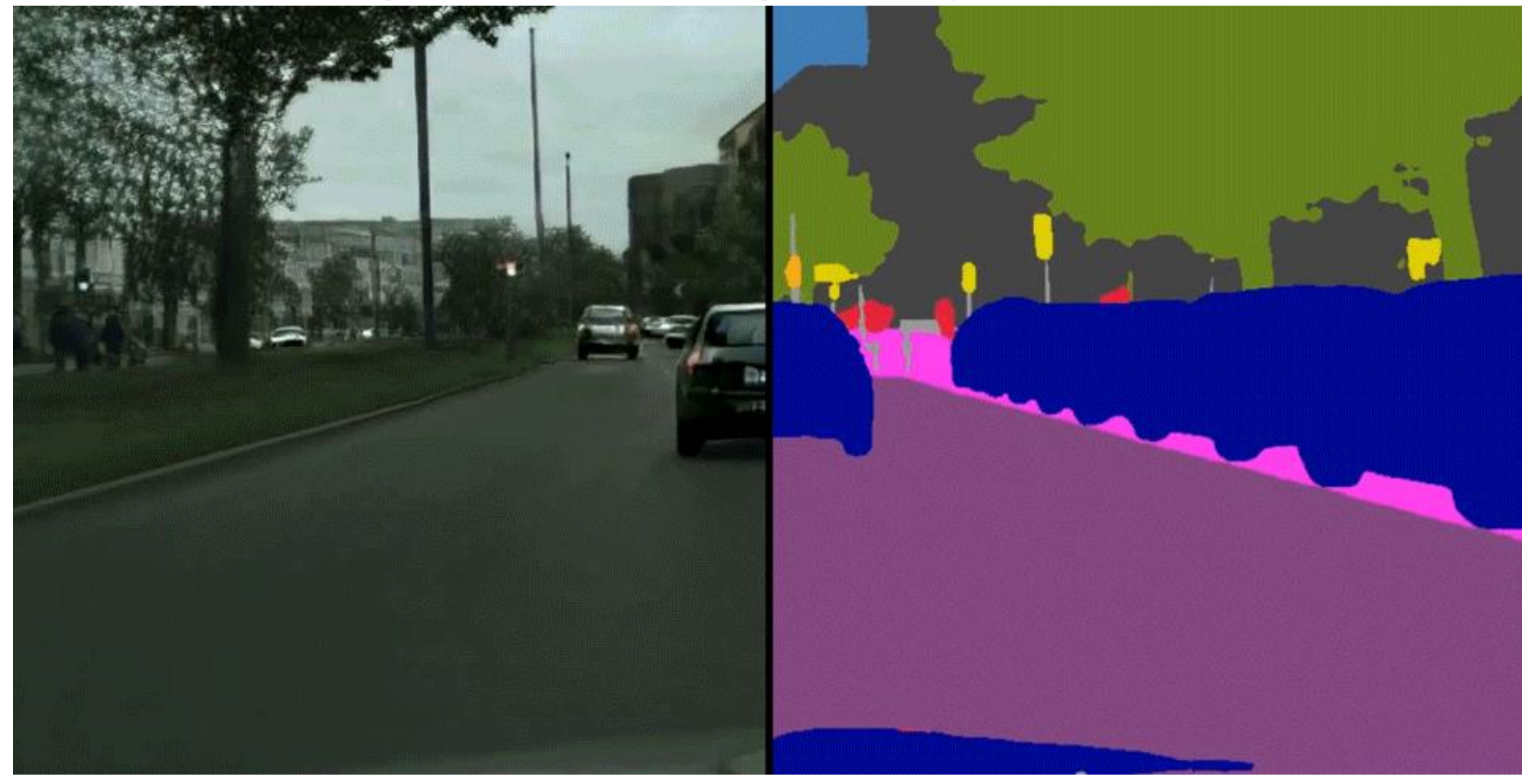
UR-60 Video-to-Video Synthesis With Semantically Segmented Video

Abstract

Our project involves studying the usage of generative adversarial networks (GANs) to translate video of semantically segmented masks to photo-realistic video in a process known as video-to-video synthesis. In our study, we implement a model that learns a mapping from semantically segmented masks to real-life images. To achieve this, we employ a conditional GAN-based learning method derived from the architecture introduced by *pix2pix*. The model produces output conditionally based on the provided segmentation mask sequence. Our model can synthesize a translated video that resembles real video by accurately replicating low-frequency details from the source. Though our model produces video that lacks some high-frequency details, this study demonstrates how simplistic conditional GANs are capable of translating semantically labeled videos.

Introduction

- Image-to-Image translation is the task of translating an image from one domain to another, using another image as the condition for the translation.
- Video-to-Video synthesis is the counter part that is tasked with synthesizing video by translating a sequence of images from one domain to another.
- Useful applications include creating efficient rendering techniques that can extract key features from videos of various scenes. This can be applied to autonomous vehicles and future video generation.
- We attempt to apply video-to-video translation to transform images of semantically segmented video frames into photo-realistic image sequences that depict the corresponding labels.



Research Question(s)

- **How to generate photo-realistic video given a semantically labeled source video?**
- **How to maintain temporal coherence across the image sequence?**

Materials and Methods

- Implemented using Keras for TensorFlow in Python
- Based on the *Pix2Pix* architecture
- Trained using the Cityscapes Dataset
- Utilizes a condition GAN that is trained on pairs of semantic masks and original source images

Results

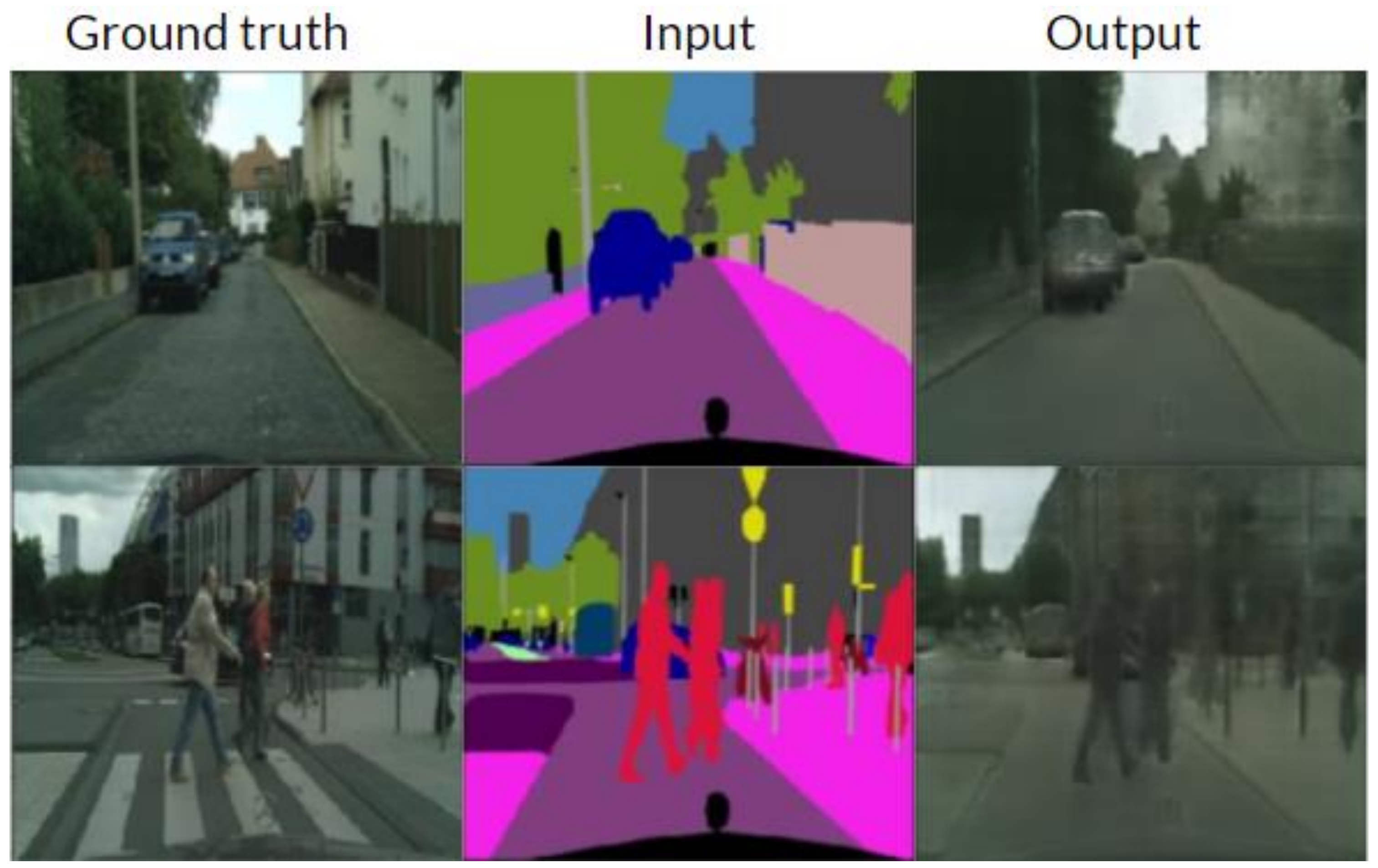


Fig. 1: Two random samples with their associated semantic labels and output synthesized by the generator

- In Figure 1, we show the output of the generator after training for 10 epochs
- The model is able to map specific labels to certain styles with well-defined edges
- The generator properly reconstructs low-frequency detail with distinct edges and shapes
- High-frequency details are not well-represented

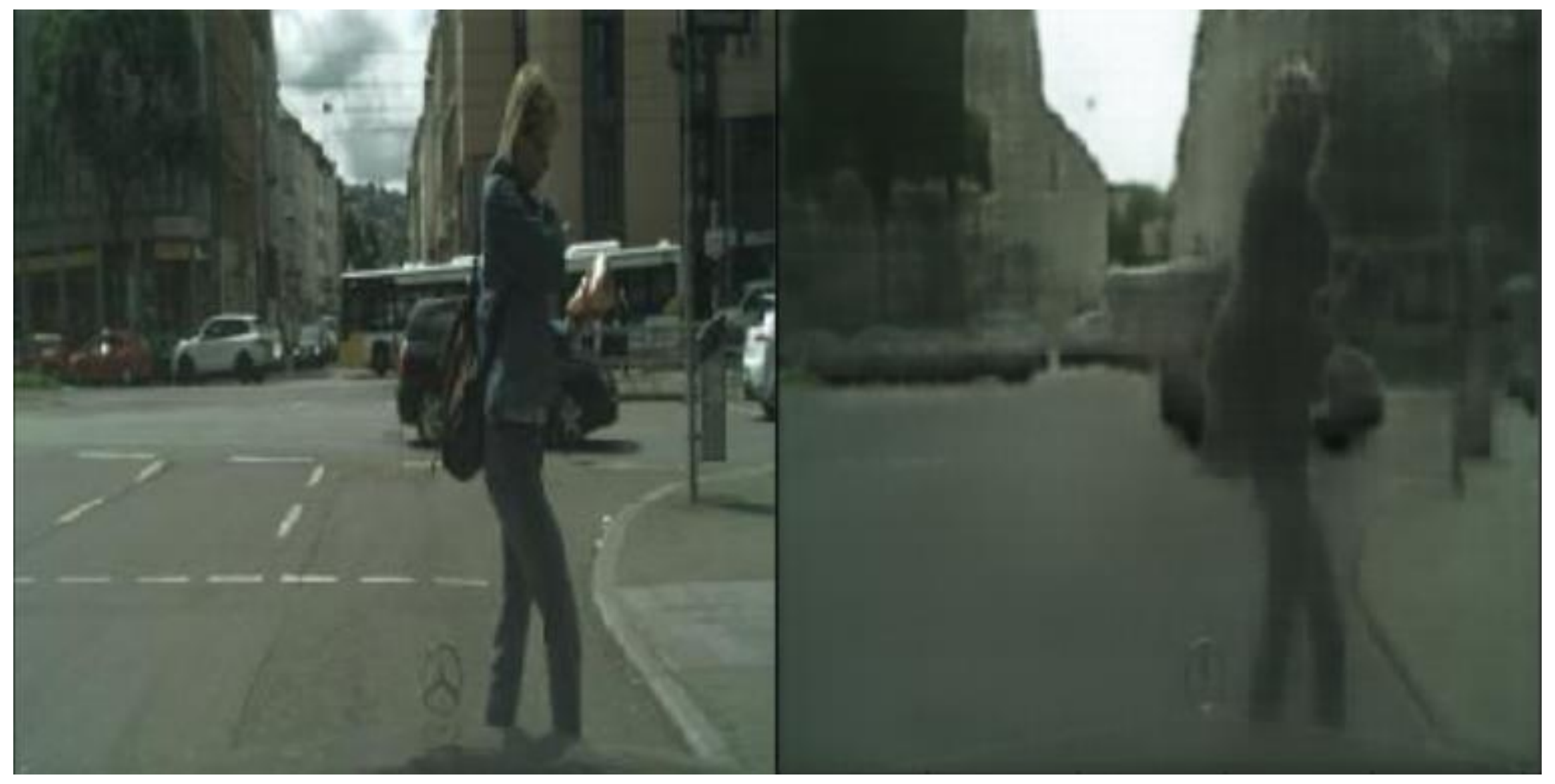


Fig. 2: Comparison between an original source image and the corresponding generated image

Data and Preprocessing Techniques

- Our model was trained using the Cityscapes Dataset
- To prepare the data for training, image pairs of the semantic label masks and original source images are scaled to a resolution of 256x256
- The images of the semantic labels and original source images are then concatenated together, producing a single 512x256 image which contains the input and desired output
- The model is trained on random samples as there were no datasets that contained ordered sequences of samples
- Therefore, the model was not trained sequentially, so it does not have information about previous frames, which causes some flickering in generated video

Future Studies

- Create unique dataset of sequences of pairs of source images and semantic masks.
- Possibly generate maks utilizing game engine software that could generate video sequences using both original textures and replacing textures with solid colors that represent specific semantic labels.
- Additionally, semantic masks can be generated using atrous convolution for dense feature extraction
- Attempt video prediction upon enhancing the model to track previous frames.
- Decrease visual artifacts and have the ability to replace labeled elements within the segmented images such as replacing trees with building or vice versa.

Conclusions

- Our model is able to generate low-frequency detail images given semantically labeled images of various street scenes.
- The current model leaves room for improvement as it currently fails to replicate high-frequency details that exist in the original source images.
- The synthesized images still suffer from noise and blur. We believed that changing the weight of the L1 loss term could potentially solve this issue, however, this introduced visual artifacts.
- We attempted to increase the quality of the generated images by updating the weights of the discriminator with a higher factor since we believed that the discriminator was not learning quickly enough, however, this change did not yield better results.
- The dataset used to train the model only consisted of 256x256 images which possibly attributed to the lack of detail in the output.
- Our current model struggles with maintaining temporal coherence between generated image frames. The video results in flickering as the model does not have information about previous frames however the overall setting of the scene appears mostly consistent.
- We believe our model can be improved with the appropriate dataset consisting of sequential images and paired semantic masks.

Acknowledgments

- Dr. Mohammed Aledhari
- KSU C-Day Administration

Contact Information

- Jordan Hasty
  - jhasty11@students,kennesaw.edu
- Aydan Mufti
  - amufti1@students,kennesaw.edu

KENNESAW STATE
UNIVERSITY
COLLEGE OF COMPUTING AND
SOFTWARE ENGINEERING

Author(s): Jordan Hasty, Aydan Mufti
Advisors(s): Dr. Mohammed Aledhari