GENERATING AUTOMATED VEHICLE TESTING SCENARIOS USING GENERATIVE ADVERSARIAL NETWORK

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF COMPUTER AND INFORMATION SCIENCES

Supervisors

Dr. Jing Ma

Prof. Edmund M-K Lai

March 2023

By

Longjiang Xue

School of Engineering, Computer and Mathematical Sciences

Abstract

In the field of autonomous driving, generating semantic segmentation images from road scene images is of utmost importance. This allows researchers to test and validate autonomous vehicle functions and algorithms using synthetic traffic scenes without actually driving on public roads. Generating traffic scenes using existing images has emerged as a significant research area, with generative adversarial networks (GANs) being a popular choice for image-to-image conversion in computer vision.

In this thesis, we have utilized both traditional conditional adversarial networks, specifically the pix2pix model, as well as the GANformer method, which is a novel and efficient type of transformer, for generating new images. We have then proposed a new methodology by combining the GANformer and pix2pix models to perform performance tests.

To train our models, we have used the Cityscapes dataset, which consists of 5,000 high-quality road scene images and their corresponding manually annotated semantic segmentation images. This dataset serves as a valuable resource for training and evaluating the performance of our proposed models.

The experimental results show that the third methodology which combines pix2pix and GANformer outperforms the other two methods in generating semantic segmentation

images with higher accuracy and less ambiguity. This experiment also demonstrates the different performances of these three methods when converting road scene images into semantic segmentation images.

The findings from our experiments highlight that incorporating pix2pix into GANformer results in higher accuracy and better performance in the context of traffic scenes. This has significant implications for areas such as autonomous driving testing and validation, where accurate and reliable results are crucial. The use of our proposed methodology has the potential to improve the accuracy and reliability of generating semantic segmentation images for traffic scenes, contributing to the development of more robust and efficient autonomous vehicle algorithms and functions.

Keyword-Autonomous Driving, Generative Adversarial Network, GANformer, pix2Pix, Traffic scene image.

Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of student

Acknowledgements

I would like to express my sincere gratitude to all the people who contributed to the success of this research project.

First, I would like to thank our supervisors, Dr Jing Ma and Professor Edmund Lai, for their guidance, support and encouragement during the year-long master's thesis. Their expertise in the field has been invaluable to our research and I could not have completed this project without their guidance.

I would also like to thank Auckland University of Technology for providing us with the necessary resources and facilities to conduct our experiments. The support of this institution has been critical to the success of our research.

Finally, I would like to thank my family for their unwavering support and encouragement throughout the project. Their understanding and patience have been invaluable to my success.

Longjiang Xue

March 2023

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 Vision Technology in Autonomous Driving

As technology continues to evolve, self-driving cars have become one of the key trends in the future of transportation. The implementation of self-driving cars requires the support of many technologies, one of which is self-driving vision technology. Autonomous driving vision technology uses a variety of sensors and computer vision algorithms to sense the surrounding environment, identify traffic elements such as roads, traffic signs, other vehicles and pedestrians, and thus enable intelligent decision-making and control of self-driving cars.

Significant progress has been made in the development of self-driving vision technology. The core task of this technology is to enable self-driving cars to perceive their surroundings, acquiring data through multiple sensors, including LIDAR, cameras, radar, ultrasound, etc., and feeding this data into computer vision algorithms for analysis and processing. In this way, self-driving cars are able to perform tasks such as accurate

location positioning, surrounding environment perception and obstacle recognition. This perception capability is the key to intelligent decision-making and control of self-driving vehicles.

Autonomous driving vision technology includes many key technologies, such as target detection, tracking, semantic segmentation, depth estimation, optical flow estimation, etc. The development of these technologies enables self-driving cars to accurately perceive and understand their surroundings, enabling more intelligent decision-making and control, and improving safety and driving experience. For example, on the road, self-driving cars can recognize different types of traffic signs and what they represent to enable intelligent decision-making. In addition, self-driving vision technology is improving in recognizing obstacles, allowing it to identify and track obstacles such as other vehicles and pedestrians, and to take timely action to avoid accidents.

However, autonomous driving vision technology still faces many challenges, such as complex weather conditions, lighting conditions, changes in road markings, and simultaneous tracking of multiple moving targets. For these problems, scientists are constantly looking for solutions. For example, for complex weather conditions, researchers are developing new algorithms to improve the visual recognition of self-driving cars so that they can maintain stable operation in adverse weather conditions such as rain and snow.(“Xuesong”, n.d.) Also, in the area of multiple target tracking, scientists are developing new algorithms and techniques so that self-driving cars can better identify and track multiple moving targets.

In conclusion, the development of self-driving vision technology is of great importance for the future of transportation. With the continuous development and improvement of self-driving car technology, self-driving cars will be safer, smarter and more convenient,

providing better choices for people's travel. At the same time, the development of self-driving vision technology will also promote the development and progress of the field of computer vision, providing important support and help for research and application in other fields.

1.1.2 Traffic Scene Generation

Traffic scene generation is a rapidly evolving technology that creates virtual environments that simulate the behaviour of vehicles and pedestrians on roads and streets. This technology has many applications, from self-driving vehicle testing to gaming and virtual reality training. In recent years, there has been a growing demand for more realistic and diverse virtual environments to test and train self-driving vehicles and driver assistance systems. As a result, traffic scenario generation has become an important tool for researchers and developers in the automotive industry.

The process of traffic scenario generation involves modelling road networks, vehicles and pedestrians. This requires the use of advanced simulation techniques that take into account various factors, such as vehicle dynamics, crowd behaviour, weather conditions, and environmental factors. Road networks are typically created using digital maps that contain information on road geometry, traffic signs, and signals. Vehicles are modelled using physics-based simulations that take into account factors such as acceleration, braking, and steering.(Rajamani, 2011) Pedestrians are modelled using crowd simulation techniques that take into account factors such as crowd density, movement patterns, and individual behaviour.

Traffic scenario generation has many benefits, including improved safety, reduced costs, and increased efficiency. By testing self-driving vehicles and driver assistance systems

in a virtual environment, researchers and developers can identify and address potential safety issues before they occur in the real world. This can help reduce the risk of accidents and improve the overall safety of road users. In addition, virtual testing can be significantly less costly and time-consuming than physical testing, allowing researchers to iterate and refine their systems more quickly. Finally, traffic scenario generation can be used to optimize traffic flow and reduce,e congestion, improving the overall efficiency of our roads and streets.

Traffic scenario generation is also used in virtual reality training simulations, allowing individuals to practice driving and learn how to react in different traffic scenarios. This is particularly useful for new drivers, people with disabilities, or those who want to practice driving in a safe and controlled environment. In addition, traffic scenario generation can be used in the field of urban planning and design. Traffic simulation can help urban planners and architects visualize the impact of new buildings or infrastructure on traffic flow and safety. This can help to make informed decisions about urban development and design.

In summary, traffic scenario generation is a technology with many important applications. As the automotive industry continues to evolve and develop new self-driving car technologies, the need for more realistic and diverse virtual environments will continue to grow. Researchers and developers in this industry will need to continue to explore new simulation techniques and technologies to keep up with this demand. Ultimately, the goal is to create a safer, more efficient, and more sustainable transportation system for all.

1.1.3 Motivation for the Research

Autonomous driving technology has received significant attention in recent years because it can improve safety, reduce traffic congestion, and prevent energy waste. However, the application of autonomous driving technology faces many challenges, one of which is how to make autonomous driving systems better understand and adapt to complex traffic scenarios. My motivation is to achieve safer and clearer fast recognition of different elements in the surrounding environment in autonomous driving scenarios by using an advanced deep learning model, GAN (Generative Adversarial Network), which is a deep learning model that can be used to generate new data, improve model performance, and implement other complex tasks.(Aggarwal, Mittal & Battineni, 2021)

In high-quality autonomous driving, GAN can be used to recognize and synthesize images and videos in traffic scenes, thus helping autonomous driving systems to better understand and adapt to real-world complexities. GAN can be used to generate realistic traffic scenes. Autonomous driving systems need to make driving decisions in complex traffic environments, such as vehicles, pedestrians, road signs, and road markings. Using GANs, generators can be trained to create realistic traffic scenarios, helping autonomous driving systems to better learn how to adapt and respond to situations. Generate environments that emulate real roads and environments, allowing autonomous driving systems to better understand various situations on the road.

In addition, GAN can be used for image restoration and enhancement. In autonomous driving, the camera may produce damaged or blurred images due to various reasons, such as weather, dust, or dirt. This may make it difficult for autonomous driving systems to understand and adapt to road conditions, thus reducing their performance and reliability. The use of GAN can help repair these damaged or blurred images so that

the system can better understand the scene and take appropriate actions. GAN can also convert low-resolution images to high-resolution images, thus improving the quality of the input data to the autonomous driving system. The use of GAN helps to make the technology of autonomous driving more widely available in the future and to make people accept the implementation of autonomous driving in their daily lives.

1.2 Aims and Objectives

Our research aims to optimize the performance of image-to-image transformation models based on Generative Adversarial Networks (GANs), specifically focusing on generating high-quality semantic segmentation maps from input images. To achieve this, we will initially conduct a comprehensive analysis of two prominent GAN-based image-to-image transformation models: Pix2Pix(Wang et al., 2018) and Ganformer(Hudson & Zitnick, 2021), which have been widely used in the field. This analysis will serve as the foundation for our research, providing insights into the strengths, weaknesses, and potential areas for improvement of these models in the context of semantic segmentation map generation.

We will then further apply Pix2Pix and Ganformer, our analyzed models, in the field of autonomous driving. By leveraging the strengths of both models, we will propose a novel approach that combines the capabilities of Pix2Pix and Ganformer to create a new model for generating high-quality semantic segmentation maps in the context of autonomous driving. This proposed model will be designed to address the specific requirements and challenges of the autonomous driving domain, with the aim of achieving improved performance and accuracy compared to existing approaches. And through these techniques, we will replace and pair segment the environmental elements in autonomous driving scenes to achieve more accurate recognition of scenes

in the autonomous driving domain. We will also analyze and evaluate the experimental data through FID(Fréchet Inception Distance), DCI(Disentanglement, Completeness Informativeness) and other scores to evaluate whether we achieve better results based on the original.

1.3 Organization of Thesis

The thesis consists of Six chapters including an introduction, literature review, research-related methodology, preparation and design of experiments, experimental analysis and discussion, and conclusion.

The first Chapter introduces the research motivation, topic, and objectives. This chapter leads out the research topic which generates automated vehicle testing traffic scenes using generative adversarial network.

In Chapter 2, we provide a literature review of the field of autonomous driving and techniques such as the application of generative adversarial networks in the field of self-driving cars. Specifically, we identify the importance of GANs in the autonomous driving domain and some of the problems that exist in the autonomous driving domain, and in this chapter, learn that GANs can help solve some of the problems in the autonomous driving domain.

In Chapter 3, links to research methods related to our research problem are highlighted. We state the specific generative adversarial network techniques we are about to use and the related methods, functions, etc. The limitations of the techniques we use are specifically highlighted in this chapter.

In Chapter 4, we will provide detailed information about the specific equipment, dataset, and experimental procedures used in our research. This chapter will serve as a comprehensive guide to replicate our experiments and understand the technical aspects of our study.

In Chapter 5, we will conduct a critical analysis of our results and discuss how they relate to research objectives. We will provide an in-depth analysis of our experiments, including the specific assessment scores used, and present the results in a comprehensive manner, allowing for a thorough understanding of the experimental outcomes.

In Chapter 6, we summarize our study and provide a summary of this study and attaches future research directions inspired by this study.

Chapter 2

Review of Automated Driving and Generate Adversarial Network

2.1 Review of Automated Driving Related Technique

The technology of self-driving cars or Autonomous Vehicles (AV) has the potential to trigger a significant transformation in the transport system with wide-ranging consequences for our economy and society. It is generally accepted that there are six levels of automation as defined by the Society of Automotive Engineers (SAE). Level 0 automation mainly provides warnings to drivers such as lane departure warning to drivers. Levels 1 and 2 automation includes steering and/or braking and acceleration support such as automatic emergency braking, and adaptive cruise control. These levels of automation can also be seen in some vehicles, especially higher-end ones, available commercially today. The current push is to implement level 3 automation, where the vehicles take over the control completely from the driver under some conditions but allow the driver to take over when requested. At level 4 and 5 automation, the driver will not be required to take over control and the vehicle can drive itself under limited and all conditions respectively. This research focuses on the testing of the automation

functions of AVs at levels 3-5. (International, 2018)

Research on autonomous driving technology dates back to the 1960s, initially to achieve the military goal of being driverless. As self-driving technology has evolved, its application scenarios have expanded from a single military use to include everyday transportation, such as driverless cars and flying vehicles. Among them, driverless cars are one of the most widely used areas, and major car manufacturers and technology companies have ventured into this field. Researchers have proposed a variety of different autonomous driving technologies, which include sensor-based technologies, vision-based technologies, machine learning-based technologies, etc.

Sensor-based autonomous driving technology focuses on acquiring information about the vehicle's surroundings through sensors such as LIDAR, cameras, and ultrasound, and then using this information to make autonomous driving decisions (Barua, Gomes, Baghe & Sisodia, 2019). In this technology, LIDAR is one of the most important sensors, which can detect objects around the vehicle with high accuracy and help the autonomous driving system to make accurate decisions. In addition, cameras and ultrasonic sensors can also provide useful information, such as the vehicle's speed, direction, and distance.

Vision-based autonomous driving technology uses cameras to acquire real-time video data and uses computer vision techniques to analyze this data to make autonomous driving decisions (Zablocki, Ben-Younes, Pérez & Cord, 2022). This technology has the advantage of enabling autonomous driving using relatively inexpensive hardware, but its reliability and robustness still need further improvement due to the complex environment and changing light conditions.

Machine learning-based autonomous driving technology is used to implement autonomous driving decisions by training deep learning models. One very promising technique is generative adversarial networks (GAN), which can generate realistic image and video data to provide more accurate input information for autonomous driving systems (Fujiyoshi, Hirakawa & Yamashita, 2019).

Since an AV integrates both the mechanics of a vehicle and the intelligence of the driver into a single entity, its automated driving functions must be adequately tested to ensure the safety of the public. The ultimate test of the safety of an AV is to drive it in real traffic and observe how it reacts to various traffic situations. However, how much driving is required to adequately evaluate its safety has been hotly debated. The authors concluded that in order to make statistical safety comparisons with the performance of human drivers, an AV needs to be driven at least hundreds of millions of miles (Koopman & Wagner, 2017). This will take tens of years and is impractical. Hence, the focus of automated vehicle testing has shifted from a driving distance-based to a scenario-based approach (Kalra & Paddock, 2016). With this approach, various traffic scenarios are constructed and simulation of the AV driving in these scenarios is carried out in a computer.

Scenario-based testing requires the design of appropriate scenarios (Neurohr et al., 2020). A driving scenario is typically made up of both static (non-moving) and dynamic (moving) elements. These include the environment (e.g. road widths, number of lanes, road curvature, road signs, pedestrian sidewalks), a set of vehicles and their initial states (position, orientation, speed, etc), pedestrians, traffic signals, and the ego AV (the AV under test) with its initial state and ultimate goal. It can also be viewed as a time sequence of scenes (Menzel, Bagschik & Maurer, 2018). A set of test scenarios could be designed for each operational design domain (ODD) - the operating conditions under

which certain functions of an AV are designed to function. Together they form a test scenario library (Feng, Feng, Yu, Zhang & Liu, 2020).

Manually creating test scenarios is a very time-consuming exercise. For this reason, some research on the automatic generation of scenarios has been reported (Yue, Shi, Wang & Lin, 2020). However, a lot of the scenarios that are generated are not able to fully test the ability of the AV to handle critical situations (Nalic, Eichberger, Hanzl, Fellendorf & Rogic, 2019). Hence, some of the research in this area has proposed methods to generate critical or challenging test cases (Mullins, Stankiewicz & Gupta, 2017). The problem is usually formulated as an optimization problem. A measure of "criticality" of a scenario is arbitrarily defined and used as the objective function. The optimization procedure will search through the space of all possible combinations of the parameters (of the dynamic elements) in the scenario and arrive at a solution (Klischat & Althoff, 2019).

There are two main issues with the optimization approach described above. Firstly, the search space increases exponentially as the traffic scene becomes more complex. The higher the dimensionality of the problem, the more computationally demanding the process becomes. Secondly, there is no guarantee that the optimization procedure will converge to a solution. The optimization problem may be non-convex. The examples given by the research cited above are relatively simple. For instance, it may involve a straight section of road with two lanes and three other cars. In this case, the number of variables is relatively small, and a solution could be reached. It remains to be seen how these procedures will scale with the complexity of the scene. Lastly, deterministic optimization procedures will arrive at only one solution while in practice, we want the system to generate many different critical test cases given a single scene.

The research proposed here aims to make use of machine learning methods to overcome these issues. A deep learning approach, which is very effective in a lot of application domains, will be taken. More specifically, the use of a Generative Adversarial Network (GAN) will be explored. A GAN can generate new data based on the statistics of the training dataset. It has been applied in many ways – to generate realistic artificial faces,(Varkarakis, Bazrafkan & Corcoran, 2020) create music,(Engel et al., 2019) and perform the text to image generation,(H. Zhang et al., 2018) among others. Thus, it can potentially be applied to test scenario generation by generating a variety of new test-critical test cases and scenarios. The outcome of this research will greatly enhance our ability to automatically provide effective test cases for AV validation, thereby enabling the safe deployment of AV in our society.

2.2 Review of Generate Adversarial Network

Generative Adversarial Networks (GANs) are a powerful class of deep learning models that can generate many types of data such as images, video, and audio with high fidelity (Goodfellow et al., 2020). In the field of autonomous driving, GANs have been widely used to generate highly realistic driving scenarios to help improve the performance and reliability of autonomous driving systems (Aggarwal et al., 2021). On the other hand, GAN is used to generate highly realistic image and video data to enhance the perception and decision-making capabilities of autonomous driving systems. In the testing and evaluation of autonomous driving systems, GAN can generate virtual driving scenario data with a high degree of realism. These data can cover various driving situations, such as city roads, highways, intersections, complex weather, etc. These scenario data can be used for performance testing, safety evaluation, and algorithm optimization of autonomous driving systems (Uricár et al., 2019).

In addition, GAN can help solve the difficulties of realistic scene data collection, for example, it is difficult to collect driving scene data in winter, night, or in dangerous areas, etc., when using GAN to generate virtual data can effectively reduce the difficulty and cost of data collection. On the other hand, in terms of the perception and decision-making of autonomous driving systems, GAN can be used to generate highly realistic image and video data to improve the perception and decision-making ability of autonomous driving systems. For instance, in complex weather or nighttime conditions, the perception ability of an autonomous driving system can be affected to a certain extent, when using high-quality image data generated by GAN can effectively improve the perception ability.

GAN can be also used to generate target objects such as vehicles and pedestrians with high realism to improve the perception and decision-making ability of autonomous driving systems for complex traffic scenarios. There is still room for the application of GAN in the field of autonomous driving, especially in data enhancement and data synthesis. Since autonomous driving requires a large amount of data to train deep learning models, and real data is often difficult or insufficient to obtain, using GAN to generate synthetic data has become a popular direction. A study shows that data generated using GANs can effectively improve the performance of autonomous driving models. The study used CycleGAN to transform a sunny driving scenario into a rainy driving scenario and then mixed the generated data with real rainy driving data to train a deep learning model (Cai, 2022). The results show that using the generated data improves the performance of the model and reduces the error of the model in the rainy weather scenario. Similar studies have shown that using GAN-generated data can improve the robustness of the model and make it more adaptable to a variety of different scenarios.

Futhermore, GAN can be used for data synthesis. For example, various driving scenarios are simulated in a simulator and synthetic data is generated, which can be used to train deep learning models. This approach can effectively reduce the need for real data, thus reducing the cost of data acquisition. In addition to data augmentation and data synthesis, GAN can be used for other aspects of autonomous driving, such as decision-making and planning. One study used GAN to generate car trajectories with different speeds and directions and then used these trajectories to train the planner of an autonomous driving system. The results show that this approach can significantly improve the performance of the planner. GAN has a wide range of applications in the field of autonomous driving, including image generation, data enhancement, target detection, and trajectory prediction (Yu, Sun, Zhou & Liu, 2019).

First, GAN is widely used for image generation. In autonomous driving, it is very useful to enhance the robustness and reliability of algorithms by generating realistic synthetic images. For example, the performance of an autonomous driving system is improved by semantic segmentation of objects in the scene and then using GAN to generate images with higher resolution and more realistic details. Many studies have also shown that using GAN to generate images can reduce the need for real data, thus reducing development costs.

Second, GAN has also been applied to data augmentation. In autonomous driving, data is the basis for training algorithms, but it is very expensive to obtain large amounts of real data in reality. Therefore, using GAN to generate virtual data to increase the number and diversity of samples in the dataset, thus improving the accuracy and robustness of the algorithm, is a feasible approach.

In addition, GAN can also be applied to target detection. Target detection is a core

task in autonomous driving and can be used to identify objects such as pedestrians and vehicles on the road. GAN can also be used for trajectory prediction. In autonomous driving, trajectory prediction is the key to achieve autonomous navigation. Using GAN to generate synthetic trajectories can improve the accuracy and robustness of the algorithm and reduce the need for real data. GAN, as a powerful generative model, has also been widely used in the field of autonomous driving. The following are some of the literature on GAN in the field of autonomous driving.

2.2.1 Autonomous Driving GAN

Xiong et al. propose an innovative approach to balance the relationship between data sharing and personal privacy in autonomous driving systems by applying GAN techniques to protect the privacy of vehicle camera data. The paper first introduces the rapid development of autonomous driving technology and highlights the importance of vehicle camera data. However, as the number of autonomous vehicles increases, protecting the privacy of vehicle camera data becomes particularly important. Current autonomous driving systems usually require vehicle camera data to be transmitted and shared in real time, which may lead to the risk of personal privacy leakage. Therefore, this paper proposes a GAN-based approach called Autonomous Driving GAN (ADGAN) to protect the privacy of vehicle camera data. The core idea of the method is to use GAN to generate synthetic camera images instead of real vehicle camera data. Specifically, the authors design a generative network and a discriminative network. The generative network is responsible for generating realistic synthetic images to replace the original images. The discriminant network is used to evaluate the differences between the generated synthetic images and the real images. By iteratively training the generative and discriminative networks, more realistic synthetic images can be obtained, thus preserving the privacy of vehicle camera data. The paper also provides

an experimental demonstration of the proposed method. The experimental results show that the synthetic images generated using this GAN-based method can hide the personal privacy information in the vehicle camera data while maintaining important features. In addition, the real-time performance of the method is verified in an autonomous driving system.(Xiong, Li, Han & Cai, 2019)

AD-GAN consists of two main modules: a generator and a discriminator. The generator is a neural network that receives some input information (e.g., vehicle speed, acceleration, lane information, etc.) and generates a simulated image that simulates various situations in a scene, such as moving vehicles, buildings, pedestrians, etc. To ensure the quality and realism of the generated images, the generator uses deep learning techniques such as convolutional neural networks (CNN) and recurrent neural networks (RNN), as well as some image processing techniques. The discriminator is another neural network that takes in the real image and the image generated by the generator and tries to distinguish between them. When training, the discriminator's goal is to minimize the number of times it incorrectly labels a generator-generated image as the real one. Meanwhile, the goal of the generator is to deceive the discriminator as much as possible so that it cannot accurately distinguish the generated image from the real one.(Xiong et al., 2019)

The advantage of AD-GAN is that it can generate highly realistic images, which is very useful for the development and testing of autonomous driving technologies. In addition, it can also generate different scenarios by adjusting the input parameters, which is also useful for testing the robustness of autonomous driving technologies in different situations. However, AD-GAN also has some limitations. For example, the generator may not be able to generate some rare scenarios or anomalous situations due to the limitation of the dataset. In addition, the training of the model requires a

lot of computational resources and time, so specialized hardware and algorithms may need to be used to optimize the training efficiency. By comparing the performance of ADGAN with the state-of-the-art, the advantage of ADGAN can be verified and can provide an effective trade-off between the recognition function and the privacy protection of camera data. While reading the paper, I discovered that I had problems understanding the model's formulas and other issues very quickly. By reading the paper, they investigate some safety flaws of autonomous driving technology by using the technique of GANs and thus propose the framework of ADGAN, but for the technical aspects of autonomous driving, this framework helps to make autonomous driving safer in practical applications. The comparison between ADGAN and the other four techniques shows that the image quality of ADGAN is indeed better than several other techniques, and the performance of ADGAN evaluated by real datasets.(Xiong et al., 2019)

2.2.2 Spatiotemporal Coherence-based GAN

STC-GAN is a spatiotemporal coherence-based generative adversarial network designed to generate high-quality images of traffic scenes for application in data enhancement of autonomous driving systems (Qi, Wang, Li & Luo, 2020). It can generate high-quality images from time-series data, and thus can be used to generate information about the road, vehicles, and pedestrians around the vehicle to improve the accuracy and robustness of autonomous driving systems. The network addresses the challenges in scene generation by introducing spatial and temporal consistency. Spatial consistency means that the generated images should be consistent with the spatial layout of the real scene, while temporal consistency requires that there should be continuity and consistency between the generated images at different times.

To achieve this goal, STC-GAN uses a structure consisting of two generators and two discriminators. One generator is used to generate images for the current time step and the other generator is used to generate images for the next time step. The two discriminators are used to evaluate the authenticity of the generated images for the current time step and the next time step, respectively. A key innovation of STC-GAN is the introduction of a temporal attention mechanism to handle time series data. This mechanism allows the network to focus on the relationship between the current time step and the previous time steps when generating images to improve the continuity and consistency of the generated images.

STC-GAN employs conditional image generation to enhance the diversity and realism of the generated images. In the experiments, STC-GAN is applied to traffic flow prediction in autonomous driving scenarios. The results show that STC-GAN has a significant advantage in generating high-quality scene images and can achieve better results in traffic flow prediction compared with other comparison methods.STC-GAN provides an effective data enhancement method for autonomous driving systems to generate high-quality, diverse, and continuous traffic scene images, thus improving the accuracy and robustness.

2.2.3 Conditional GAN

Kumar et al. investigated the conditional GAN (cGAN) machine cognitive framework for image transformation for self-driving cars (Kumar & Birajdar, 2018). This technique contributes to some extent to the development of autonomous driving technology, firstly in the paper a comparison between L1 and CGAN is used to identify contaminated images in experiments that can better trigger the linguistic cut of dirt and turn on the cleaning function of the lens, which contributes to the safety of smart driving cars on

the road.

In CGAN, the inputs of both the generator network and the discriminator network contain conditional information. In the generator network, the conditional information is connected with the noise vector and passed to the hidden layer of the generator network. In the discriminator network, the conditional information is connected to the input image and passed to the hidden layer of the discriminator network. The training process of CGAN is similar to that of traditional GAN, where the network is continuously optimized by alternately training the generator and discriminator networks. In each training iteration, the generator network generates some fake samples based on the conditional information, and then the discriminator network evaluates the truthfulness of these fake samples and gives a decision. Then, the network updates the weights of the generator network and the discriminator network based on the results of the determination. Through continuous iterative training, CGAN can generate more realistic images and generate the corresponding images based on the given conditional information. This literature contributes to a large extent to the safety of smart driving technology on the road and its identification in collisions.

Conditional GAN(CGAN) is a variant of generative adversarial networks in which both the generator and the discriminator receive a conditional vector as input. This conditional vector can be used to control the properties or features of the data generated by the generator. The following are some advantages and disadvantages of CGAN: First advantage is that control the generation process: CGAN allows the introduction of conditional information in the generation process, which makes it possible to precisely control the attributes or features of the generated data. By passing different information in the condition vector, it is possible to generate samples that meet specific conditions, such as generating specific kinds of images or images with specific features. Improved

sample quality: Because CGAN is able to use conditional information in the generation process, it is usually able to generate higher quality samples. By providing additional conditional information, the generator can better understand the structure and properties of the data, thus generating more realistic and accurate samples. Multimodal generation: CGAN is capable of generating multimodal data, i.e., generating multiple possible outputs under a given condition. This is useful for some tasks such as image-to-image translation, text-to-image generation, etc. By introducing different conditions in the condition vector, multiple samples with different styles, angles or variations can be generated. The next disadvantage is: increased training difficulty: the training process of CGAN is more complex compared to traditional generative adversarial networks. Due to the need to train both generators and discriminators and the additional introduction of conditional vectors, training CGAN may require more computational resources and time. Conditional information requirement: The performance of CGAN is highly dependent on providing accurate and informative conditional vectors. If the conditional vector does not have sufficient information or incorrect information, the generator may not be able to correctly generate samples that meet the desired conditions. Restricted to conditional information: The CGAN generation process is limited by the information provided by the conditional vectors. If the condition vector does not contain information about certain key features or attributes, the generator may not be able to generate the corresponding samples. This may lead to limitations in the generated results.

2.2.4 Cyclic Consistent GAN

Ahmad et al. proposed a Cycle GAN (Sallab, Sobh, Zahran & Essam, 2019), which is a model for modeling LiDAR. cycleGAN is an image translation model that does not require paired data and is able to translate images in one set of domains to images in another set of domains. Unlike traditional GAN models, CycleGAN does not require

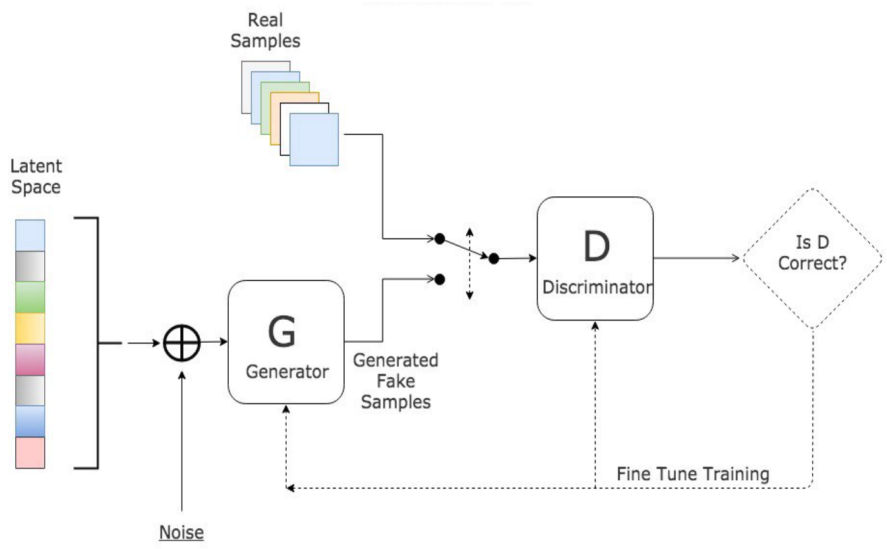

Figure 2.1: Conditional GAN

each input image to correspond to a target image. Instead, CycleGAN requires only a large amount of data from two domains for training, and then can transform images in one domain to images in the other domain without the need for pairs of data.

CycleGAN is based on the GAN architecture, which uses two generators and two discriminators for training. One of the generators converts an image in one domain to an image in another domain, while the other one converts the image in the other domain back to the original domain. This two-way conversion ensures that the translation process is a one-to-one correspondence. Two discriminators are responsible for determining whether the images generated by the generators are true or not. By using the carla simulator to sample the experimental data for that model, the experiments were made easier to complete by transforming the data collected from the carla and KITTI datasets, saving a lot of time and cost. In that article, the experiments were not completed and no task loss was derived for CycleGAN, which is a drawback of that article, but at the end of the article, a specific next step for the experiments is given. It is also proposed that a method will be discussed to evaluate the quality of LiDAR and

to extend CycleGAN's task feature loss.

Isaac Ogunrinde et al. argue that autonomous driving technology still faces some huge challenges in adverse weather conditions (Ogunrinde & Bernadin, 2021). Object detection in adverse weather conditions has become one of the most critical issues facing autonomous driving technology. A large number of existing technologies do not pay attention to this and tend to perform superior only in good weather conditions. Based on this problem, they focused on foggy weather conditions. And they investigated whether the quality of defogging and recovering foggy weather images affects the performance of the system. They used a cyclic consistent generative adversarial network (CycleGAN) based image defogging technique to defog and improve the visibility and quality of haze images. After several comparisons under fog-free, medium fog, and heavy fog conditions, the main idea of CycleGAN is to train two GAN models: one to transform images from domain A to domain B (called generator G) and one to transform images from domain B to domain A (called generator F). These two generators are simultaneously trained by discriminator D, which is used to evaluate whether the images generated by the generators are realistic or not.

To ensure that the transformed images retain the semantic information of the original images, CycleGAN introduces cyclic consistency loss. This means that the transformation from A to B and then to A should be very similar to the original image A, while the transformation from B to A and then to B should be very similar to the original image B. In this way, CycleGAN can generate images that are realistic and retain semantic information. The results show that the defogging effect is superior and the detection performance is significantly improved under the medium fog condition. However, it is not as effective under foggy conditions. This proves that CycleGAN still has room for improvement. If I want to test this network in my future research, I need to make

appropriate modifications to improve the performance.

2.2.5 Conditional Multi-Generator GAN

Lekic et al. investigated the use of the Conditional Multi-Generator GAN (CMGGAN) technique to help generate more accurate data from radar sensors in smart driving technology (Lekic & Babic, 2019). Conditional multi-generator GAN (CM-GAN) is a type of GAN designed to improve the generation of high-resolution images with large variations. The key idea of CM-GAN is to use multiple generators, each specializing in generating a certain subset of the data. This allows for greater diversity in the generated images and reduces the possibility of pattern collapse.

In CM-GAN, each generator is conditioned on a different subset of the latent code. This allows each generator to focus on a specific aspect of the data, such as texture or shape. The generators are trained together with a shared discriminator, which is responsible for distinguishing between real and false images. During training, the generator and discriminator are updated in an alternating fashion. The purpose of the generator is to deceive the discriminator by generating an image that is indistinguishable from the real image. On the other hand, the discriminator aims to correctly classify real and false images.

One of the main advantages of CM-GAN is that it can generate a variety of high-quality images. This is particularly useful in applications that require a large number of different images, such as in computer vision tasks. In addition, the use of multiple generators can improve the robustness of the model to changes in the input data. This is an unsupervised machine-learning method. In the experiments of this paper, the images generated by the CMGGANs technique are used to enhance the images or enhance the

robustness of the camera by semantic segmentation. By comparing the experiments, CMGGAN exists lower values compared to CGAN and InfoGAN. That is, CMGGAN has the ability to learn to generate differentiated features, which makes the generated images high quality.

2.2.6 Lane-GAN

Lane-GAN is an image processing technique based on a generative adversarial network (GAN) for road marker generation in autonomous driving (Liu, Wang, Li, Li & Zhang, 2022). Lane-GAN mainly implements image generation and discrimination by training two neural networks. The generative network generates images from random noise, and the discriminative network is responsible for discriminating whether the generated images are real road marker images. In Lane-GAN, the generative network adopts the idea of conditional GAN (CGAN), which means that the input conditions (e.g., road shape and lane line type) control the properties of the generated images. During the training process, the generative network continuously generates images and feeds them into the discriminative network to discriminate them, while the discriminative network is responsible for distinguishing whether the generated images are real road marking images or not.

By continuously optimizing the parameters of the two networks, Lane-GAN can gradually improve the quality of the generated images and thus obtain more realistic road marking images. Compared with the traditional rule-based road marker generation methods, Lane-GAN has the following advantages. First of all is adaptive. Traditional methods often need to rely on accurate map data and road information, while Lane-GAN can generate corresponding road marker images based on real-time scenes. Second is robustness. Traditional methods tend to fail for complex scenes, while Lane-GAN

can adapt to various changes in road shapes and traffic markings. Third is efficient. Traditional methods consume a lot of computational resources and time, while Lane-GAN's generation process can be completed in a shorter time. Lane-GAN has a wide range of application prospects in the field of autonomous driving and can be used in the generation of road markings, detection of lane lines, and production of high-precision maps. The continuous development and optimization of this technology will hopefully provide powerful technical support for realizing highly automated driving.

2.2.7 Recurrent Conditional GAN

The architecture of Recurrent Conditional GAN(RCGAN) consists of two networks, a generator, and a discriminator, that are trained together in an adversarial way (Arnelid, 2018). The generator takes as input a noise vector and a sequence of condition vectors that encode the desired output and generates a sequence of images. The discriminator takes as input a sequence of images and a corresponding sequence of condition vectors and outputs a probability that the sequence is real or fake. The key innovation of RCGAN is the use of RNNs to model the temporal dependencies in the image sequences. Specifically, the generator network uses a variant of the long short-term memory (LSTM) architecture to generate the image sequence, and the discriminator network uses a similar RNN architecture to classify the real and fake sequences.

As Arnelid (2018) in his research developed and used Recurrent Conditional GAN (RCGAN) to model the sensor errors, firstly the experiments were implemented through Tensorflow in python to build the model and in the final results, it can be seen that the RCGAN framework produces good results in training and can learn the distribution. In this literature, all the data are real data from Volvo cars, so it is helpful for smart driving technology. Moreover, the experiments in the literature need further improvement so

that the RCGAN framework can be further improved. In the process of reading the literature, I do not have a deep understanding of this model, so I need to learn it more deeply in the next research.

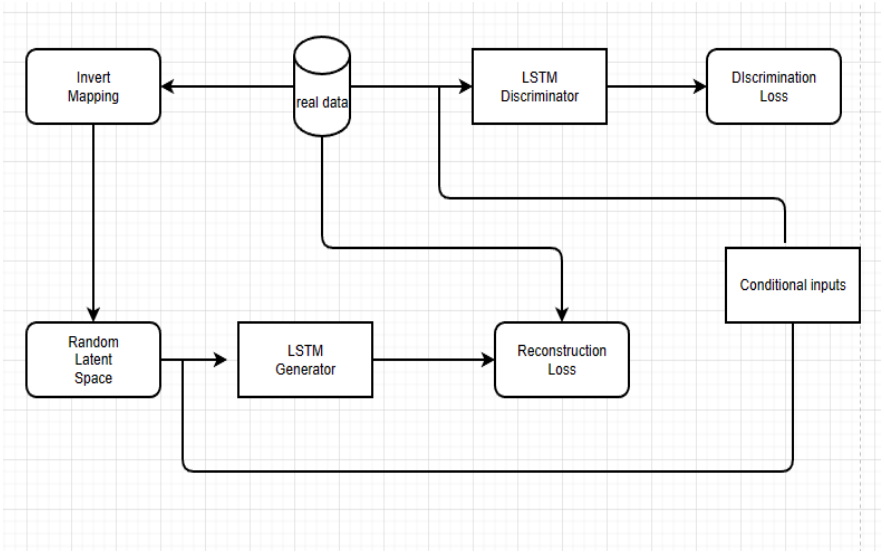

Figure 2.2: Model of Recurrent Conditional GAN

2.2.8 Driving GAN

Cameron Fabbri proposed a GAN-based Driving-GAN (D-GAN) neural network framework in 2018 (Fabbri & Sharma, 2018). The main idea behind D-GAN is to generate synthetic driving scene images and videos that can be used to train and test autonomous driving systems. D-GAN consists of two main components: a generator and a discriminator. The generator is responsible for generating synthetic images and videos that are similar to real driving scenes. It takes random noise as input and generates images and videos that are then fed to the discriminator. The discriminator's job is to distinguish between real and synthetic images and videos. It is trained to correctly identify real images and videos and to reject synthetic ones. D-GAN also includes a novel loss function that encourages the generator to produce images and videos that are not only visually similar to real driving scenes but also physically realistic. This is achieved by

incorporating a physics-based loss term that penalizes the generator if the generated images and videos violate physical laws, such as the conservation of momentum.

Overall, D-GAN is a promising approach to generating synthetic driving scene data for autonomous driving applications. By providing a way to generate large amounts of diverse and physically realistic data, D-GAN can help to improve the performance and safety of autonomous driving systems. The network can generate an approximate policy from the data and generate realistic actions based on the framework. In addition, they also introduced some game theory perspectives. It is worth noting that the technology is applied in the game, the purpose is that the game behavior is close to real human behavior. This is helpful to autonomous driving technology to a certain extent. First of all, they use inverse reinforcement learning in the article to achieve a method that approximates human behavior in the GAN network. It is worth noting that the experiment was not aimed at finding a perfect driver. Therefore, it is necessary to pay attention to this problem in the final application. The literature goes a long way towards helping intelligent driving technology to resemble a human more closely on the road.

2.2.9 StackGAN++

StackGAN++ is a generative adversarial network (GAN) model that was proposed for realistic image synthesis by Han Zhang et al. in 2017 (H. Zhang et al., 2018). The model is an extension of the original StackGAN model that was introduced earlier in the same year. The StackGAN++ model generates high-resolution images from text descriptions, where the text is first converted into a low-resolution image, which is then progressively refined into a high-resolution image.

The StackGAN++ architecture consists of two stages of GANs, where each stage

generates an image of a different resolution. The first stage generates a low-resolution (64x64) image from the text description using a text-conditional GAN. The second stage takes the low-resolution image as input and generates a high-resolution (256x256) image using a conditional GAN with a novel conditioning augmentation technique that improves the quality of the generated images. The conditioning augmentation technique involves generating multiple random noise vectors and using them to condition the generator network. This helps to diversify the output images and improves the quality of the generated images. The authors also introduce a perceptual similarity loss, which encourages the generated images to match the perceptual features of real images, resulting in more realistic images. The StackGAN++ model was evaluated on several benchmark datasets, including Oxford-102 flowers and CUB-200-2011 birds, and achieved state-of-the-art results in terms of image quality and diversity. The model was also able to generate novel and diverse images that were not present in the training dataset. StackGAN++ is a powerful and effective model for realistic image synthesis that can generate high-resolution images from text descriptions with high quality and diversity.

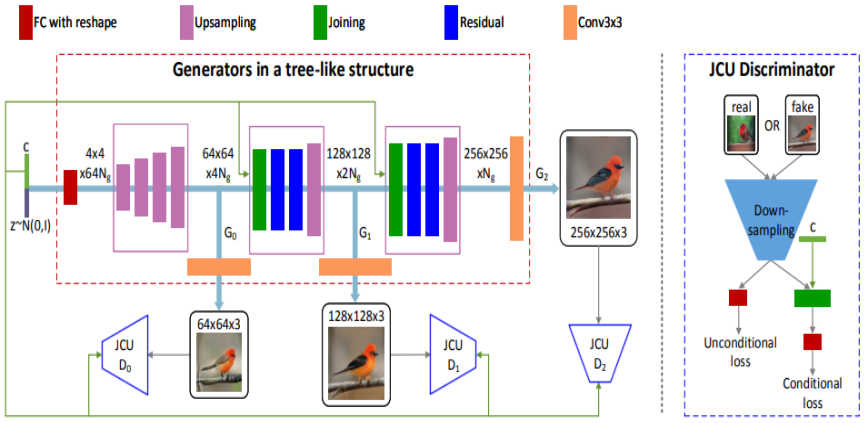

Figure 2.3: Structure of Stackgan

2.2.10 EL-GAN

EL-GAN is a generative adversarial network-based lane detection method that aims to solve the robustness problem of traditional lane detection methods under complex backgrounds, lighting changes, weather changes, etc. EL-GAN employs embedding loss to guide the generator to generate more accurate lane lines (Ghafoorian, Nugteren, Baka, Booij & Hofmann, 2018). In EL-GAN, the generator is trained to map the input image to the lane line image, and the discriminator is trained to distinguish the generated lane line image from the real lane line image.

In addition, EL-GAN introduces an embedding loss to constrain the distance between the generated lane line image and the real lane line image in the feature space to enhance the accuracy of the lane line image. The generator of EL-GAN consists of two parts: an encoder and a decoder. The encoder converts the input image into an intermediate representation, and the decoder converts this intermediate representation into an output image. The intermediate representation is shared between the encoder and decoder, which allows EL-GAN to learn how to extract useful features from the input image. The discriminator of EL-GAN also uses a traditional GAN structure, but during training, it classifies and reconstructs the image simultaneously and uses the reconstructed image to help distinguish between the real image and the generated image. This approach can effectively mitigate the pattern collapse problem in GAN. The training process of EL-GAN is divided into two stages: the first stage is to train the encoder and decoder to generate low-quality images, and the second stage is to use the SSIM loss function to fine-tune the generator to generate higher-quality images. In addition, EL-GAN uses an adaptive learning rate adjustment strategy to accelerate the training process. The main advantage of EL-GAN is that it can improve the accuracy and robustness of lane detection, especially in complex environments. However, the disadvantages of EL-GAN

are a large amount of training data and computational resources required, as well as the complicated hyperparameter tuning process. Overall, EL-GAN provides a new and effective approach to lane detection task with great potential for application.

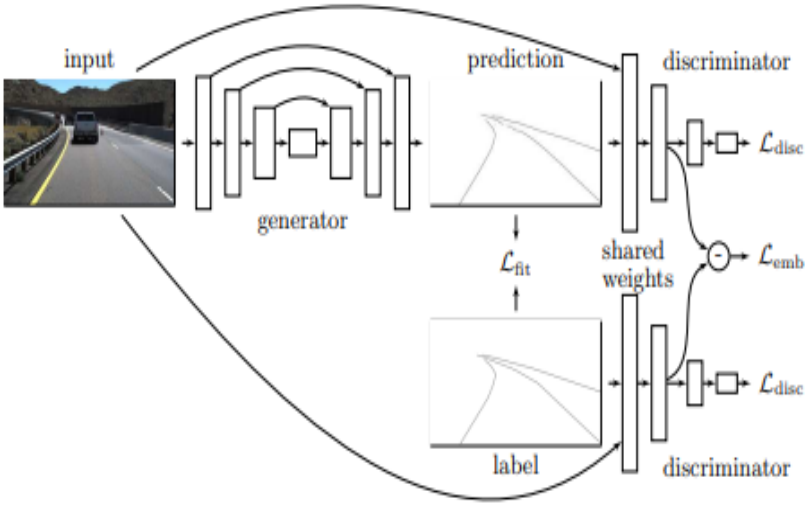

Figure 2.4: Structure of EL-GAN

2.2.11 Other GANs

Han et al. propose a new method for detecting vehicles in images using deep learning techniques (Han, Su & Zhang, 2019). The paper addresses the challenges of detecting vehicles in complex environments, such as occlusions, changing lighting conditions, and different vehicle sizes. The proposed method combines multiple convolutional layers and generative adversarial networks (GANs) to improve vehicle detection performance. Multiple convolutional layers are used to extract layered features from the input images, while generative adversarial networks are used to generate synthetic images to enhance the training dataset and improve the generalization ability of the model. The multi-convolutional layers consist of several convolutional layers with different kernel sizes, which enable the model to capture features at different scales. The outputs of these layers are concatenated and fed into a fully connected layer for classification.GANs are used to

generate synthetic images that can be used to augment the training dataset. The generator network of GANs is trained to generate realistic images, while the discriminator network is trained to distinguish between real and synthetic images. By using synthetic images to train the model, the proposed method can improve the generalization ability of the model and reduce the risk of overfitting. Experimental results show that the proposed method outperforms existing methods in terms of detection accuracy and robustness.

The proposed method achieves an average accuracy (mAP) of 91.7 percent on the KITTI dataset, which is a widely used benchmark for vehicle detection. In conclusion, this literature is a promising approach to improving the accuracy and robustness of vehicle detection in complex environments. The proposed method innovatively combines deep learning techniques and achieves state-of-the-art results on a benchmark dataset. The method has the potential to be applied in various fields, such as autonomous driving, traffic monitoring, and security surveillance.

In (M. Zhang, Zhang, Zhang, Liu & Khurshid, 2018), Mengshi Zhang et al. tested the GAN-based deformed autonomous driving technique. DeepRoad is a GAN model for road image generation in autonomous driving scenarios. It is based on the Pix2Pix architecture and uses a residual network to improve the effectiveness and speed of image generation.DeepRoad is proposed by Alibaba AI Lab to generate high-resolution road images for decision-making and planning in autonomous driving scenarios. The training dataset of DeepRoad is derived from the fusion of high-precision maps and actual scene images, which can accurately simulate road scenarios under different weather, time of day, and traffic conditions. DeepRoad's training process uses adversarial training, where a generator and a discriminator are trained simultaneously, allowing the generator to generate more realistic road images and the discriminator to discriminate between real road images and generated road images.DeepRoad's generator uses a residual network

to improve the effectiveness and speed of image generation. A residual network is a neural network connected across layers that learns the residuals (i.e., the difference between the output and the input) to predict the output more accurately. The test results show that many automatic generation techniques have safety problems and generate images with low realism, which ultimately affects the safety and accuracy of the results. For this reason, they proposed an unsupervised learning framework called Deep Road. The framework can form a large number of accurate driving scenarios to help test the safety of autonomous driving systems. They tested three different autonomous driving models and produced different results. This demonstrates that their proposed framework can effectively test the accuracy of autonomous driving technology. It proves that the framework has good application prospects.

Daiki Shiotsuka et al. proposed an autonomous driving technology that can better recognize natural night scenes (Shiotsuka et al., 2022). They believe that the robustness and high accuracy of the entire system are achieved through the extensive use of deep learning in existing autonomous driving technologies. However, many systems are based on CNN methods, and these methods often require a large amount of data to be implemented. However, most of the existing datasets that can be used for autonomous driving training are constructed based on daytime scenarios, and few datasets are constructed based on adverse conditions such as nighttime driving.

Furthermore, Generative Adversarial Network (GAN) models perform well in various image translation tasks. Considering that data information may be lost in time transformation, for example, the conversion of day and night will bring some information loss. Therefore, they propose an image translation framework based on GAN. The framework maintains semantic consistency across changes by migrating the semantic segmentation network to GAN. The final experimental results show that their proposed

method can produce natural night scene images better than various past methods. When reading the paper, I found that the formula of the model can be understood relatively easily. Only a small number of formulas need to be consulted before they can be understood. By reading the paper, I understood their idea of using the technology of GANs to generate nighttime images. This kind of thinking can better help me understand other autonomous driving technologies.

As Weihuang Xu et al. Put forward in 2021, the rapid development of auto-driving technology benefits from the improvement of artificial intelligence technology (Xu, Souly & Brahma, 2021). The main improvement of artificial intelligence technology lies in some machine learning technologies and deep learning technologies. These technologies can help self-driving systems learn and update themselves. In fact, a lot of training is often required before an autonomous driving system is officially launched. People will build many models and special cases based on the actual situation for the system to learn. However, these models often have problems in practical use, such as model scenarios, small data sets and other problems. These problems are collectively referred to as the long tail problem.

In order to reduce the occurrence of these problems, GANs can be used to generate real data in autonomous driving perception tasks. In this paper, Xu et al. demonstrate the reliability and usability of GANs in addressing long-tailed problems. By setting up different experimental procedures such as nighttime and daytime scene training, scene classification, etc. Finally, it is basically reliable to apply the data generated by GANs to train and evaluate the vision module for autonomous driving. While reading the paper, I found that there were some small problems in the part where I understood the experimental design, and I didn't quickly understand the purpose of their experiment design at the beginning. But after reading the article many times, it was finally solved.

It is clear from reading the paper that they verified the reliability of GANs technology in autonomous driving through specific experiments, thus proposing a new loss function to help GAN focus on the improvement of the target detection module. Additionally, they show how to use GANs to simulate rare traffic scenarios. This has made a huge contribution to the study of automata as technology.

2.3 Summary

In conclusion, GAN has become a promising technology in the field of autonomous driving. It has been used for various applications such as image generation, data enhancement, and simulation of real-world scenarios. GAN-based models, such as pix2Pix, CycleGAN, and Driving-GAN, have shown good results in generating realistic and diverse driving scenarios.

GAN-based data augmentation techniques have been shown to improve the performance of object detection and lane detection models, which are key components of autonomous driving systems. In addition, GAN-based simulation models can be used to generate large datasets of synthetic driving scenarios that can be used to train autonomous driving systems safely and cost-effectively.

However, there are still several challenges that need to be addressed before GANs can be fully integrated into autonomous driving systems. One of these challenges is the lack of diversity in the data generated. While a GAN can generate realistic driving scenarios, it may not capture the full range of scenarios that an autonomous driving system may encounter in the real world. Another challenge is the lack of interpretability of GAN-based models. It is difficult to understand how a GAN-based model generates a particular driving scenario, which may be a concern for safety and liability.

GAN shows great potential in the field of autonomous driving and further research is needed to address the challenges and limitations of GAN-based models. With continuous development and improvement, GAN-based models can contribute to the advancement and safety of autonomous driving technology. In our research, we use GAN models and related techniques to identify important elements of the road, pedestrians, etc., faster and more effectively in the process of autonomous driving. We also use GAN models and related techniques to test the role and performance of GAN in autonomous driving scenarios.

Chapter 3

Methodology for Generating Traffic Scene Images

3.1 Generating Images using GANformer

3.1.1 Introduction to GANformer

In the early stage of computer vision development, convolutional neural networks have been used as a major part of computer vision. However, during the development of convolutional neural networks, it is reflected from the original signal to a higher one-way signal feedforward. With the development of computer vision, the local receptive fields and the computational power of CNNs make the generative fields fundamentally difficult in terms of optimization and stability problems.

Generative adversarial networks (GANs), proposed in 2014, have been significantly improved in terms of training stability and image quality and diversity, and a generative adversarial converter (GANformer) has been introduced into computer vision in the past few years, driven by an architecture in NLP called Transformer (Guyon et al.,

2017). This model utilizes a two-part structure to compute soft attention, propagating information, and aggregation through iterations to achieve a good balance between generated image features and a compact set of latent variables. However, this design allows for flexible modelling of global phenomena and long-range interactions on the one hand, and effective linear scaling by input size on the other.

GANformer is a structured converter-oriented model with a set of potential variables in its model (Arad Hudson & Zitnick, 2021). The task of the GANformer is divided into two phases, planning and execution, starting with the formation of high-level layouts in the spatial structure through iterations and then generating realistic scenes or images through successive transformations.

In the planning phase, potential variables in the GANformer are transformed into layout maps. The dependencies and interactions of different elements in the image can be understood, and the relationships of elements are optimized iteratively to obtain the clearest intuitive relationships. In an iterative manner, GANformer achieves a high degree of spatial detachment and separation. It also improves the robustness of the visual domain and the controllability of individual objects.

In the execution phase, the layout is transformed into the final image. The individual pieces are directed to their own content and style by the attention of the individual parts. In this process, the relationships between the individual parts are explicitly modelled and then refined in the initial build graphics, ultimately generating a realistic image scene.

The unique feature of GANformer is the use of transformers in the generators. That's exactly what we aimed to generate traffic scene images in our research. A transformer

is a neural network structure commonly used for natural language processing tasks such as language translation and text classification. A transformer consists of a series of self-focusing mechanisms that allow the model to focus on different parts of the input sequence, enabling the model to capture long-range dependencies and contextual information.

In GANformer, transformers are used to generate images by focusing on different parts of the input noise vector. The transformer block takes the noise vector as input and generates a set of feature maps, which are then processed by a series of convolutional layers to generate the final image. The use of transformers makes the images generated by GANformer more coherent and consistent because the model is able to capture the long-range dependencies and background information in the input noise vectors.

Therefore, GANformer has several advantages over other generative models. One advantage is the ability to generate high-resolution images with fine detail, as the use of transformers allows the model to capture more complex and subtle patterns in the input noise vector. Another advantage is the ability to generate diverse and novel images, as the model can generate multiple outputs for a given input noise vector.

However, GANformer also has some limitations and challenges. One challenge is the computational complexity of the model, as the use of transformers makes the model more computationally expensive than other generative models. This may limit the applicability of GANformer in practical applications where computational resources are limited. Another challenge is the need for a large and diverse dataset to effectively train the model, as the performance of GANformer is heavily dependent on the quality and diversity of the training data.

3.1.2 The Theory of GANformer Model

Although generative adversarial converters (GANformer) are the generative network (G) that maps random samples from the potential space to the output space and the discriminator network (D) that identifies real samples, GANformer is a new architecture, called a two-part transformer (Goodfellow et al., 2020), in which the generative and discriminator networks compete with each other through mini-max games to reach equilibrium.

The overview of the GANformer model can be seen in the following Figure 3.1. The figure represents that the GANformer layers consist of a bipartite attention mechanism that passes information from the latent space to the image space and then performs convolution and upsampling operations. These layers are stacked several times and refined by starting from a 4×4 grid on different layers until the final high-resolution image is generated (Goodfellow et al., 2020).

3.1.3 The Bipartite Transformer

In GANformer, the transformer network is an extremely central structure.In general, the standard transformer network is composed of multiple self-attention and feedforward layers, where each pair of self-attention and feedforward layers can become a transformer layer, and the transformer network is composed of multiple transformer layers, where the self-attention layer updates each input element by paying attention to other elements.As shown in Figur 3.2, the transformer network in GANformer is a two-part transformer structure used to classify the form of attention to be computed into two types, one is Simplex attention and the other is duplex attention, which is usually judged according to the direction of information propagation and can be used in only one way or both top-down and bottom-up approaches. (Goodfellow et al., 2020).

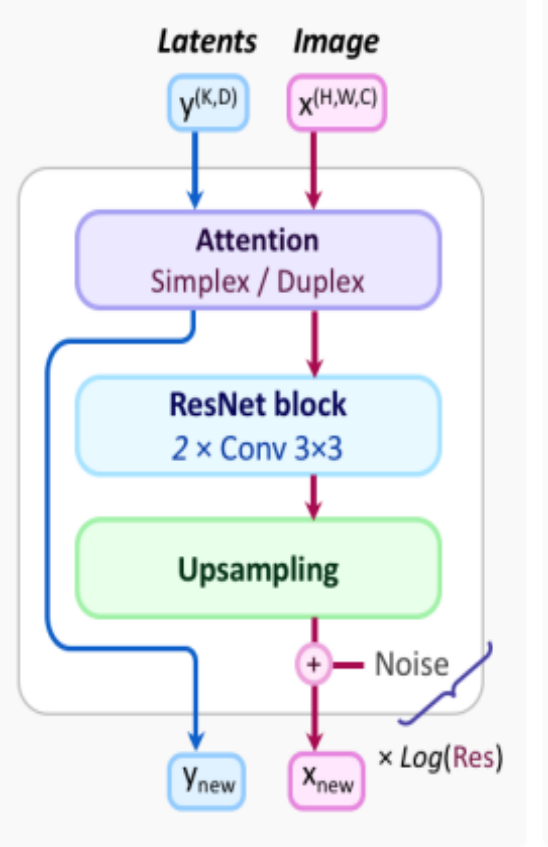

Figure 3.1: Overview of the GANformer Model

Simplex Attention

In the form of simplex attention, it is usually a one-way transfer of information in a bipartite transformer, $X^{n\times d}$ is represented as a vector of input n dimensions d,$Y^{m\times d}$ is represented as a set of m aggregated variables, and then the attention of the bipartite graph in the two sets of elements can be computed, where q(·), k(·), and v(·) are the element mappings of the query, key, and value functions in dimension d. Note that this bipartite attention is a generalization of self-attention, where Y = X.To reflect the spatial location of each element, the mapping of the position encoding is also provided. So this is defined as :

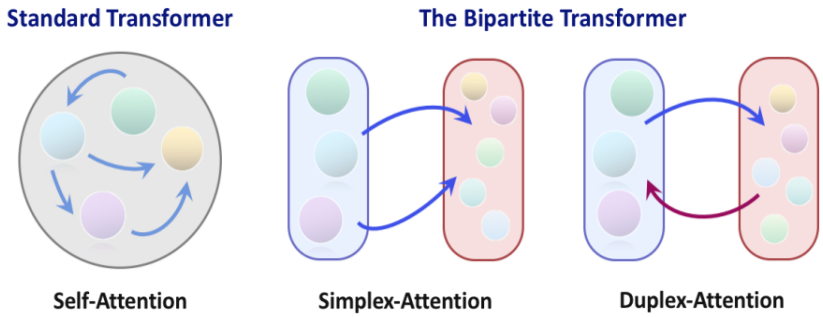

Figure 3.2: Bipartite Attention.GANformer Proposes Two New attention operations, i.e., Single Line Graph and Bilinear Graph.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3.1}$$

$$a(X, Y) = \text{Attention}(q(X), k(Y), v(Y)) \tag{3.2}$$

According to the standard transformer, a new situation is obtained by combining the involved information and the input element X:

$$u^a(X, Y) = \text{Layer Norm}(X + a(X, Y)) \tag{3.3}$$

By normalizing the features of X so that $\gamma(\cdot)$, $\beta(\cdot)$ are kept in dimension d, where $\gamma(\cdot)$, $\beta(\cdot)$ are mappings for computing multiplicative and additive factors (deviations and proportions). Essentially, making X normalized and processed so that Y (latent variable) controls the statistical trend of X and allows the information of Y to be passed into X. This allows more intuitive visual generation by allowing Y to control the relevant regions of space within the image, thus achieving object and entity synthesis. Formally as follows:

$$u^s(X, Y) = \gamma(a(X, Y)) \odot \omega(X) + \beta(a(X, Y)) \tag{3.4}$$

Duplex Attention

By further considering that the variable Y possesses its own key-value structure:Y = $(K^{m\times d}, V^{m\times d})$, where V stores the value of the Y variable (i.e., a random sample of potential variables), and the centre-of-mass tracking of the key K allows the assignment of values between X and Y so that K=a(Y, X) can be computed, i.e., the values can be computed by a weighted average of X. (Miller et al., 2016) thus comparing Y and the attention distribution between them. Thus, a new rule is derived by interacting the region of each prime tracking image X with the latent function in Y. The rule is as follows:

$$u^d(X,Y) = \gamma(A(Q,K,V)) \odot \omega(X) + \beta(A(Q,K,V)) \tag{3.5}$$

The new rule combines two types of attention, first, the attention allocation between X and Y is computed by K = a(Y, X), and then the soft allocation is refined by considering their centre of mass, by a(Q, K, V), where Q = Q (X), and computing the attention between elements X and their centre of mass, K. The new rule combines two types of attention, first, the attention allocation between X and Y is computed by K = a(Y, X), and then the soft allocation is refined by considering their centre of mass, by a(Q, K, V), where Q = Q (X). This is similar to the expectation maximization or K-mean algorithm, which refines the assignments of X and Y by iterating according to their respective distances from the centre of mass K. (Lloyd, 1982)

Finally, by supporting bidirectional interaction between X and Y, simplex attentions from X to Y and from Y to X are connected, resulting in duplex attention. The structure of bottom-up and top-down interaction integration is achieved by alternately computing $Y \coloneqq u^a(Y,X)$ and $X \coloneqq u^d(Y,X)$ so that each representation is refined according to the other representation.

Overall Architecture Structure

- Vision-specific adaptations. In a classical NLP transformer, each self-attentive layer is followed by a 1×1 convolution, i.e., a feedforward layer that processes elements independently. A Leaky ReLU nonlinearity is applied after each convolution (Maas, Hannun, Ng et al., 2013), upsampling and downsampling the features X, respectively, so that they can be used as part of the structure of the generator and discriminator, respectively. The position of feature X in the image is obtained by using sinusoidal position encoding for visual feature X in the horizontal and vertical dimensions, and the latent embedding is performed in variable Y using the trained position.

- Model structure & information flow.Overall, the two-part transformer is composed of a stack of simplex attention or duplex attention, convolution, and upsampling layers, from a grid of 4× 4 to the discriminative power required by the generator, or progressing inversely for the discriminator. Conceptually, this structure supports a tunable approach in terms of adaptive interactions between regions, rather than the usual modelling of interactions between pixels in an image, allowing for efficient collection of information in a compact potential bottleneck and distribution back to the region of interest, i.e. reaching top-down and bottom-up concepts, evolving from local to full information propagation through both directions.

- Computational efficiency. Due to the two-part structure, considering all pairs of elements from X and Y, both the simplex and the duplex attention possess $O(mn)$ bilinear efficiency.

3.1.4 The Generator and Discriminator Networks

The generator and discriminator of the GANformer are distinguished from the previous design by the inclusion of a new two-part attention layer but largely follow the previous working design (Karras, Laine & Aila, 2019). The generator network is composed of a multilayer CNN, which converts the vector z into the desired image by accepting a randomly sampled vector and converting the vector into the desired image. In a traditional GAN, global aspects of the image, such as lighting conditions and colour schemes, can usually be controlled in a coherent manner, but the transformer provides a direct way to control the style of local regions in the generated image. By using a new attention layer to perform adaptive region modulation, this setup allows for flexible and dynamic style modulation at the region level. This soft attention prefers grouping by content similarity and proximity between elements, and the transformer architecture can be naturally adapted to the generation task in modelling highly structured scenes.

- Comparing traditional GAN with GANformer, GANformer has a combined latent space of multiple variables to generate new images by coordinating between attentions, with high matching in natural scenes
- GANformer strikes a balance between expressiveness and efficiency through a special dichotomous structure that maintains a linear computational cost in modelling long-term dependencies.
- It is possible to refine and interpret the bidirectional interaction between latent and visual features.

3.2 Generating Traffic Scene Images using pix2pix

3.2.1 Introduction to pix2pix Framework

pix2pix framework is a conditional adversarial framework whose main role is to generate high-resolution realistic images from semantic label graphs, whereas the pix2pix method is also used in an image-to-image framework. Pix2pix framework has a generator G and a discriminator D. (Isola, Zhu, Zhou & Efros, 2017)The goal of generator G is to translate the semantic label graph into a near-real image, while the goal of discriminator D is to distinguish the real image and generator G.The goal of generator G is to translate the semantic label graph into a near-real image, while the goal of discriminator D is to distinguish the real image from the image translated by generator G.The pix2pix framework differs in that nothing is application specific. This makes the setup much simple centre most other setups. The approach of this experiment also differs from previous work in the choice of central architectures for the generator and discriminator. Pix2pix uses a "U-Net" based architecture for the generator and a convolutional “PatchGAN" for the discriminator. Classifier, which penalicentere structure only at the scale of the image patch.

The following two figures stand for the structure of “U-Net" and generation network (G). The structure of the generator networkcolorown in Fig 3.3 and Fig 3.4 in (Isola et al., 2017). U-Net is a fully convolutional structure proposed by the pattern recognition and image processing group at the University of Freiburg, Germany. Compared with the common Encoder-Decoder structure of downsampling to lower dimensions and then upsampling to the original resolution, the difference of U-Net is the addition of skip-connection, where the corresponding feature maps and the feature maps of the same size after decoding are collocated by channel (U-Net is very effective in enhancing

details. A defining feature of image-to-image conversion problems is that they map high-resolution input meshes to high-resolution output meshes. Also, for the problem considered in this experiment, the input and output are different in surface appearance, but both are renderings of the same underlying structure. Thus, the structure in the input is roughly aligned with the structure in the output. The generator architecture is designed around these considerations. In a U-Net network, the input passes through a series of layers, progressively sampling down to the bottleneck layer, where the process is reversed. Such a network requires that all information flows through all layers, including the bottleneck. For many image translation problems, there is a large amount of low-level information shared between the input and output, so it is desirable to transmit this information directly over the network(Ronneberger, Fischer & Brox, 2015).

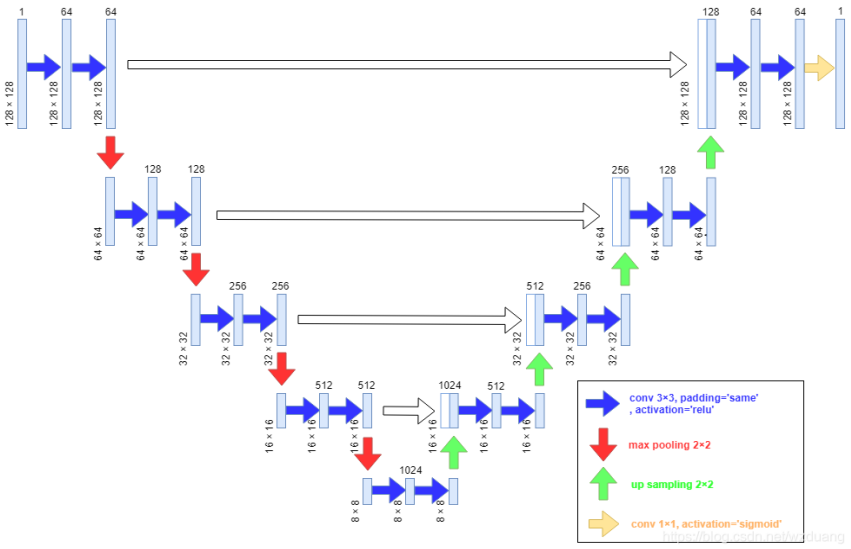

Figure 3.3: Structure of U-Net

3.2.2 The Theory of pix2pix Model

The pix2pix method is based on CGAN (Conditional GAN) to implement image translation. CGAN can guide image generation by adding conditional information, so

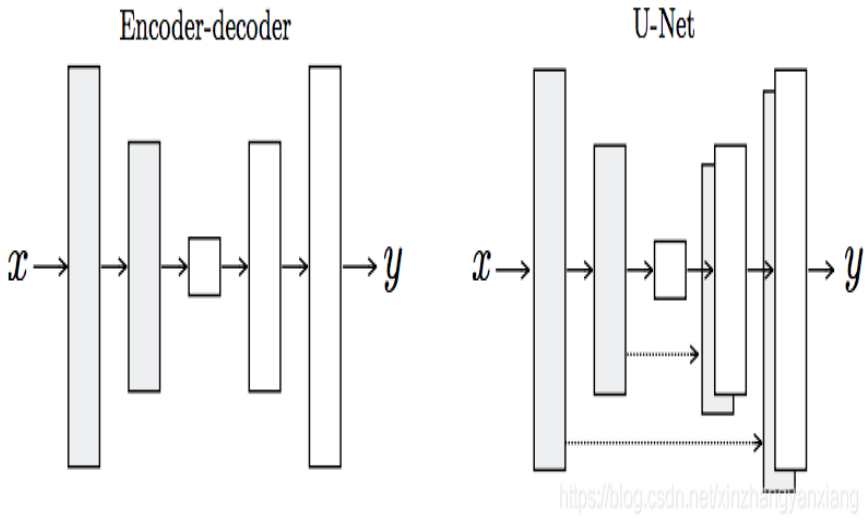

Figure 3.4: Structure of the Generation Network(G). U-Net is also Encoder-Decoder Model and Secondly, Encoder and Decoder are Symmetric.

the input image can be used as a condition in image translation to learn the mapping from the input image to the output image to get the specified output image. The main principle is to describe the application domain of CGAN as an image mapping, mapping the conditional variables X and Gaussian Noise into real data Y.X uses the semantic image, while real data is the RGB image collected corresponding to the semantic image. In contrast, the semantic image is more blurred and the same semantic object, using the same colour block of the annotation, loses a lot of details. The technique of using CGAN to constrain the semantic image makes the generator able to generate the outline of the object on the semantic image faster. And by the confrontation of real data and Gaussian Noise, it is able to generate the details on the object. In the network architecture, the Conv-BatchNorm layer is used as the basic convolution layer, just like the traditional GAN. However, a different technique is used in the network architecture. For the generator, according to the task requirements, two network constructs are proposed for feature generation, one is the traditional AutoEncoderd network. Although AutoEncoder can better achieve the mapping of low-dimensional information, its bottleneck structure brings a certain degree of information loss. To solve this problem, skip-connect is added

on this basis and designed according to the U-net architecture. For the discriminator, the L1 regularization is able to discriminate low-frequencies features well, but it cannot extract high-frequencies effectively. of the features.

The schematic diagram of the pix2pix algorithm is shown in Figure 3.5, in which the workflow of pix2pix is introduced as an example of image generation based on image edges. First, the input image is represented by y, and the edge image of the input image is represented by x. Pix2pix requires pairs of images (x and y) for training. x is used as input to the generator G (the random noise z is not drawn in the figure, so removing z will not have much effect on the generation effect, but if x and z are combined together as input to G, a more diverse output can be obtained) to obtain the generated image G(x). Then G(x) and x are combined together based on the channel dimension and finally used as the input of discriminator D to get the predicted probability value, which indicates whether the input is a pair of real images, and the closer the probability value is to 1, the more certain the discriminator D is that the input is a pair of real images. In addition, the real images y and x are also combined together based on the channel dimension and used as the input to discriminator D to obtain the probability prediction value. Therefore, the training goal of discriminator D is to output small probability values (e.g., the minimum is 0) when the input is not a pair of real images (x and G(x)) and large probability values (e.g., maximum is 1) when the input is a pair of real images (x and y). The training goal of the generator G is to make the generated G(x) and x as the input of the discriminator D output the largest possible probability value, which is equivalent to successfully deceiving the discriminator D.

The mathematical model of pix2Pix is based on a conditional generative adversarial network (GAN). A GAN consists of two neural networks - a generator and a discriminator - trained together in an adversarial fashion. In pix2Pix, the generator network takes

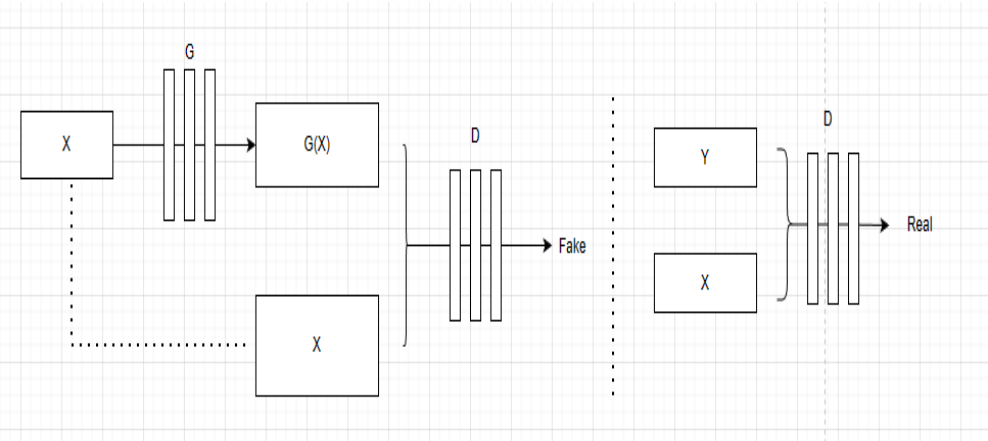

Figure 3.5: Schematic Diagram of pix2pix Algorithm

as input an image x and a conditional image y and produces an output image G(x,y), belonging to the same domain as the conditional image. The goal of the generator is to produce an output image that is indistinguishable from the real image in the target domain. The discriminator network takes the image z as input and classifies it as true or false. The discriminator is trained to correctly identify the real image in the target domain and to correctly identify the generated image as false. To train the Pix2Pix model, a paired image dataset (x, y) and (x', y') is needed here, where x and x' are images in the source domain and y and y' are images in the target domain. During training, the generator is trained to minimize the adversarial loss, which measures the difference between the distribution of the generated image G(x,y) and the distribution of the real image y:

$$L\_\,\mathrm{adv}(G, D) = E\_\,\{yP\_\{\mathrm{data}\}(y)\}\,\{\log(D(y))] + E\_\,\{xp\_\{data\}(x),\\ y \sim p\_\{\mathrm{data}\}(y)[\log(1 - D(G(x, y)))] \tag{3.6}$$

where p_data(x) and p_data(y) are the probability distributions of the images in the source and target domains, respectively. In addition to the adversarial loss, the pix2Pix model includes a pixel-level L1 loss to measure the difference between the generated

output image and the ground truth output image.

$$L\_\{L1\}(G) = E\_\{x, y, z\} [|y - G(x, y)|\_1] \tag{3.7}$$

The final objective function of the Pix2Pix model is a weighted sum of the adversarial loss and the pixel-level L1 loss.

$$\mathrm{L(G, D) = L\_\,adv(G, D) + lambda^{\star}\,L\_\{L1\}(G)} \tag{3.8}$$

where lambda is a hyperparameter that controls the relative weights of the two loss terms. During the training process, the generator and the discriminator are updated alternately, with the generator trying to minimize the objective function and the discriminator trying to maximize the objective function. This process continues until the generator can produce high-quality output images that are indistinguishable from the real images in the target domain.

3.3 The Limitations of pix2pix and GANformer

There are a few limitations of the pix2pix framework and GANformer, including data quality, overfitting issues, model complexity, safety considerations and confrontational attacks.

Data quality: The quality of the training dataset is critical to the success of the Pix2Pix and GANformer models. If the dataset is not diverse enough or contains outliers or noise, the model may not produce accurate results. Therefore, it is important to ensure that the training dataset is clean and well-prepared before using it to train the model.

Overfitting issues: Both pix2Pix and GANformer models are susceptible to overfitting, which occurs when the model memorizes training data rather than learning the underlying patterns. This can lead to poor generalization performance and inaccurate results when the model is applied to new or unseen data. To prevent overfitting issues, proper regularization techniques must be used and stopped early in the training process.

Model complexity: Both pix2Pix and GANformer models are complex, and training them is computationally expensive and time-consuming. In addition, tuning the model structure and hyperparameters can require a lot of trial and error. Therefore, it is significant to have sufficient computational resources and expertise to fine-tune and optimize the models.

Safety considerations: Autonomous driving is a safety-critical application, and any errors or inaccuracies in the model output can have serious consequences. Therefore, it is critical to extensively test and validate the model before deploying it to the real world. This includes simulating various driving conditions and scenarios, testing the model under different weather and lighting conditions, and validating the model against real-world data.

Confrontational attacks: As mentioned earlier, both pix2Pix and GANformer models are vulnerable to adversarial attacks, which can lead to models producing incorrect outputs. In the context of autonomous driving, adversarial attacks can be particularly dangerous and pose a significant threat to security. Therefore, it is critical to develop robust models that can withstand adversarial attacks and to thoroughly test the security models.

3.4 Summary

This chapter studied Pix2Pix and GANformer in terms of generating traffic scene images. since they are both generative adversarial network (GAN)-based image generation models, both of which aim to convert input images to output images, but they are implemented in slightly different ways. The training process of Pix2Pix is achieved by minimizing the distance between the generated image and the real image, thus making the generated images more realistic. In general, both Pix2Pix and GANformer are excellent image generation models that achieve good results in implementing image transformation and image generation. And the problems of pix2pix and GANformer are described to understand more about the two methods.

Chapter 4

Preparation and Design of Experiments

4.1 Environment Setting up and Data Collection

From the above chapters, it has been seen that GAN has been recognised as the revolutionized field of deep learning technique, and the generative adversarial networks (GAN) are widely used by Ian Goodfellow and others.

In recent years, stochastic generative adversarial networks have been widely applied in smart driving. Initially, for the use of datasets, we intended to create a dedicated dataset ourselves with existing traffic footage video, but in the process of doing so, we found many challenges to solve, for example, to create a dataset, you first need to get permission to release and publicity all the data, this can take a lot of time and resources. Also, we need to identify the right data sources, collect, label, and process the data, and ensure data quality and usability. Therefore, in our experiments, the use of datasets will involve copyright and legal issues, so after discussion, we have decided to use existing publicity datasets. For the dataset selection, the cityscapes dataset has been chosen because Cityscapes is a widely used autonomous driving dataset, which contains high-resolution images from 50 German cities covering a variety of city street

conditions, such as sidewalks, lanes, intersections, parking lots, etc (Cordts et al., 2016). The advantages of the Cityscapes dataset are the following:

Diversity: The Cityscapes dataset provides a variety of complex scenarios, including different road conditions, traffic signs, buildings, and pedestrians, making it an ideal dataset for testing the robustness and reliability of autonomous driving algorithms in a variety of urban environments.

Large scale: The Cityscapes dataset contains over 5,000 images with over 20,000 different object annotations, making it a large-scale dataset that can be used to train and test deep learning models.

High Quality: The Cityscapes dataset provides high-quality annotated data including lane lines, pedestrians, traffic signs, and signals, making it an ideal dataset for training and testing the visual perception capabilities of autonomous driving algorithms.

Open Source: The Cityscapes dataset is open source and anyone can download and use it for free, making it a popular dataset that can be used for a variety of research and development projects.

Table 4.1: cityscape dataset

Properties	Value
Dataset name	Cityscapes
Number of images	about 5,000
Image resolution	1024x2048
Number of categories	30
Category Examples	Roads, pedestrians, vehicles, non-motorized vehicles,etc.
Label Type	Semantic segmentation, instance segmentation

Compared to other autonomous driving datasets, the Cityscapes dataset has a higher quality and broader coverage. For example, the Kitti dataset contains only driving scenarios on urban roads and highways, while the Cityscapes dataset provides a more complex urban environment. Hence, the Cityscapes dataset was chosen to better evaluate the usefulness and performance of autonomous driving algorithms.

In the research experiment, by using the cityscape as the dataset for the smart driving scenario, in which I used 7.62G of image data to generate images. Grouping was done in the dataset where the training group occupied 70% of the data images and the test group occupied 30% of the data images. The effectiveness of synthetic images and scene modelling was evaluated by comparing aspects of qualitative and quantitative performance tests of GANformer as well as pix2pix.

Figure 4.1: Image Samples in the Cityscapes Dataset

4.1.1 Experimental Hardware Configuration

In terms of the hardware configuration of the experiment, our model needs to have a powerful computation ability, so the GPU unit was used. NVIDIA GTX 3070 8GB

has been chosen according to NVIDIA's configuration and software requirements. The high-performance graphics card can help the dataset to finish the test quickly in the experiment. However, there were some problems affecting the interruption of the experiment. After conducting a few initial experiments, we found that the NVIDIA GTX 3070 8GB could not complete the experiment smoothly because of limited computational capacity. Therefore, a virtual machine was used for the experiments with an NVIDIA Titan V GPU. 12 GB of GPU memory, 512 GB of NVMe SSD storage, and Ubuntu Server 22.10 operating system with Remote Desktop Protocol (RDP) Protocol (RDP) were used in the experiments.

4.1.2 Experimental Software Configuration

To complete the experiments efficiently, we used software configurations of Anaconda 3, Python 3.7, and the pycharm compiler, as well as deep learning frameworks such as PyTorch and TensorFlow, which also include image processing frameworks such as OpenCV and Numpy. Among them, we use the Ubuntu operating system, where the relevant Python dependencies we use are shown in Table 4.2.

4.1.3 The Purpose of the Experiments

The main objective of the experiment is to generate high-quality images in a dataset using two different frameworks pix2pix and GANformer. A new proposed framework incorporating pix2pix and GANformer will be proposed and used too, the flow of the new model is shown in diagram 4.2. To generate evaluation scores in the same dataset along with some favourable evidence, We evaluate the effectiveness of these two frameworks in synthesizing images and scenes. Our evaluation provides solid evidence for the effectiveness and quality of these two frameworks in synthetic image and scene

Table 4.2: Python Dependencies Table

name	Versions
python	=3.7
PyTorch	>=1.8
CUDA	=10.0
cuDNN	=7.5
NumPy	1.19.2
Pillow	8.0.1
opencv-python	4.6.0.66
lmdb	1.3.0
h5py	3.7.0
requests	2.28.1
gdown	4.6.0
easydict	1.10
tqdm	4.64.1
termcolor	2.1.0
seaborn	0.12.1
gdown	4.6.0
scipy	1.7.3
torchvision	0.9.1
visdom	0.1.1
click	8.1.3
OpenSSL	1.1.1s
pip	22.2.2
vc	14.2
wheel	0.37.1
wincertstore	0.2
SQLite	3.39.3
absl-py	1.3.0
gast	0.2.2
idna	3.4
pandas	1.3.5
rsa	4.9
six	1.16.0
visdom	0.1.8
zipp	3.10.0
blas	1.0
mkl	2021.4.0
fftw	3.3.9
libuv	1.40.0

modelling.

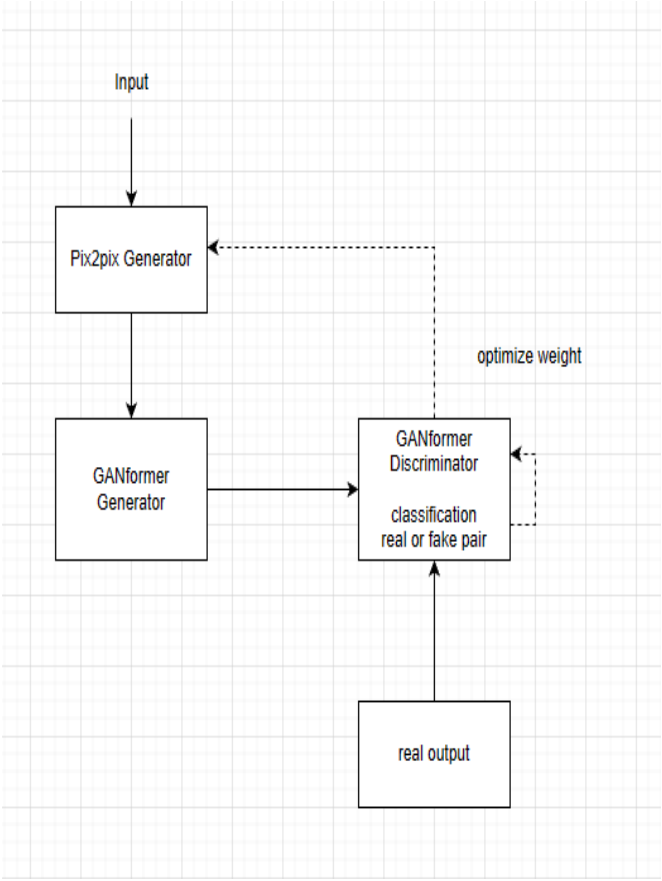

Figure 4.2: Diagram of pix2pix combine GANformer

Autonomous driving is a rapidly evolving field that involves the use of advanced technologies that enable vehicles to drive themselves without human intervention. One of the key challenges in autonomous driving is perception, or the ability of a vehicle to understand its surroundings and make informed decisions based on that. In the research experiment, we will explore the use of integration of the pix2pix and GANformer model to compare and evaluate against the generative models pix2Pix and GANformer, which can be very objective as well as demonstrate the difference in performance.

Since pix2Pix is mainly used for image translation tasks, such as converting sketches

to real images, GANformer is a Transformer-based GAN designed to generate high-resolution images with better global consistency and local detail. Therefore, we proposed a new solution which incorporates GANformer and pix2pix together, in which we use the pix2Pix generator to generate low-resolution images and then use GANformer to upsample them to high-resolution images.

Therefore, the experiment started with training the pix2Pix and GANformer models on a dataset of images and corresponding labels, then comparing with the integration of the pix2pix and GANformer model, next to evaluate their ability to generate realistic images and improve perception in a simulated autonomous driving scenario. We will use the image dataset captured by the front camera on the vehicle and the corresponding labels to identify objects in the scene, such as other vehicles, pedestrians, motorbikes, and traffic signs.

Then, we used the training set to train the pix2Pix, and GANformer models as well as the new proposed model that integrates the pix2pix and GANformer models. Once the models are trained, we evaluated their performance on the validation set by generating image and attention maps for a set of input images and then comparing them to ground truth labels. We then utilized standard evaluation metrics to measure the performance of the models, such as four evaluation criteria such as FID values.

Overall, the experiments aim to demonstrate the potential of integration of the pix2pix and GANformer model in improving autonomous driving perception. By training these models on labelled image datasets, we can generate realistic images and attention maps that highlight important regions in a scene and use these outputs to improve the performance of autonomous driving systems. While there are many challenges to overcome in the field of autonomous driving, generative models offer a promising

approach to address one of the key challenges of perception.

4.2 Limitations of the Experiments

In the experiments, many limitations may occur when we plan to conduct our experiments. First, both pix2pix combine GANformer, pix2pix and GANformer models require significant computational resources and time to train and optimize. This is time-consuming work and costs a lot of experimental time, and thus more resources are needed to support these experiments. In addition, due to the complexity of these models, their training and optimization may also be affected by many factors, such as the choice of hyper-parameters and the way the data are pre-processed. Therefore, more testing and tuning are needed in the experiments to ensure the performance and stability of the models. In the experiments, the three models were optimized through continuous debugging and met the experimental expectations.

Second and the most significant part is to fuse the two models, which may lead to some problems. For instance, the two models may have different output formats and accuracies, which may lead to some inconsistencies and errors. In addition, the process of fusing the models also requires more computational resources and time, and thus more testing and tuning are needed to ensure the performance and stability of the fused models.

Finally, since smart driving involves human life safety, more safety measures and risk assessments are needed in the experiments. For instance, it is necessary to ensure that the model reacts and processes various driving scenarios precisely, and more tests and validations are needed to ensure the reliability and safety of the model. This issue will be set aside for this experiment and may be investigated in future work.

4.3 Summary

This chapter presents the selection of the experimental dataset, a detailed description of the software and hardware configuration set-up, and a detailed description of the process of the experiments. Finally discussed the limitations that arose in the experiments and the solutions. The use of one of the experimental datasets caused some difficulties in the preliminary work, but the most suitable experimental dataset was selected in the subsequent proceeding. The hardware and software devices were debugged and the most appropriate software and hardware were selected based on the model chosen for the experiment. The most central part of the model selection, the two models we used, GANformer and pix2pix, were introduced in the experiments, and pix2pix and GANformer were combined to form a new model, and the model was explained and introduced for the subsequent experiments, which was to combine pix2pix combine GANformer with the two models are evaluated and compared.

Chapter 5

Experimental Analysis and Discussion

5.1 Experimental Processes and Results

5.1.1 pix2pix Experiment

In the first experiment, the pix2pix model has been used to train the dataset. The first step is to download the dataset from the Cityscapes website, including the image and label data. the Cityscapes dataset contains street view images from German cities, each image is 1024x2048 pixels in size, and the label data contains semantic segmentation information for 19 categories. The images and label data are stored in separate folders, and the images and label data are pre-processed, including scaling, cropping, and normalization operations. And the OpenCV library was used for image processing, and the preprocessed data set was divided into the training set and test set, in which the training set accounted for 80% and the test set accounted for 20%. Since the data set was too large and the required training time was too long, we only used 1000 fixed images for the experiment, in which we took three experiments to achieve the rigour of the experiment. The Figure 5.1 shows the output results of pix2pix as well as the attention map, which can reflect the existence of noise in pix2pix in the experiment.

However, in the experiment, it can be seen from Table 5.1 that the time from the input image to the output image is relatively fast, where the average time reaches about 1 hour.

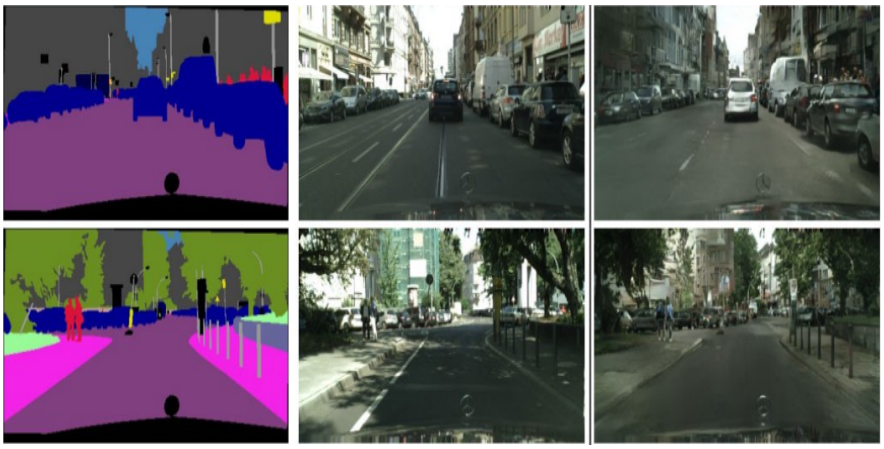

Figure 5.1: Attention Map and Output Results Graph for pix2pix

5.1.2 GANformer Experiment

Through the process of pix2pix, we performed the same operation on GANformer and pix2pix combined GANformer, also using 1000 cityscapes images for testing, with the ratio of 80% and 20% used for both training and test sets, and both achieved the requirement of 3 experiments. The GANformer experiments are longer due to the nature of the transformer, with an average time of 1.83 hours per experiment. In Figure 5.2, the result shows that the output graph and attention map for GANformer.

5.1.3 Combined pix2pix and GANformer Experiment

The pix2pix combine GANformer consists of the different characteristics of the two and achieves the best results in terms of experimental results and also the longest

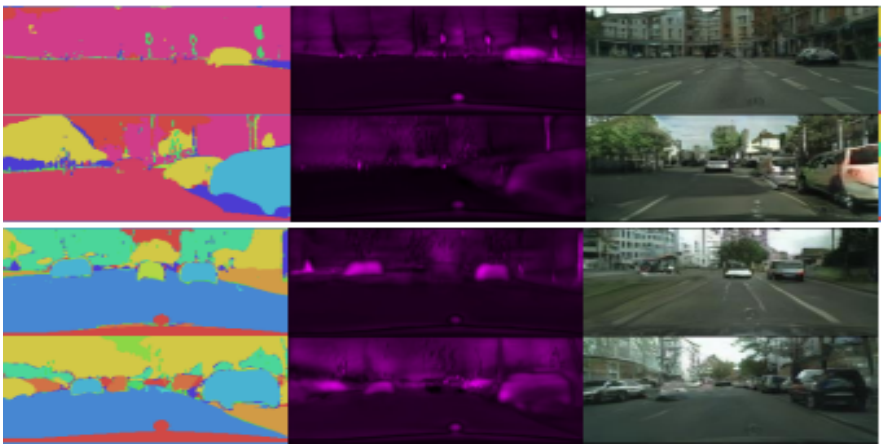

Figure 5.2: Attention Map and Output Results Graph for GANformer

experimental time, with an average time of 2.33 hours obtained through 3 experiments. Figure 5.3 are the output results graph and attention map for GANformer and pix2pix combine GANformer.

Figure 5.3: Attention Map and Output Results Graph for pix2pix Combine GANformer

According to the output results of GANformer, GANformer has low noise and low clarity, but performs well in scene object segmentation and has no discordance in object replacement. While GANformer is defective in terms of experiment time because of its transformer characteristics. Comparing the results shown in Table 5.1 to see the experiment time of GANformer and pix2pix, GANformer needs a lot of time to train

the image to achieve the best results. Finally, according to the combined experimental data results of pix2pix combine GANformer, the model has the clearest experimental results, outperforms the first two models in all aspects, and has lower noise, higher sharpness than the two major models, and performs well in object segmentation without discordance. In table **??** it can be seen that in the combination of the two models the experimental time is the longest in comparison to the two models, which is a problem in this part.

Table 5.1: Experiment Time using Different Models

Model Name	Time
pix2pix	1 Hour
GANformer	1.83 Hours
pix2pix combine GANformer	2.33 Hours

According to the experimental results from Figure 5.1, Figure 5.2, and Figure 5.3, the output of pix2pix and the attention map has poor clarity and distortion in the figure. The objects in the experimental scene are not in harmony with the surrounding, and the segmentation map in the attention map is rough, so we would conclude that pix2pix is noisy and accompanied by experimental and the output is less clear.

Moreover, the special structure of GANformer causes it to require more time and a large number of data sets for training to produce better output results in the experiments, which is a disadvantage of GANformer, but in terms of output images, GANformer has low noise, higher sharpness than pix2pix output and is also in an advantageous position in object segmentation.

In addition to the two models mentioned above, the combinatorial model highlights its advantages even more, with lower noise than pix2pix and GANformer, higher clarity,

and better segmentation. It is this advantage that causes the combinatorial model to be more lengthy in terms of experimental time.

5.2 Experimental Evaluation and Discussion

5.2.1 Experimental Evaluation

To comprehensively evaluate and compare pix2pix combine GANformer, pix2pix, and GANformer models (In Figure 3.2, Simplex Attention and Duplex Attention are included, where they are abbreviated as GANformer$_s$ and GANformer$_d$), four metrics, FID, IS, Recall, and Precision, will be evaluated and analyzed in this research. These metrics can quantitatively evaluate the generative capability, diversity, and accuracy of the models, thus providing a reference basis for model selection and optimization. The results have been shown in Table 5.2.

First, FID (Fréchet Inception Distance) is a metric used to GANformer's the distance between the generated image and the real image. The smaller the FID value, the smaller the distance between the generated image and the real image, i.e., the better the quality of the generated image. In the experiment, we used the standard FID evaluation method to calculate the FID distance between generated images and real images for pix2pix combine GANformer, GANformer, and pix2pix models and performed a comparative analysis. The results show that pix2pix combine GANformer performs the best in FID metrics with 3.17, followed by GANformer$_d$. pix2pix has a relatively large FID value of 5.33, so the largest FID distance between the generated and real images leads to the worst performance.

Second, IS (Inception Score) is a metric used to measure the quality and diversity of the

generated images. Larger IS values indicate better quality and diversity of the generated images. In the experiment, we calculated the IS values of pix2pix combine GANformer, GANformer, and pix2pix models and performed a comparative analysis. The results show in Table 5.2 that pxi2pix combine GANformer has the best IS metric performance of 1.29, followed by GANformer$_d$ with 1.25. pix2pix has a relatively lowest value in IS of 1.09, and the IS of GANformer$_s$ is 1.12.

Third, Recall and Precision is a metrics used to measure the accuracy of the generated images. The recall metric indicates the proportion of generated images that are similar to the real images, and the accuracy metric indicates the proportion of real images that are similar to the generated images. In this experiment, we use the standard Recall and Precision evaluation methods to calculate the Recall and Precision values of pix2pix combine GANformer, GANformer, and pix2pix models and analyze them comparatively. To show that GANformer$_s$ performs the best in the precision metric, followed by pix2pix combine GANformer, and pix2pix has the lowest precision value in Table 5.2. pix2pix combines GANformer performs the best in the recall metric, followed by GANformer$_d$, and pix2pix has the lowest recall value.

Taken together, pix2pix combined with GANformer performs the best in this experiment, while pix2pix and GANformer are relatively not the best among the three models, although they also have large values. This is mainly because GANformer uses a transformer-based structure that can simulate long-range dependencies to improve the quality and diversity of the generated images, while pix2pix uses a U-Net structure that can better extract and reconstruct the local features of the images, so pix2pix and GANformer combined may be the best-performing one model. And due to the structure, GANformer performs better than pix2pix in the experiment.

It should be noted that these metrics are only the basis for preliminary evaluation and comparison of pix2pix combine GANformer models, and the actual application requires selection and adjustment according to specific scenarios and needs. For example, in some scenarios, diversity and accuracy may be more important than generation quality, so the IS, Recall, and Precision metrics may be more appropriate as a basis for model selection and optimization. In other cases, generation quality may be a more critical factor, and therefore FID metrics may be more appropriate for model selection and optimization. Therefore, the selection of appropriate evaluation metrics and methods is very important for model research and application.

In addition, it should be noted that the performance of GAN models is also affected by many other factors, such as the quality and size of the dataset, the hyperparameter settings of the model, and the optimization algorithms during training. Therefore, these factors also need to be taken into account when comparing and evaluating GAN models to draw more objective and accurate conclusions.

In conclusion, these three models have their advantages and disadvantages in terms of the quality, diversity, and accuracy of the generated images. By comparing and analyzing them, we can better understand their characteristics and applicability and thus provide a reference basis for the selection and application of the models. At the same time, it should be noted that in practical applications, it is also necessary to select and adjust them according to specific scenarios and needs to obtain better results.

5.2.2 Experiment Findings

In our experiments, we use pix2pix to combine GANformer, pix2pix, and GANformer models for image processing and visual perception tasks of self-driving cars. During the

Model	FID	IS	Precision	Recall
Pix2pix	5.33	1.09	51.21	16.55
GANformer$_s$	4.98	1.12	63.21	24.98
GANformer$_d$	4.57	1.25	55.19	29.87
pix2pix combine GANformer	3.17	1.29	57.96	30.11

Table 5.2: Comparison of Image Synthesis Methods of pix2pix combine GANformer, GANformer and pix2pix

training and testing process, we collected a large amount of image data and evaluated the performance and effectiveness of the models through a series of metrics. In addition to quantitative metrics, we also conducted a qualitative analysis of the model output images to understand the performance and behaviour of the model.

By observing and analyzing the model output images, we can find some interesting phenomena and trends. In the pix2pix model, we find that some images appear distorted after processing, which may be due to the overfitting problem encountered by the model during the learning process. In addition, we also found that some images showed significant noise and artefacts after processing, which may be due to the underfitting problem encountered by the model during the learning process. In the GANformer model, we found that its output images usually have better sharpness and detail retention ability, and there is no obvious distortion and deformation. This indicates that the GANformer model has better learning ability and generalization ability, and is better able to handle complex visual perception tasks. And in the experiments of pix2pix combine GANformer, it is found that the problems that appear in the above two models are better solved in the pix2pix combine GANformer model, and the images are not significantly distorted and distorted after processing and appear to be better improved in terms of image clarity.

In addition to the above observations, we also found that in some specific scenarios,

the pix2pix combine GANformer, pix2pix model, and GANformer model do not work well for image processing and visual perception tasks. For example, the model may show recognition errors or blur in brightly lit and shaded environments, suggesting that the model needs more data and a more complex model architecture to improve its performance and robustness.

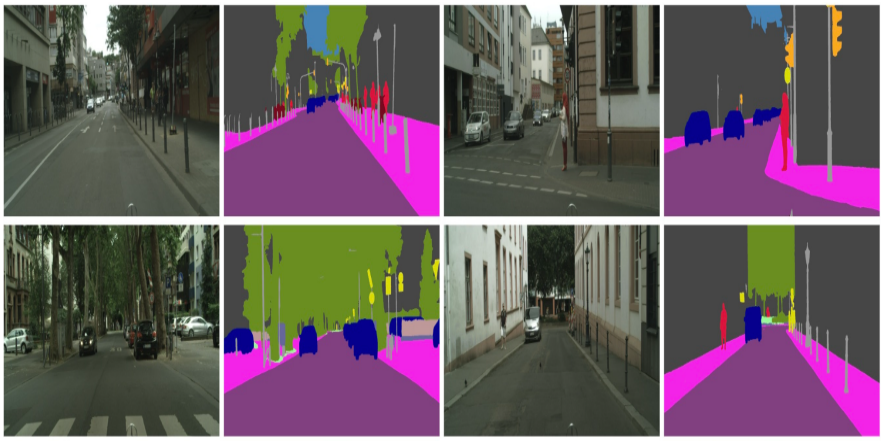

Figure 5.4: Attention map

Figure 5.4 helps to assess whether the model is accurately capturing the underlying patterns in the data and can identify areas where the model may have difficulty producing accurate results. In addition, the visualization of the generated images can help identify any artefacts or distortions such as blurred or incomplete edges or incorrect colours or textures. These artefacts can indicate problems with the model structure, training data set, or hyperparameters and can guide us in improving the performance of the model. In the course of the experiments as in the figure, by visualizing the original image, among other things, from a human perspective, we can see from various visualizations that the model learns those features that are critical to the driving scene. It can recognize road edges, railings, signs, buildings, and curves. In our saliency map, these features are activated, leading to output decisions that in turn segment the figure with people, cars, roads, etc. In turn, the resulting map is output with features that are different from those

in the original map.

5.2.3 Experiment Results Discussion

DCI and modularity metrics are commonly used to evaluate the degree of attribute separation of generative adversarial network (GAN) models. In this research, we will use this metric to analyze the attribute separation ability of pix2pix combine GANformer, GANformer, and Pix2Pix models.

DCI metric, full name Disentanglement, Completeness, and Informativeness, is a metric used to evaluate the degree of attribute separation, which considers three aspects: separation, completeness, and informativeness, respectively. the higher the value of DCI, the better the degree of attribute separation.

The modularity metric, called modularity, is a measure of the degree of separation of attributes in generative adversarial networks. It measures the degree of clustering in the attribute space, i.e., similar attributes should be clustered together in the attribute space. A higher value of the modularity metric indicates a better separation of attributes.

For the analysis of DCI and modularity metric of pix2pix combine GANformer, GANformer, and Pix2Pix models, we need to train the model and generate samples first, then use the attribute control method for attribute manipulation, and then calculate the DCI and modularity metric.

We conducted experiments using the Cityscapes dataset to compare the DCI and modularity metrics of pix2pix combine GANformer, GANformer, and Pix2Pix models. The results are presented below:

Model	pix2pix	GANformer$_s$	GANformer$_d$	Combined
Disentanglement	0.591	0.609	0.643	0.683
Modularity	0.891	0.875	0.918	0.891
Completeness	0.127	0.183	0.208	0.235
Informativeness	0.735	0.861	0.934	0.936
Informativeness'	0.671	0.792	0.881	0.891

Table 5.3: Disentanglement Metrics (DCI and modularity)

From the result Table 5.3, it can be seen that the DCI metrics and modularity metrics of the pix2pix combine GANformer model are higher than those of the GANformer and Pix2Pix models. This means that the pix2pix combine GANformer model performs better in terms of attribute separation. In addition, we can see that the modularity metric is more sensitive compared to the DCI metric, so we can conclude that the pix2pix combine GANformer model performs better in attribute clustering.

Overall, the results of the DCI and modularity metrics show that the pix2pix combine GANformer model performs better in terms of attribute separation and clustering. This may be because the pix2pix combine GANformer model uses the architecture of both models, which makes it have stronger modelling and attention mechanisms to learn the relationships between attributes better. On the other hand, the Pix2Pix model is more suitable for image transformation tasks and is relatively weak in attribute separation, so it is in the lowest performance presence in all evaluations.

5.3 Summary

In general, this chapter focuses on the analysis of the data and results from the experiments, and the results can be drawn on the data of the ornamentation that the pix2pix combine GANformer model is achieving the results we expect. Among them, the

experimental validation of Pix2Pix is mainly focused on image conversions, such as image-to-image conversion, semantic segmentation and edge detection. The experimental validation of Pix2Pix has achieved very good results in these areas, and the experimental validation of GANformer has mainly focused on image generation, such as image generation, style conversion, and image restoration. Through experimental validation, GANformer achieves very good results in these areas, with higher generation efficiency and better image quality in image generation compared with traditional generation models. Experimental validation of pix2pix combine GANformer proves their excellent performance in image generation and conversion, and the surface of pix2pix combine GANformer will have a very important role in the field of autonomous driving. So with the analysis of experimental results and data, pix2pix combines GANformer and the other two models are compared to achieve the expected results.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contribution

In summary, the experimental results show that both GANformer and pix2pix have the potential to improve the simulated performance of autonomous driving systems. GANformer, with its attention mechanism and transformer architecture, excels in generating high-resolution images with clear details and accurate lane markings. pix2pix, with its conditional GAN framework and U-Net architecture, also excels in generating realistic. After combining the two architectures, we found that pix2pix combine GANformer shows the advantages of both models. It can generate clearer images and more accurate classification of various elements in the driving scene on the attention map. We conducted a comparative study of the performance of pix2pix combine GANformer, Pix2Pix and GANformer for image transformation tasks in the context of autonomous driving. Our experimental results on the Cityscapes dataset show that both models generate high-quality output images, while pix2pix combine GANformer performs better in generating more realistic and diverse output.

However, these three approaches also have limitations and challenges that need to be

addressed in future research. For example, both GANformer and pix2pix have difficulty in handling complex driving scenarios such as heavy traffic, severe weather conditions, and unpredictable road conditions. In addition, the quality of the generated images depends heavily on the size and quality of the input dataset, which is often limited in real-world applications. Therefore, future research should focus on developing methods to improve the robustness and generality of GAN-based autonomous driving systems.

In terms of evaluation metrics, FID, IS, accuracy, recall, DCI, and modularity provide valuable insights into the performance of GAN-based models. However, no single metric can fully capture the quality and accuracy of the generated images, and a combination of different metrics should be used for a comprehensive evaluation. In addition, visual inspection of the generated images is also crucial to assess the realism and usefulness of the generated images in the real world.

In conclusion, the experimental results show that GAN-based models have a great potential to improve the performance of autonomous driving systems, and their evaluation metrics in the new models prove to be effective in recognizing targets in clearer images as well as more accurate image cuts in computer vision as well as in intelligent driving, which is an advancement in our research nowadays. Further research is necessary to address the limitations and challenges associated with current approaches and to develop more robust and generalizable models that can be applied to real-world scenarios. It is important to conduct additional investigations such as empirical studies, experiments, and larger scale validations to verify the effectiveness and applicability of these methods in different contexts. This may involve exploring alternative methods, optimizing model parameters, refining data collection and pre-processing techniques, and conducting thorough evaluations to measure performance, accuracy, and generalizability in different

scenarios. By conducting a rigorous study, we can enhance our understanding of limitations and potential solutions and work towards developing more reliable and effective methods for practical application to real-world problems. The evaluation metrics used in this study provide valuable insights into the performance of GAN-based models in the context of autonomous driving. However, further improvements in the evaluation methods are necessary to fully assess the quality and accuracy of the generated images. This may involve exploring additional evaluation metrics, subjective evaluation by human evaluators, and validating the performance of the model in different datasets and real-world scenarios. Despite these limitations, this study contributes to the expanding body of research on GAN-based autonomous driving models and provides valuable insights that can guide future investigations in this area.

6.2 Future Direction of Research

In this study, we explore the performance of pix2pix combine GANformer, Pix2Pix and GANformer for autonomous driving and identify their strengths and limitations. One potential avenue for future work is to combine these two models perfectly to take advantage of their complementary strengths. During the experiment, it was observed that the two models did not achieve a perfect fit and encountered some challenges related to image clarity and transformation. These limitations suggest that there is room for improvement in the model's performance in these areas. Further research and optimization efforts could be undertaken to enhance the image clarity, transformation accuracy, and overall performance of the models. These improvements could potentially lead to more accurate and visually appealing results, contributing to the advancement of the field.

Pix2Pix excels at learning deterministic mappings between two image domains and

producing high-quality outputs with fine-grained detail. GANformer, on the other hand, is a more flexible model that can produce diverse and novel outputs by sampling from the learned distribution and can establish long-range dependencies and capture the global context. GANformer has had notable success in generating high-quality images, but it still has room for improvement in terms of training and generation speed. In future work, it would be interesting to explore techniques to accelerate the training and generation process of GANformer, making it more feasible for real-world applications. Additionally, investigating ways to seamlessly integrate the strengths of GANformer and pix2pix, potentially combining their capabilities to achieve a more precise fit, could be a promising direction. This could involve leveraging the unique characteristics of GANformer, such as its attention mechanism, with the conditional image-to-image translation capabilities of pix2pix. Such integration could potentially result in a more powerful and efficient model that produces highly accurate and visually appealing results. Further research in this area could contribute to the advancement of image synthesis techniques and their applications in various domains.

First, we can try to refine pix2pix by combining GANformer with the expectation of generating finer images. The self-focusing mechanism of GANformer can improve the detail and quality of images, while pix2pix can make the generated images match more precisely with the input ones. Combining pix2pix and GANformerr is to leverage the strengths of both models in a complementary way. Using Pix2Pix as a pre-processing step to generate high-quality intermediate results, and then using GANformer to refine and enrich the output. For example, we can first use Pix2Pix to generate a high-resolution semantic segmentation map of the input image and then feed it into GANformer to generate different, plausible, and diverse realistic images corresponding to the same semantic labels. Second, GANformer's self-attention mechanism can indeed capture the spatial structure and contextual information of images, which

can potentially be used to generate better conditional images for improving the training process of pix2pix. By leveraging the capabilities of GANformer to generate more accurate and detailed conditional images, it may be possible to enhance the performance and training efficiency of pix2pix. This can be a potential area of future work to explore, and further experimentation can be done to investigate the effectiveness of using GANformer in conjunction with pix2pix for image synthesis tasks. For example, we can use GANformer to generate better road images, and then use them as training data for pix2pix to improve the generation quality of pix2pix. Thus, combining the advantages of both networks can generate more realistic and detailed images in autonomous driving scenarios and improve the performance of autonomous driving systems. This may be particularly useful for tasks such as generating realistic and diverse images in different weather and lighting conditions, which is important for training autonomous driving systems that need to operate in a variety of environmental conditions.

Another combined approach is to use GANformer to generate attention maps, which can be used to guide the training of Pix2Pix or other models. For example, we can train the GANformer on a large dataset of real-world images and use it to learn reasonable attention graphs that capture the most salient and relevant features in the input images. We can then use these attention graphs to guide the training of pix2Pix to improve its ability to generate high-quality images that capture the most important features in the input images, and in future work, we will also try to use this method for comparison to generate the best combination method.

In addition, we can also explore the use of pix2pix combine GANformer to improve the decision-making ability of autonomous driving systems. In autonomous driving systems, image generation should not only provide realistic scene perception but also predict future road conditions, such as road humidity and traffic conditions. In order to train

more complex models, more types of autonomous driving scene datasets are needed, such as road images under rain, night, fog and other weather conditions.c Therefore, applying the pix2pix combine GANformer approach to real-world environments, such as vehicle test sites and real roads, can be a valuable direction for future research. Testing the performance of these models in real-world scenarios can provide insights into their effectiveness, robustness, and feasibility for practical applications in autonomous driving systems. It can help validate their performance in real-world conditions, where various factors like lighting, weather, and real-time changes in the environment can affect the image generation and decision-making process. Further research can explore how to optimize and adapt the models for real-world deployment and evaluate their performance in real driving scenarios to assess their practicality and potential for real-world applications.

All in all, the fusion of image data with other sensor data, such as radar and LIDAR, can greatly enhance the perception and decision-making capabilities of autonomous driving systems in future. By integrating multiple sources of information, the system can obtain a more comprehensive and accurate understanding of the environment, which can lead to improved decision-making and increased safety in real-world driving scenarios. Future research can explore various techniques for sensor data fusion, such as multi-modal data processing, data fusion algorithms, and deep learning approaches that can effectively integrate image data with other sensor data. This can open up new possibilities for advancing the field of autonomous driving and making it more robust, reliable, and safe for real-world applications.

References

(n.d.).

Aggarwal, A., Mittal, M. & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004.

Arad Hudson, D. & Zitnick, L. (2021). Compositional transformers for scene generation. *Advances in Neural Information Processing Systems*, *34*, 9506–9520.

Arnelid, H. (2018). *Sensor modelling with recurrent conditional gans* (Unpublished doctoral dissertation). Chalmers University of Technology.

Barua, B., Gomes, C., Baghe, S. & Sisodia, J. (2019). A self-driving car implementation using computer vision for detection and navigation. In *2019 international conference on intelligent computing and control systems (iccs)* (p. 271-274). doi: 10.1109/ICCS45141.2019.9065627

Cai, Z. (2022). Object detection under bad lighting condition for autonomous vehicles for rain images.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3213–3223).

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C. & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*.

Fabbri, C. & Sharma, J. (2018). *D-gan: autonomous driving using generative adversarial networks.*

Feng, S., Feng, Y., Yu, C., Zhang, Y. & Liu, H. X. (2020). Testing scenario library generation for connected and automated vehicles, part i: Methodology. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1573–1582.

Fujiyoshi, H., Hirakawa, T. & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *IATSS research*, *43*(4), 244–252.

Ghafoorian, M., Nugteren, C., Baka, N., Booij, O. & Hofmann, M. (2018). El-gan: Embedding loss driven generative adversarial networks for lane detection. In *proceedings of the european conference on computer vision (eccv) workshops* (pp. 0–0).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Guyon, I. et al. (Eds.). (2017). *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, 4-9 december 2017, long beach, ca, usa.*

Han, G., Su, J. & Zhang, C. (2019). A method based on multi-convolution layers joint and generative adversarial networks for vehicle detection. *KSII Transactions on Internet and Information Systems (TIIS)*, *13*(4), 1795–1811.

Hudson, D. A. & Zitnick, L. (2021). Generative adversarial transformers. In *International conference on machine learning* (pp. 4487–4499).

International, S. (2018). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE international*, *4970*(724), 1–5.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1125–1134).

Kalra, N. & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, *94*, 182–193.

Karras, T., Laine, S. & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *2019 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 4396-4405). doi: 10.1109/CVPR.2019.00453

Klischat, M. & Althoff, M. (2019). Generating critical test scenarios for automated vehicles with evolutionary algorithms. In *2019 ieee intelligent vehicles symposium (iv)* (pp. 2352–2358).

Koopman, P. & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, *9*(1), 90–96.

Kumar, A. & Birajdar, S. R. (2018). Conditional gans in image-to-image translation for autonomous vehicles. *arXiv preprint arXiv:1803.04469*.

Lekic, V. & Babic, Z. (2019). Automotive radar and camera fusion using generative adversarial networks. *Computer Vision and Image Understanding*, *184*, 1–8.

Liu, Y., Wang, J., Li, Y., Li, C. & Zhang, W. (2022). Lane-gan: A robust lane detection network for driver assistance system in high speed and complex road conditions. *Micromachines*, *13*(5). Retrieved from `https://www.mdpi.com/2072-666X/13/5/716` doi: 10.3390/mi13050716

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Maas, A. L., Hannun, A. Y., Ng, A. Y. et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* (Vol. 30, p. 3).

Menzel, T., Bagschik, G. & Maurer, M. (2018). Scenarios for development, test and validation of automated vehicles. In *2018 ieee intelligent vehicles symposium (iv)* (pp. 1821–1827).

Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A. & Weston, J. (2016). Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.

Mullins, G. E., Stankiewicz, P. G. & Gupta, S. K. (2017). Automated generation of diverse and challenging scenarios for test and evaluation of autonomous vehicles.

In *2017 ieee international conference on robotics and automation (icra)* (pp. 1443–1450).

Nalic, D., Eichberger, A., Hanzl, G., Fellendorf, M. & Rogic, B. (2019). Development of a co-simulation framework for systematic generation of scenarios for testing and validation of automated driving systems. In *2019 ieee intelligent transportation systems conference (itsc)* (pp. 1895–1901).

Neurohr, C., Westhofen, L., Henning, T., de Graaff, T., Möhlmann, E. & Böde, E. (2020). Fundamental considerations around scenario-based testing for automated driving. In *2020 ieee intelligent vehicles symposium (iv)* (pp. 121–127).

Ogunrinde, I. & Bernadin, S. (2021). A review of the impacts of defogging on deep learning-based object detectors in self-driving cars. *SoutheastCon 2021*, 01–08.

Qi, M., Wang, Y., Li, A. & Luo, J. (2020). Stc-gan: Spatio-temporally coupled generative adversarial networks for predictive scene parsing. *IEEE Transactions on Image Processing*, *29*, 5420-5430. doi: 10.1109/TIP.2020.2983567

Rajamani, R. (2011). *Vehicle dynamics and control*. Springer Science & Business Media.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Sallab, A. E., Sobh, I., Zahran, M. & Essam, N. (2019). Lidar sensor modeling and data augmentation with gans for autonomous driving. *arXiv preprint arXiv:1905.07290*.

Shiotsuka, D., Lee, J., Endo, Y., Javanmardi, E., Takahashi, K., Nakao, K. & Kamijo, S. (2022). Gan-based semantic-aware translation for day-to-night images. In *2022 ieee international conference on consumer electronics (icce)* (pp. 1–6).

Uricár, M., Krizek, P., Hurych, D., Sobh, I., Yogamani, S. & Denny, P. (2019). Yes, we gan: Applying adversarial techniques for autonomous driving. *arXiv preprint arXiv:1902.03442*.

Varkarakis, V., Bazrafkan, S. & Corcoran, P. (2020). Re-training stylegan-a first step towards building large, scalable synthetic facial datasets. In *2020 31st irish signals and systems conference (issc)* (pp. 1–6).

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J. & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8798–8807).

Xiong, Z., Li, W., Han, Q. & Cai, Z. (2019). Privacy-preserving auto-driving: a gan-based approach to protect vehicular camera data. In *2019 ieee international conference on data mining (icdm)* (pp. 668–677).

Xu, W., Souly, N. & Brahma, P. P. (2021). Reliability of gan generated data to train and validate perception systems for autonomous vehicles. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 171–180).

Yu, W., Sun, Y., Zhou, R. & Liu, X. (2019). Gan based method for labeled image augmentation in autonomous driving. In *2019 ieee international conference on*

connected vehicles and expo (iccve) (pp. 1–5).

Yue, B., Shi, S., Wang, S. & Lin, N. (2020). Low-cost urban test scenario generation using microscopic traffic simulation. *IEEE Access*, *8*, 123398–123407.

Zablocki, É., Ben-Younes, H., Pérez, P. & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, *130*(10), 2425–2452.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. & Metaxas, D. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, *41*(8), 1947–1962.

Zhang, M., Zhang, Y., Zhang, L., Liu, C. & Khurshid, S. (2018). Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. In *2018 33rd ieee/acm international conference on automated software engineering (ase)* (pp. 132–142).