Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.059594

ARTICLE

A Perspective-Aware Cyclist Image Generation Method for Perception Development of Autonomous Vehicles

Beike Yu[1], Dafang Wang[1,*], Xing Cui[2] and Bowen Yang[1]

[1]School of Mechatronics Engineering, Harbin Institute of Technology, Weihai, 264200, China

[2]China North Artificial Intelligence & Innovation Research Institute, Beijing, 100000, China

*Corresponding Author: Dafang Wang. Email: wangdf@hit.edu.cn

Received: 10 September 2024 Accepted: 22 November 2024 Published: 17 February 2025

ABSTRACT

Realistic urban scene generation has been extensively studied for the sake of the development of autonomous vehicles. However, the research has primarily focused on the synthesis of vehicles and pedestrians, while the generation of cyclists is rarely presented due to its complexity. This paper proposes a perspective-aware and realistic cyclist generation method via object retrieval. Images, semantic maps, and depth labels of objects are first collected from existing datasets, categorized by class and perspective, and calculated by an algorithm newly designed according to imaging principles. During scene generation, objects with the desired class and perspective are retrieved from the collection and inserted into the background, which is then sent to the modified 2D synthesis model to generate images. This pipeline introduces a perspective computing method, utilizes object retrieval to control the perspective accurately, and modifies a diffusion model to achieve high fidelity. Experiments show that our proposal gets a 2.36 Fréchet Inception Distance, which is lower than the competitive methods, indicating a superior realistic expression ability. When these images are used for augmentation in the semantic segmentation task, the performance of ResNet-50 on the target class can be improved by 4.47%. These results demonstrate that the proposed method can be used to generate cyclists in corner cases to augment model training data, further enhancing the perception capability of autonomous vehicles and improving the safety performance of autonomous driving technology.

KEYWORDS

Realistic cyclist generation; perspective-aware image synthesis; autonomous vehicle; artificial intelligence

1 Introduction

Nowadays, self-driving is a thriving cutting-edge technology in the automotive industry. However, it is challenging to perceive corner cases due to the difficulty of collecting and labeling challenging scenarios and rare objects, like cyclists. In this case, scene generation becomes especially important for training and testing self-driving cars to deal with extreme situations.

Currently, the most widely used image simulation method is Computer Graphics (CG), which models scenes using physics. Simulation engines like Carla [1], AirSim [2], and LGSVL [3] can provide a wide range of viewpoint changes and flexible constraints for objects. However, cyclists are relatively

complex in structure, making creating their 3D assets difficult. Generally, CG has some flaws, like the high price and the lengthy manual modeling process, not to mention the realism gap between CG and the real world, which results in a decline in performance for identification tasks [4].

Data-driven image editing models have recently gained attention for reducing the realism gap and enabling tasks such as object modification, weather changes, and so on. Choi et al. [5] perform image-to-image translations in different domains using a single model. Lv et al. [6] designed a fine-grained part-level semantic layout descriptor for synthesizing images conditioned on semantic labels. RESAIL [7] segments semantic labels into patches by class, retrieves corresponding patch images, and synthesizes new images holistically. Image editing models can produce high-resolution and realistic images and continuously generate data once training is completed. Nevertheless, these methods are primarily designed for the generation of vehicles, lacking the necessary realism for cyclists. Additionally, the perspective controlling mechanism remains an open issue. Since cyclists are typical traffic participants who frequently interact with self-driving cars and their movements are unpredictable, synthesizing realistic images of them for training and testing of perception models is crucial for the safety of autonomous driving technology.

Under this condition, this paper proposes a realistic and perspective-aware cyclist generation method via object retrieval. The objectives of this research are to generate realistic cyclists and to control their perspective accurately. In general, the major contributions of this paper are as follows:

1. A pipeline is built by using the cyclists and the backgrounds from existing datasets to produce new urban scenes. It reduces the realism gap between simulation and real-world data, accurately controls cyclists' perspective through object retrieval, produces images without complex modeling and rendering, and facilitates the creation of rare and challenging scenarios for training and testing the perception models in autonomous vehicles.
2. A method based on the principle of imaging for calculating perspective is proposed. It is used to label the cyclists and compute the desired angles in the newly generated images, ensuring accurate retrieval and effective control over image generation.
3. Extensive experiments are conducted to prove the ability of perspective control and realistic generation of the proposed pipeline. The applications of the building of specific scenes and data augmentation show the potential of the presented method in practice.

2 Related Work

2.1 Urban Scene Generation

Urban scene generation aims to synthesize traffic environment and participants with realistic appearance, accurate perspective, and diverse relationships so that the simulated images can be used to train and test the perception algorithms in autonomous vehicles. Simulator platforms, such as LGSVL [3] and AirSim [2], rely on virtual engines to create scenes, making progress in realistic expression. However, more vivid scenes require more financial and manual inputs.

Recently, with the rise of image synthesis methods, conditioned 2D generation has been studied widely. Depth-SIMS [8] couples a generation model and an inpainting network to transform Red, Green, and Blue (RGB) canvases and well-aligned semantic labels into RGB images and dense depth maps. Konushin et al. [9] generate traffic signs on authentic urban scene images for data augmentation. Li et al. [10] leverage local and global consistency to automatically produce rare objects in urban scenes. NeuralField-LDM [11] leverages a latent diffusion model to synthesize multiple viewpoints of a scene by compressed 2D images and pose pairs. SGI-Net [12] generates target objects in the square masks

given a desired mask and class ID. OASIS [13] synthesizes urban scenes conditioned on semantic labels. Vobecký et al. [14] use person images and key points maps to generate new persons with specific poses in the background to augment datasets. Even though 2D synthesis models can generate more realistic images, there is still a problem of uncontrollability in perspective, as well as a low quality of appearance when utilized in cyclist synthesis. In this case, our work in this paper provides a solution.

2.2 Image Synthesis

Four primary types of generative models are studied in image synthesis. One of them is autoregressive models like Pixel-Stega [15]. They are conditional distributions developed over a convolution neural network or recurrent neural network. The feature of generating pixel by pixel makes it very slow and impractical. DALL-E [16] and CogView [17] predict image tokens from text tokens by combining a variational autoencoder (VAE) with a transformer. This kind of VAE model compresses the image into lower dimensions with less computational cost but still has the weaknesses of unidirectional bias.

Versions of StyleGANs [18–20] utilize latent space to control the level of the style of the generations. CLIP [21] trains text-image pairs to enable zero-shot image synthesis tasks. These models are classified as the well-known GAN (generative adversarial network) [22], which can generate excellent-quality images. However, the training process of GAN is difficult, which limits its application.

Lately, after Dhariwal et al. [23] presented a different method leveraging noise diffusion and removal, the diffusion models have become the most popular research direction because of their excellent performance and convenience of training. Many attempts have been made on different tasks [24], including image generation. GLIDE [25] and DALL-E2 [26] combine CLIP with diffusion models to achieve text-to-image generation. Imagen [27] adds a dynamic threshold and noise adjustment scheme into the model to produce more photorealistic images. Stable Diffusion [28] encodes inputs into a latent space to reduce computation complexity. Multidiff [29] units multiple diffusion models trained on multiple sub-tasks to solve combined tasks. Universal-Guided-Diffusion [30] presents a universal algorithm that enables diffusion models to generate images with arbitrary guidance without needing to be trained on specific conditions. In summary, diffusion models have developed rapidly, and the effect is continuously improving, which makes it possible to be applied in traffic scene generation.

3 Method

The entire process is composed of three stages, as shown in Fig. 1: 1) Creating a cyclist bank using images, semantic labels, and depth maps, categorizing the separated cyclists, and computing their perspectives. 2) Training an image synthesis model conditioned on edge maps and semantic labels. 3) Determining dynamic object locations and angles to be synthesized, matching and pasting cyclist candidates with similar perspectives, and inputting these new labels and image edges into the model to generate target objects.

3.1 Perspective System

A perspective system is first designed based on the principle of imaging. As demonstrated in Fig. 2a, which is a simplified top-view-diagram of a rider on a bicycle moving in front of a camera, φ is used to measure the object's pose and θ is put forward to determine the object's position relative to the camera. For clarity, the image plane, the camera coordinate system, the image coordinate system, and the pixel coordinate system are illustrated in Fig. 2b.

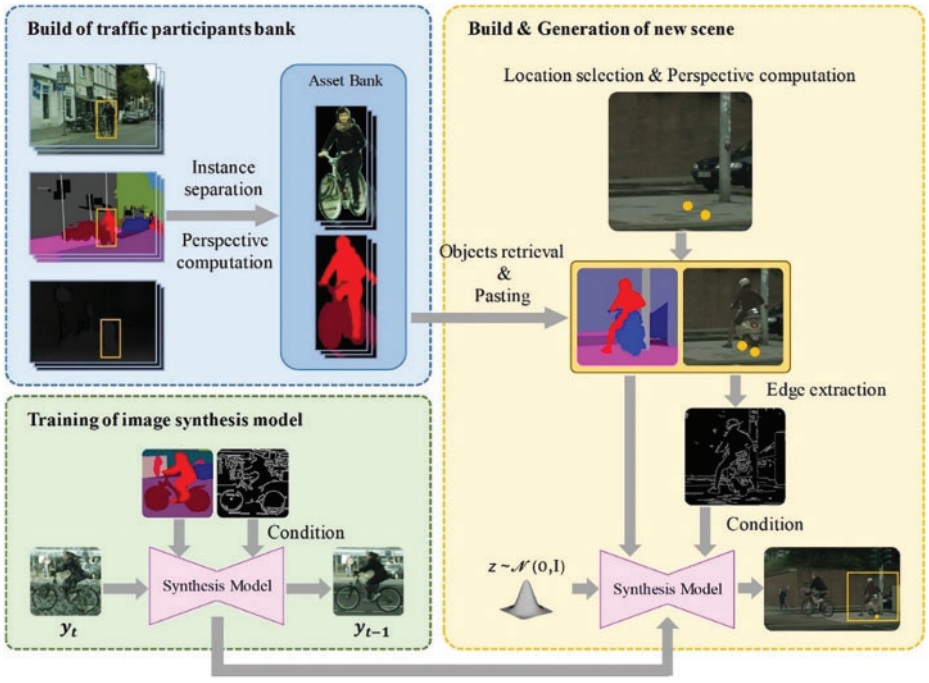
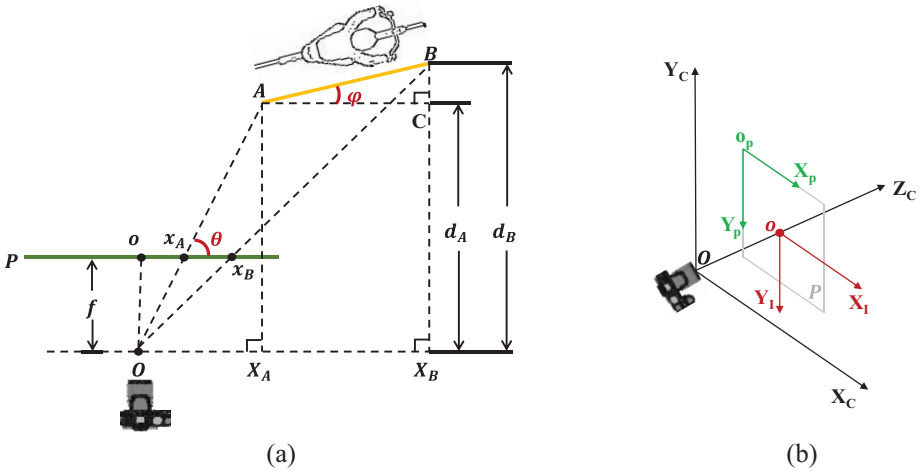


Figure 1: The pipeline of the proposed perspective-aware and realistic scene simulation method



(a) (b)

Figure 2: (a) Simplified top-view-diagram of a cyclist moving in front of a camera; (b) coordinate systems: the gray P is the image plane, the black $O - X_C - Y_C - Z_C$ is the camera coordinate system, the red $o - X_I - Y_I$ is the image coordinate system, and the green $o_p - X_p - Y_p$ is the pixel coordinate system

Given O as the camera and P as the image plane, the distance between O and P is the focal length f. Suppose that $\overline{AB}$, where A is the tail point (back wheel), and B is the head point (front wheel), represents a cyclist, then d_A and d_B can be obtained from the depth map. The angle φ in ΔBAC can be computed as follows:

$$\varphi = \arctan \frac{d_B - d_A}{X_B - X_A} \tag{1}$$

where X_A and X_B are the x coordinates of A and B, respectively, in the camera coordinate system.

Next, connecting A and O brings out x_A, which is the image coordinate of A in the image coordinate system. Moreover, x_B can be obtained similarly. Geometrically, ΔAX_AO and ΔOox_A are similar triangles, as well as ΔBX_BO and ΔOox_B, so that:

$$\frac{d_A}{X_A} = \frac{f}{x_A} \tag{2}$$

$$\frac{d_B}{X_B} = \frac{f}{x_B} \tag{3}$$

According to the transformation relationship between the image coordinate system and the pixel coordinate system,

$$u_A = \frac{x_A}{k} + c_x \tag{4}$$

$$u_B = \frac{x_B}{k} + c_x \tag{5}$$

where u_A and u_B are the x pixel coordinates of A and B in the pixel coordinate system, k is the size of every pixel in the x direction, and c_x is the x coordinate of the camera optical center in the pixel coordinate system. Bringing Eqs. (4) and (5) into Eqs. (2) and (3) and making $f_x = f/k$ can get:

$$X_A = \frac{d_A (u_A - c_x)}{f_x} \tag{6}$$

$$X_B = \frac{d_B (u_B - c_x)}{f_x} \tag{7}$$

Bringing Eqs. (6) and (7) into Eq. (1) and eliminating X_A and X_B can get:

$$\varphi = \arctan \frac{(d_B - d_A) f_x}{d_B (u_B - c_x) - d_A (u_A - c_x)} \tag{8}$$

In addition, θ is defined as the angle between $\overline{OA}$ and P. In Δox_AO, θ can be gained by:

$$\theta = \arctan \frac{f}{x_A} = \arctan \frac{f_x}{u_A - c_x} \tag{9}$$

Practically, as long as u_A and u_B are set, d_A and d_B can be obtained by the points with the exact pixel coordinates in the depth map, as shown in Fig. 3. Since f_x and c_x are the settled internal parameters of the camera, φ and θ can be calculated by Eqs. (8) and (9), respectively.

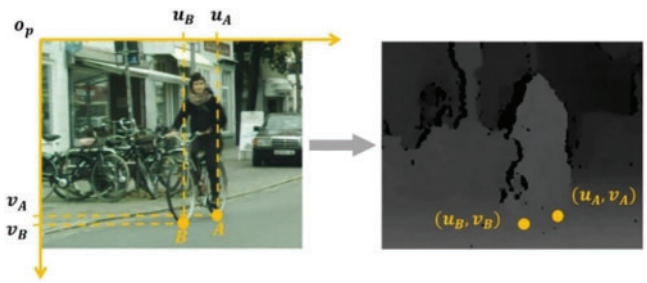


Figure 3: The cross point of the front wheel and the ground, B, and the cross point of the back wheel and the ground, A, are selected in the image to get the coordinates (u_B, v_B) and (u_A, v_A) with o_p as the origin of the pixel coordinate system, according to which the depth of B and A can be obtained respectively

There are also possible positions and poses of the objects other than Fig. 2a, several of which are shown in Fig. 4, but the calculations are all the same as discussed above. Special situations should be noticed as well. Assuming $f_x \neq 0$, when A and B are not overlapped in the image and $d_B = d_A$, two exceptional cases do not need to be calculated by Eq. (8). One is when $u_B > u_A$, φ is 0°. The other is when $u_B < u_A$, φ is 180°. If the head point is nearer to the camera, meaning $d_B < d_A$, φ should be the sum of the computed result and 180°. Also, when the tail point and the head point are overlapped, $u_A = u_B$, θ can be computed typically, and φ is equal to θ when the tail is visible, as shown in Fig. 4f, or equal to θ plus 180° if the head is visible.

In summary, the perspective of the objects can be calculated as below:

$$\theta = \begin{cases} \theta_c & \text{if } u_A > c_x \\ 90 & \text{if } u_A = c_x \\ 180 + \theta_c & \text{if } u_A < c_x \end{cases} \tag{10}$$

$$\varphi = \begin{cases} 0 & \text{if } d_B = d_A, u_B > u_A \\ 180 & \text{if } d_B = d_A, u_B < u_A \\ \varphi_c & \text{if } d_B > d_A, u_B \neq u_A, \varphi_c > 0 \\ 180 + \varphi_c & \text{if } d_B > d_A, u_B \neq u_A, \varphi_c < 0 \text{ or } d_B < d_A, u_B \neq u_A, \varphi_c > 0 \\ 360 + \varphi_c & \text{if } d_B < d_A, u_B \neq u_A, \varphi_c < 0 \\ \theta & \text{if } d_B > d_A, u_B = u_A \\ 180 + \theta & \text{if } d_B < d_A, u_B = u_A \end{cases} \tag{11}$$

where θ_c and φ_c are used as intermediate variables to keep the equations concise; they are equal to θ and φ in Eqs. (8) and (9), respectively.

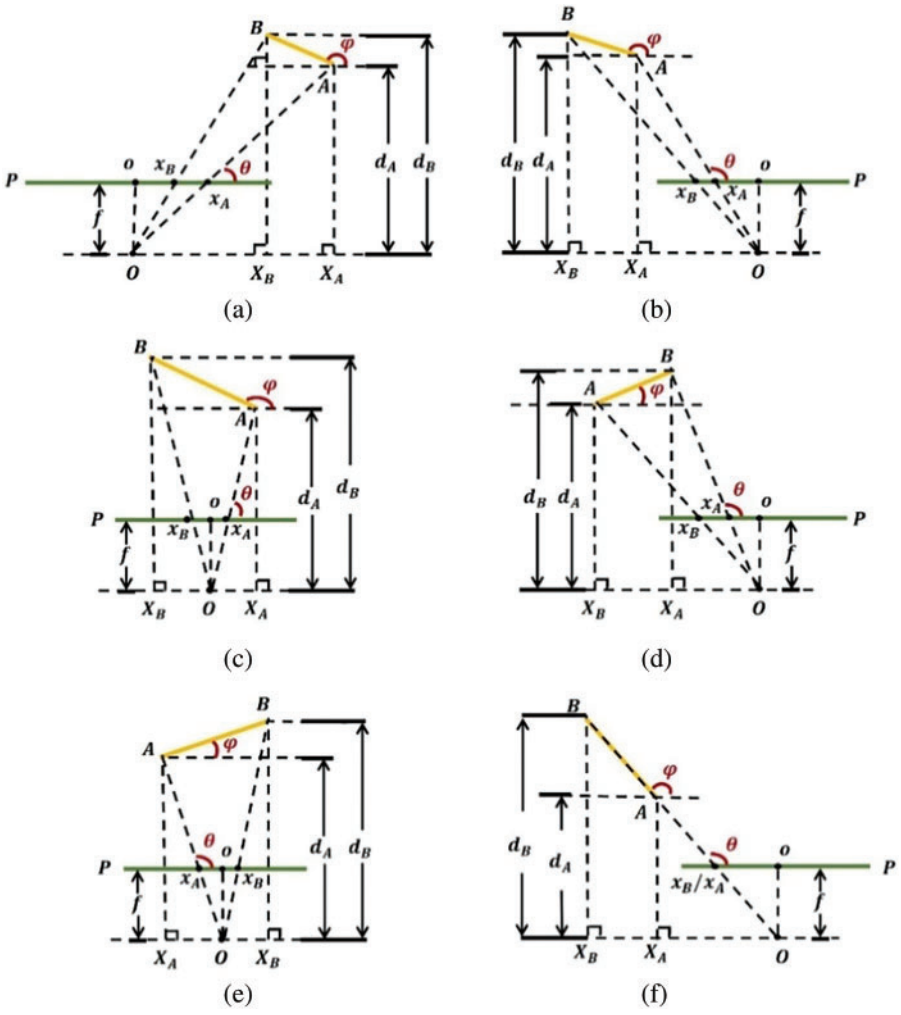


Figure 4: Some possible poses and positions of the objects: (a) when $x_A > x_B > o$; (b) when $o > x_A > x_B$; (c) when $x_A > o > x_B$; (d) when $o > x_A > x_B$; (e) when $x_B > o > x_A$; (f) when $o > x_B = x_A$

3.2 Build of the Bank of Cyclists

Existing urban scene datasets with semantic labels and depth/point cloud information can be used to build a cyclist bank. Cyclists are isolated and categorized using semantic maps. Their head/tail points are identified as key points (e.g., front/back wheel-ground intersections as shown in Fig. 5). A fine-tuned key point detection model [31] extracts these points' coordinates as shown in Fig. 5, which are further used for depth retrieval, enabling perspective calculation given camera parameters. Finally, backgrounds of dynamic object images are masked, leaving only cyclists, each labeled with category, φ, θ, and tail point depth.

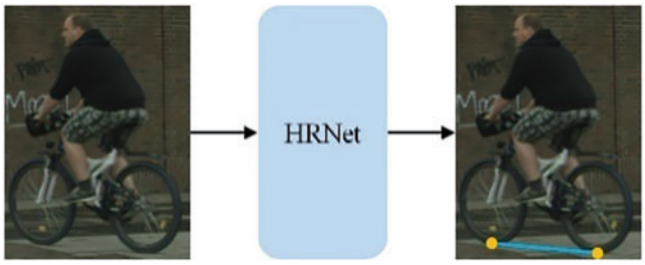


Figure 5: Head and tail points extraction

3.3 Design of New Scenes

The locations of cyclists' critical points in the new background are sampled from the area labeled "road" in the semantic map to calculate φ and θ. As shown in Fig. 6, C is chosen to be the location of the tail point, and D is selected so that $\overline{CD}$ can indicate the object's pose, noting that D is not the location of the heading point but only the point used for the calculation. Based on the coordinates and depths of C and D, the pose (φ) and location (θ) of the target object are determined. Alternatively, the value of θ and φ can also be directly assigned by the user.

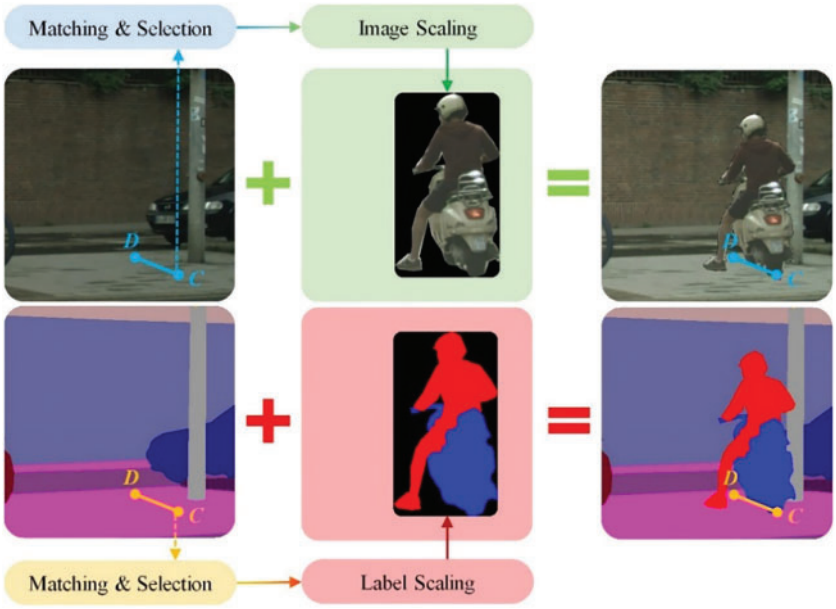


Figure 6: The process of the design of new scenes

The next step is to retrieve proper objects from the bank according to the desired category, φ and θ. Once picked, an object is scaled to the appropriate size for the correct view in the new background. The new height of the object is:

$$h_C = \frac{d_{ori}}{d_C} h_{ori} \tag{12}$$

where d_C is the depth of C, d_{ori} is the original depth of the chosen object's tail point, h_C is the new height of the object, and h_{ori} is the original height of the object from the bank.

Afterward, the processed object is pasted into the background by overlapping the tail point of the scaled object with that of the new background to prepare the new images and semantic labels. To this end, the design of the new scenes is finished.

3.4 Image Synthesis Model

The composition of the proposed new scenes utilizes the state-of-the-art diffusion model, which converts Gaussian noise into samples from a learned data distribution via a gradual denoising process. Given a diffusion model, the training process is based on a denoising objective of the form as follows [27]:

$$E_{(x,c,\epsilon,t)}\left[\|\epsilon - \epsilon_\theta\left(x_t, t, c\right)\|_2^2\right] \tag{13}$$

where x is the input, c is the condition, and $\epsilon \sim \mathcal{N}\left(O, I\right)$. $\epsilon_\theta\left(x_t, t, c\right)$ represents a sequence of denoising autoencoders, which is the denoising U-Net in Fig. 7, with t uniformly sampled from $1, \ldots, T$. It is trained to predict the added noise with the noisy version of x, x_t, as the input and c as the condition. When it comes to sampling, a noise x_t is input into a sampler like DDIM [32] to generate denoised images x step by step.

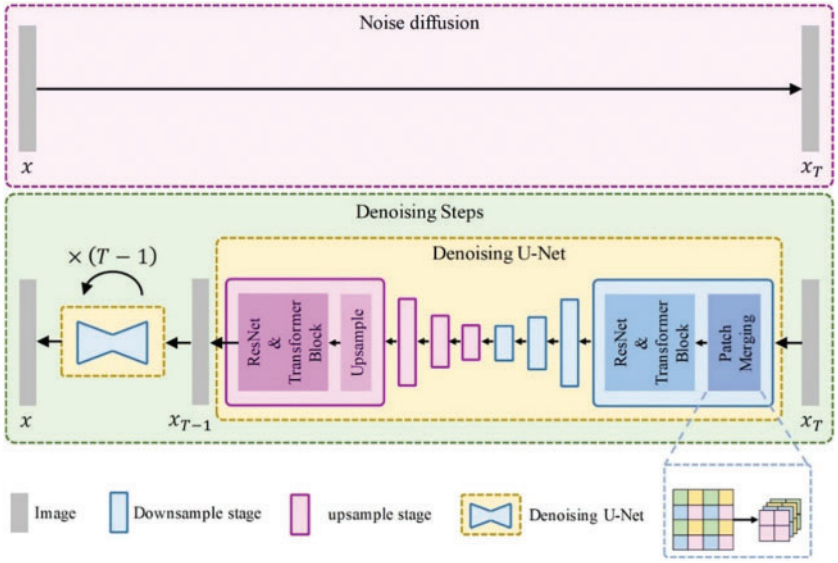


Figure 7: Visual depiction of the presented U-Net in the diffusion model

The base denoising architecture presented in this paper is a U-Net [33], which has a structure similar to Imagen [27] but conditioned on both semantic labels and edge maps. These additional conditions are concatenated with the input at the initial stage, differing from the original text condition. As is shown in Fig. 7, in order to reduce the size of the model, the original two convolutions at the beginning of each downsampling stage are replaced by an overlapped patch merging process containing an unfold operator with a kernel size of 3 and a convolution with a kernel size of 1.

3.5 Generation of New Scenes

Finally, the pasted images and the trained synthesis model are used to generate new scenes. As shown in Fig. 8, a square image centered by the pasted object is cropped out, along with its semantic label, forming a black mask of the cyclist. The semantic label and the edge map extracted from the cut

image are fed into the synthesis model as conditions to generate a new covered image that only keeps the moving object. This mask and the generated object are then merged into the original scene.

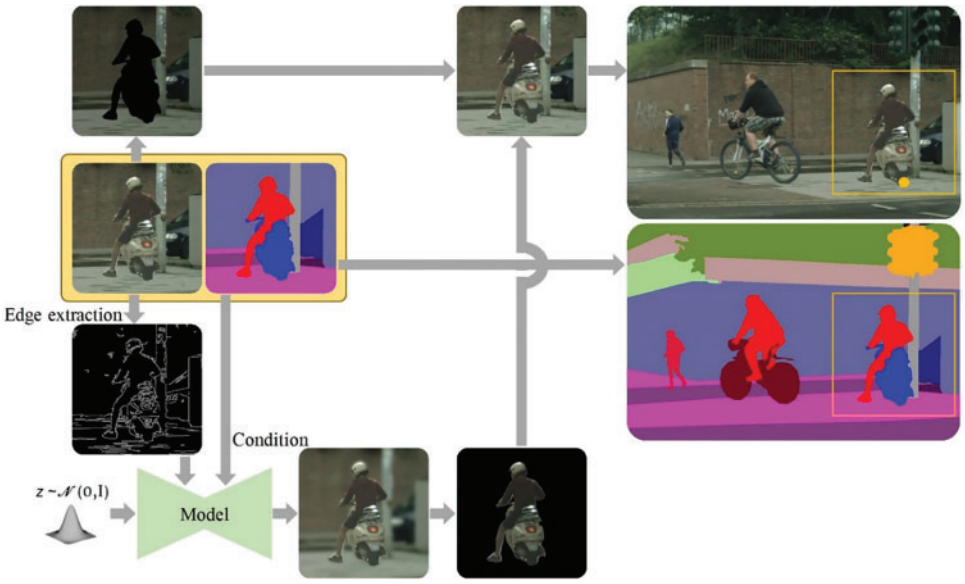


Figure 8: Generation process of the new scenes

4 Experiments

In this section, the experimental settings are first introduced. Then, the proposed method is compared with other image simulation baselines in visual realism to show its advantages. Next, ablation studies are carried out on the synthesis model used in our pipeline. Finally, the effects of downstream tasks are demonstrated to prove that the method presented is effective and practical.

4.1 Experimental Implementation Details

4.1.1 Datasets

Datasets used in the training of the synthesis model are Cityscapes [34], BDD100K [35] and Mapillary Vistas [36]. The asset bank and the backgrounds used in the comparisons are from Cityscapes.

1. Cityscapes focuses on urban street scenes, and its semantic labels and depth labels are essential for this research. Only fine annotated images are utilized in this paper.
2. BDD100 k is an all-weather and all-light large dataset of urban street scenes. It includes 10000 annotated images segmented in the same way as Cityscapes.
3. Mapillary Vistas is the largest self-driving dataset crowd-sourced worldwide. It contains 25000 high-resolution images with 124 annotated object categories.

4.1.2 Training of the Synthesis Model

Images and semantic maps with all cyclists in the center are cropped and collected from the three datasets mentioned in the previous subsection to assemble the training data containing 4600 examples.

The training is operated based on Pytorch and Python. The inputs of the model are the annotation and the edge map of the image, which are resized to squares with a size of 128 and randomly flipped horizontally. The U-Net structure in the diffusion model is the same as that of Imagen [27], except for our modified parts. AdamW, with a learning rate of 0.0001, is used as the optimizer. The learning rate is firstly warmed up with LinearWarmup, which increases the learning rate from 0 to the setting value linearly for 2600 steps and then decayed by CosineAnnealingLR, which reduces the learning rate gradually. The changing curve is in the shape of the cosine function, with 80 thousand T-max steps. The model is trained for 800 thousand iterations with a batch size of 1.

4.1.3 Cyclist Asset Bank

Candidates larger than 200 in size and with apparent appearances are processed by the proposed pipeline and registered in the asset bank by category, pose angle, position angle, and tail depth. Finally, 300 pairs of images and semantic maps with the unwanted area blacked out, and their corresponding depth maps are contained in the asset bank utilized in this paper's experiments.

4.2 Comparisons

4.2.1 Baselines for Comparison

Two deep-learning-based image synthesis baselines are utilized to conduct the comparison. Unlike the presented method, they cannot address realism and perspective at the same time. For the appearance comparison, objects with different perspectives are synthesized and inserted into the corresponding reasonable positions in the backgrounds to build new scenes that do not exist in the original Cityscapes. For the evaluation of realism, the customarily used metric FID (Fréchet Inception Distance, which calculates the distance of distribution of real images and that of fake ones to evaluate the quality of the generated images) [37] is calculated on images with the size of 1024×2048, containing cyclists from the validation set of Cityscapes. The target objects are more significant than 32 in height or width. If there is more than one eligible target in an image, only the largest one is selected for synthesis, and the rest of the background retains the original ground truth. There are 253 images in the final test dataset for the comparisons. The baselines are introduced as below:

1. SGI-Net [12] can reconstruct the masked area of an image and insert desired objects into it by providing the model with a class ID. Official codes are retrieved from Github, and the model is trained according to the author's instructions, except that the class IDs of interest are changed from car and pedestrian into rider, bicycle, and motorcycle.
2. OASIS [13] is a state-of-the-art 2D synthesis model generating images based on semantic maps. Since it cannot control perspective and only performs synthesis, semantic maps generated by our proposed method are used as the input of the official checkpoint of the OASIS model to generate images.

4.2.2 Appearance Comparison

In this section, the presented method is compared with the baselines in terms of the appearance of the simulated images. As shown in Fig. 9, it is hard for SGI-Net to generate complicated objects like cyclists, and the results can barely be recognized. Meanwhile, OASIS, as the representative of typical 2D synthesis models, cannot form perspective-controllable semantic labels for guidance, and the generated cyclists exhibit an artificial appearance and are short of details. In short, our method can effectively control both the perspective and the appearance of the produced objects.

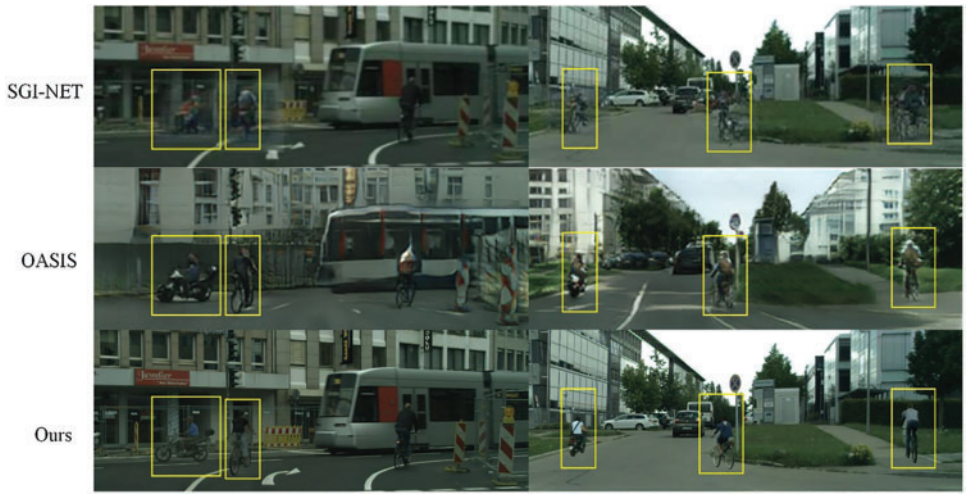


Figure 9: Appearance comparison of image simulation approaches

4.2.3 Perceptual Comparison

Perceptual comparisons are further conducted by calculating FID between the synthesized images and the ground truth to measure the generations' quality. The results listed in Table 1 show that our method surpasses all the other competing methods with an FID of 2.36, while SGI-Net gets 10.74 and OASIS gets 3.45, noting that the original images replace all the backgrounds of the three methods, and only the cyclists are considered in the computation of FID.

Table 1: Perceptual and computational comparisons of the synthesis methods. The smaller the FID, the better the performance

Method	SGI-Net	OASIS	Ours
FID	10.74	3.45	2.36
FLOPs (G)	13.32	208.12	175.32

4.2.4 Computational Comparison

Although the objectives of this work are perspective control and the realistic appearance of the synthesized cyclists, computational comparisons are presented by calculating FLOPs (floating point operations) of the models with the input size of $1 \times 3 \times 256 \times 256$. The results in Table 1 indicate that our method has 175.32 GFLOPs, while OASIS has 208.12 GFLOPs, which means our work can achieve better performance with less computational cost than OASIS. Regarding the comparison with SGI-Net, even though our calculation consumption is higher than it, our visual appearance is significantly superior. All in all, our proposed model has an excellent effect with relatively fast computing speed.

4.3 Ablation Study

Ablation studies are carried out to validate the necessity of the model design. Firstly, the edge map is removed from the inputs to demonstrate the difference. Then, the input is set as only the edge map to showcase the rationality of semantic maps as the condition. The other settings of the model's training,

except inputs for the two ablation studies above, are the same as those of the official one described in Section 4.1.2.

As displayed in Table 2, when only the semantic map or the edge map is the input, the FIDs are 5.99 and 4.30, respectively. Both results are worse than our official method, which has an FID of 2.36, which means that our design can surely improve the performance of the synthesis.

Table 2: Perception comparison of ablation studies

Method	W/o edge map	W/o semantic map	Ours
FID	5.99	4.30	2.36

4.4 Downstream Tasks

Two applications are investigated in this section to show the practical value of the proposed method.

4.4.1 Build of Scenarios

The most suitable downstream task goes to the build of desired scenarios. Nowadays, self-driving cars are well-trained for everyday situations, as they are readily available. However, acquiring some corner cases like accident ahead scenarios is impractical and dangerous due to their rarity and difficulty in replication. In this case, our method enables the selection of cyclists from a specific perspective and puts them in desired locations in the backgrounds. The realistic appearance obtained by the proposed approach can fulfill the training and testing of self-driving cars. Examples are depicted in Fig. 10.

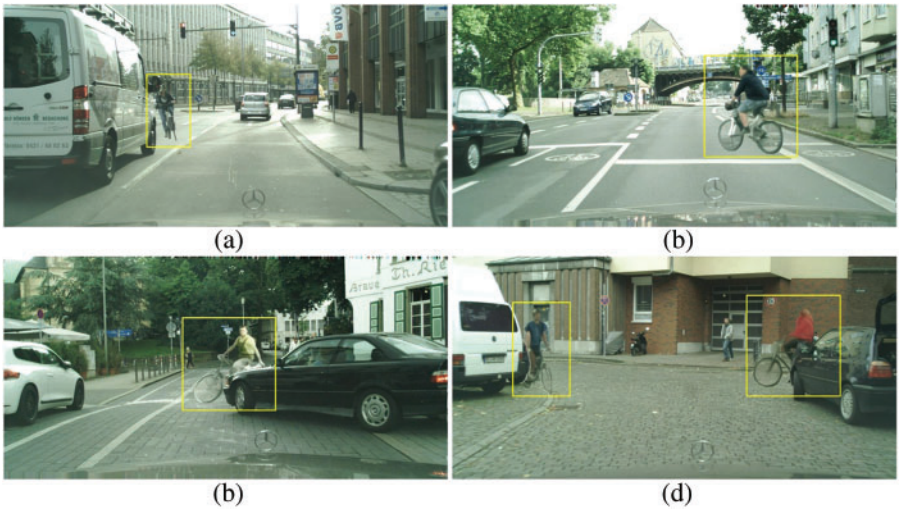


Figure 10: Scenarios of accidents ahead: (a) and (b) depict a cyclist crossing the road recklessly against the traffic lights; (c) shows the scene immediately before the crash, and (d) illustrates bicycles approaching, covered by objects

4.4.2 Data Augmentation

Another common usage is to augment training data to improve the performance of perception algorithms. Firstly, the classic segmentation model ResNet-50 [38] is trained on the official training set of Cityscapes containing 2975 images. Then, cyclists with motorcycles are inserted into the original 2975 images by the pipeline presented in this paper to obtain 2550 additional training images gathered together with the initial training set to form the augmentation for the training of ResNet-50. Both training processes are conducted for the same iterations with the same hyper-parameters.

The evaluations are operated on the validation set of Cityscapes and estimated by the standard metric, mean of class-wise intersection over union (mIoU), which is the higher, the better. It is observed in Table 3 that the semantic segmentation performance of the added category, motorcycle, is enhanced by 4.47%. Apparently, the proposed perspective-aware realistic image simulation method can upgrade the perception model by augmenting training data without requiring additional data collection and annotation since the materials are from existing datasets, and the presented pipeline can assemble the semantic maps and generate images simultaneously.

Table 3: Semantic segmentation results of ResNet-50 trained on the training set of Cityscapes and augmented data

Result (mIoU%)	Motorcycle
Training set	54.33
Augmentation	56.76

5 Conclusion

This paper proposes a perspective-aware and realistic cyclist generation pipeline via object retrieval. Cyclists are gathered from existing datasets, and their perspectives are calculated using the presented imaging approach to build the asset bank. Then, locations in the background are picked on demands, and the desired perspectives of the inserted objects are computed so that the matched objects can be retrieved from the bank to be placed in the background, forming semantic maps and pasted images. With the help of the cropped semantic maps and edges extracted from the pasted images, the synthesis model described in this paper can generate realistic images with appropriate angles. It is shown in the substantial experiments that the presented method can produce images with outstanding realism compared to the competing alternatives and qualitatively regulate the generated cyclists' perspectives at the same time. Furthermore, the practical usages, including the building of challenging scenarios and data augmentation, prove that the proposed method can be utilized to design and build specific scenes with real-world appearance, significantly benefiting the autonomous vehicle industry in perception model training and testing.

Acknowledgement: None.

Funding Statement: This work was supported by the Cultivation Program for Major Scientific Research Projects of Harbin Institute of Technology (ZDXMPY20180109).

Author Contributions: The authors confirm their contributions to the paper as follows: study conception and design: Beike Yu, Dafang Wang; data collection: Beike Yu, Xing Cui; analysis and

interpretation of results: Beike Yu, Xing Cui, Bowen Yang; draft manuscript preparation: Beike Yu, Bowen Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Dafang Wang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Conf. Robot Learn.*, PMLR, 2017, pp. 1–16.

[2] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, Springer, 2018, pp. 621–635.

[3] G. Rong *et al.*, "Lgsvl simulator: A high fidelity simulator for autonomous driving," in *2020 IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, IEEE, 2020, pp. 1–6.

[4] H. S. Li, "Simplified unsupervised image translation for semantic segmentation adaptation," *Pattern Recogn.*, vol. 105, no. 10, 2020, Art. no. 107343. doi: 10.1016/j.patcog.2020.107343.

[5] J. Choi, D. Ha Kim, S. Lee, S. H. Lee, and B. C. Song, "Synthesized rain images for deraining algorithms," *Neurocomputing*, vol. 492, no. Jul. 1, p. 492, 2022. doi: 10.1016/j.neucom.2022.04.034.

[6] Z. Lv, X. Li, Z. Niu, B. Cao, and W. Zuo, "Semantic-shape adaptive feature modulation for semantic image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11214–11223.

[7] Y. Shi, X. Liu, Y. Wei, Z. Wu, and W. Zuo, "Retrieval-based spatially adaptive normalization for semantic image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11224–11233.

[8] V. Musat, D. De Martini, M. Gadd, and P. Newman, "Depth-SIMS: Semi-parametric image and depth synthesis," in *Int. Conf. Robot. Autom. (ICRA)*, IEEE, 2022, pp. 2388–2394.

[9] A. Konushin, B. Faizov, and V. Shakhuro, "Road images augmentation with synthetic traffic signs using neural networks," *Comput. Opt.*, vol. 45, no. 5, pp. 736–748, 2021. doi: 10.18287/2412-6179-CO-859.

[10] N. Li, F. Song, Y. Zhang, P. Liang, and E. Cheng, "Traffic context aware data augmentation for rare object detection in autonomous driving," in *2022 Int. Conf. Robot. Autom. (ICRA)*, IEEE, 2022, pp. 4548–4554.

[11] S. W. Kim *et al.*, "NeuralField-LDM: Scene generation with hierarchical latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8496–8506.

[12] P. Ardino, Y. Liu, E. Ricci, B. Lepri, and M. De Nadai, "Semantic-guided inpainting network for complex urban scenes manipulation," in *2020 25th Int. Conf. Pattern Recognit. (ICPR)*, 2020, pp. 9280–9287.

[13] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele and A. Khoreva, "OASIS: Only adversarial supervision for semantic image synthesis," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 2903–2923, 2022. doi: 10.1007/s11263-022-01673-x.

[14] A. Vobecký, D. Hurych, M. Uřičář, P. Pérez, and J. Sivic, "Artificial dummies for urban dataset augmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, pp. 2692–2700, 2021. doi: 10.1609/aaai.v35i3.16373.

[15] S. Zhang, Z. Yang, H. Tu, J. Yang, and Y. Huang, "Pixel-Stega: Generative image steganography based on autoregressive models," 2021, *arXiv:2112.10945*.

[16] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8821–8831.

[17] M. Ding *et al.*, "CogView: Mastering text-to-image generation via transformers," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 19822–19835, 2021.

[18] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8110–8119.
[20] T. Karras *et al.*, "Alias-free generative adversarial networks," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 852–863, 2021.
[21] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.
[22] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.
[23] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 8780–8794, 2021.
[24] F. Daneshfar, A. Bartani, and P. Lotfi, "Image captioning by diffusion models: A survey," *Eng. Appl. Artif. Intel.*, vol. 138, no. 1, 2024, Art. no. 109288. doi: 10.1016/j.engappai.2024.109288.
[25] A. Nichol *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741*.
[26] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, *arXiv:2204.06125*.
[27] C. Saharia *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 36479–36494, 2022.
[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[29] N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Unite and conquer: Plug & play multi-modal synthesis using diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 6070–6079.
[30] A. Bansal *et al.*, "Universal guidance for diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 843–852.
[31] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
[32] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Med. Image Comput. Comput.-Assist. Interv.-MICCAI 2015: 18th Int. Conf.*, Munich, Germany, Springer, Oct. 5–9, 2015, pp. 234–241.
[34] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[35] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2636–2645.
[36] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4990–4999.
[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 6626–6637, 2017.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.