2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)
October 8-12, 2022, Macau, China

SST-GAN: Single Sample-based Realistic Traffic Image Generation for Parallel Vision

Jiangong Wang, Yutong Wang, *Member, IEEE,* Yonglin Tian,
Xiao Wang, *Senior Member, IEEE,* and Fei-Yue Wang, *Fellow, IEEE*

2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) | 978-1-6654-6880-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITSC55140.2022.9921999

***Abstract*— To improve their adaptability to various kinds of driving situations, deep learning-based vision algorithms need images from rare scenes, such as extreme weather conditions and traffic congestions. However, most datasets collected from physical driving environments are lack of such images, making vision models trained on these datasets do not work well in scarce scenes. Thus, we design an SST-GAN method for controllably generating realistic images of scarce driving scenes based on the framework of parallel vision. Trained on only a single sample, SST-GAN can produce hundreds of rare scene images from two directions: style transfer and content generation. Specifically, a transition retraining method is designed to transfer the weather and lighting styles from common scenes to scarce scenes, and a structural similarity index loss is used as reconstruction loss to guarantee the trained network can obtain more realistic content modification and generation during the image reconstruction. Experimental results show that SST-GAN outperforms the state-of-the-art method on expanding the amount of scarce scene images from both style and content. The method is highly adaptable and works flexibly on handling image generation problems for various types of rare scenes.**

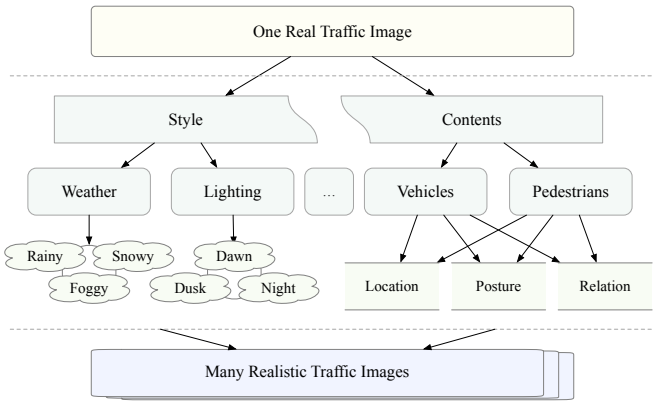

Fig. 1. The main elements that make up a traffic image can be roughly divided into two parts: scene style and scene contents. By modifying one or more elements in a traffic image, a variety of realistic traffic images can be generated.

I. Introduction

In recent years, autonomous driving has gradually become a hot research topic where more and more advanced technologies such as deep learning-based computer vision are researched and applied [1]–[4]. The autonomous vehicles access to most information by vision-based scene perception and understanding [5], [6]. Deep learning-based computer vision has excellent performance in many fields and is naturally used in vision perception tasks for autonomous driving [7]–[9]. So, many companies and research institutes such as Tesla and Baidu Research have invested and studied much in deep learning-based computer vision for autonomous driving [10].

It is well known that deep learning has a huge demand for large-scale and diverse datasets with annotations but such a dataset is difficult to obtain. The traditional method of image collection is by manually photographing and labeling images from real scenes. With the expansion of the autonomous driving industry and the investment of significant funding, many larger-scale traffic image datasets have been collected by traditional methods [11]. However, since the images in these datasets are basically from common traffic scenes, they cannot meet the demand for the diversity of datasets.

In particular, the diversity and completeness of training data is so important in areas such as autonomous driving, where safety is strongly required [12]. The rarer the scenario, the more likely it is to interfere with the automated driving system and cause a traffic accident while the vehicle is in motion [13], [14]. In order to solve the problem of visual perception and understanding in rare scenes, we must first obtain as many images from these scenes as possible and feed them to the model during the training process.

As mentioned above, the contradiction between rare scenarios in reality and the need for numerous learning samples cannot be solved by traditional methods of image collection, and virtual data generation based on parallel vision can solve it. In previous studies, it is a good method to build a virtual world that can change at will using game engines such as Unity 3D, Unreal and Carla [15]. Examples include SIM10K [16], a virtual traffic dataset built on the computer game Grand Theft Auto V, and Virtual KITTI [17], a virtual dataset corresponding to KITTI [18], built by Unity 3D. In addition, we can even use the game engine to build a virtual world that reproduces the real world, such as ParallelEye and ParallelEye-CS [19]. In summary, building virtual datasets based on game engines can generate a large number of highly customized images by changing the parameter settings of the

This work was supported partly by the National Natural Science Foundation of China under Grant U1811463, and the Key Research and Development Program 2020 of Guangzhou under Grant 202007050002.

Jiangong Wang, Yutong Wang, Yonglin Tian, Xiao Wang and Fei-Yue Wang are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Jiangong Wang is also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

Xiao Wang is also with Qingdao Academy of Intelligent Industries, Qingdao 266000, China.

Fei-Yue Wang is also with Institute of Systems Engineering, Macau University of Science and Technology, Macau 999078, China. feiyue.wang@ia.ac.cn

virtual world. That is to say, lots of corresponding images can be obtained by simulating rare real-world traffic scenarios in the virtual world. This already seems to be a perfect method for controllable image generation, but an obvious drawback is that the images generated by game engines are clearly recognizable to be fake nowadays. Compared to actual or realistic images, it is more difficult to train visual networks on such images generated by game engines.

In order to further solve the problem that the generated images are not realistic enough, we propose a controllable method, SST-GAN to generate realistic traffic images for parallel vision. In contrast with virtual image generation methods based on game engines [19], the proposed method is based on generating adversarial network (GAN) and other computer vision methods. The network is trained on a single image from scarce scenes, and processes images captured in the real world to controllably obtain realistic generated images. The generated new images not only have the scene characteristics relevant to the task requirements, but also maintain the similarity to the original real images.

In general, the difference between rare scenes and common scenes in autonomous driving is mainly reflected in two aspects: scene style and scene contents, as shown in Fig. 1. The difference in scene style is primarily reflected in the weather (i.e., sun, rain and fog) and lighting (i.e., dawn, dusk, night), and the scene contents mainly include the position or postures of traffic subjects such as vehicles and pedestrians and the relation between them [20]. The proposed method can controllably modify the scene style and scene contents of real images so that the generated images can be used as a supplement to make up for the lack of rare scenes in the original dataset.

Overall, this paper has the following main contributions.

- The controllable image generation method, SST-GAN trained on only a single sample can generate a wide variety of scarce images from the perspective of scene style and scene content as needed.
- We introduce new transition retraining methods for transferring the style of images while a reconstruction loss function focusing on image structural similarity for content generation into SST-GAN.
- Compared with the state-of-the-art method trained under the same conditions, SST-GAN performs better.

The rest of the paper is organized as follows. In section II, the related work such as parallel vision is summarized and introduced. Section III proposes the approach to generating realistic traffic images. And section IV shows the experimental results and analysis. Lastly, section V presents the conclusion of this paper.

II. Related Work

A. Parallel Vision

The theory of parallel vision [21] originates from parallel systems [22]. With the rapid development of computer vision and the research of parallel systems being enriched and improved in practice [23]–[26], parallel vision was proposed as a new intelligent method for visual computing to solve the visual problems of complex scenes. ACP (Artificial systems, Computational experiments, and Parallel execution) is the core idea of parallel vision. Among them, image generation is the focus of artificial systems, realistic images are the basis of computational experiments, and controllability is the key to parallel execution.

Parallel vision has been widely applied in the field of autonomous driving. Based on the artificial system in parallel vision, virtual worlds can be built with game engines. ParallelEye and ParallelEye-CS [19] are generated in this way and both of them are virtual traffic datasets with annotations. These images can be used to train and evaluate visual models in various traffic vision tasks. In addition, parallel vision can solve the visual long-tail problem in autonomous driving [13]. LoTR (Long-tail Regularization) and PVAS (Parallel Vision Actualization System) based on parallel vision are proposed and applied in IVFC (Chinese Intelligent Vehicles Future Challenge), the longest-lasting autonomous driving competition in the world. Moreover, it is worth mentioning that controllable image generation techniques based on game engine have been required and plays a key role in this work.

B. Scene Style Transfer

Style transfer has been studied as a separate topic within the field of computer vision. It can convert an image to the target style while preserving the basic contents of the image [27]. Foggy Cityscapes [28] adds fog to the normal images in the Cityscapes [29] dataset by building an optical model of fog and combining it with the depth information of the image. Since images acquired under foggy conditions are severely inadequate, Foggy Cityscapes has been widely used for numerous related tasks and studies, even though it is a dataset consisting of manually generated images.

Style transfer based on traditional image processing requires the construction of a specific mathematical model or image processing algorithm for each target style. GAN-based style transfer networks, on the other hand, are able to generate images for multiple extreme weather conditions or lighting conditions simultaneously in a data-driven manner. WeatherGAN [30] is specifically designed for the task of transferring the weather conditions of an image from one category to another. It enables the transfer of weather conditions between sunny, cloudy, foggy, rainy and snowy days. In addition, Vinod et al. [31] also implemented the transfer of light conditions from day to night.

C. Scene Contents Generation

For simple images containing only a single object such as face, a lot of sufficiently realistic images can usually be generated with GAN due to the simplex contents of the scene [32], [33]. However, traffic scenes often have complex contents, and a single image contains a complex background and numerous foregrounds. The existing deep learning techniques such as GAN are unable to directly generate realistic images, and they also cannot achieve the

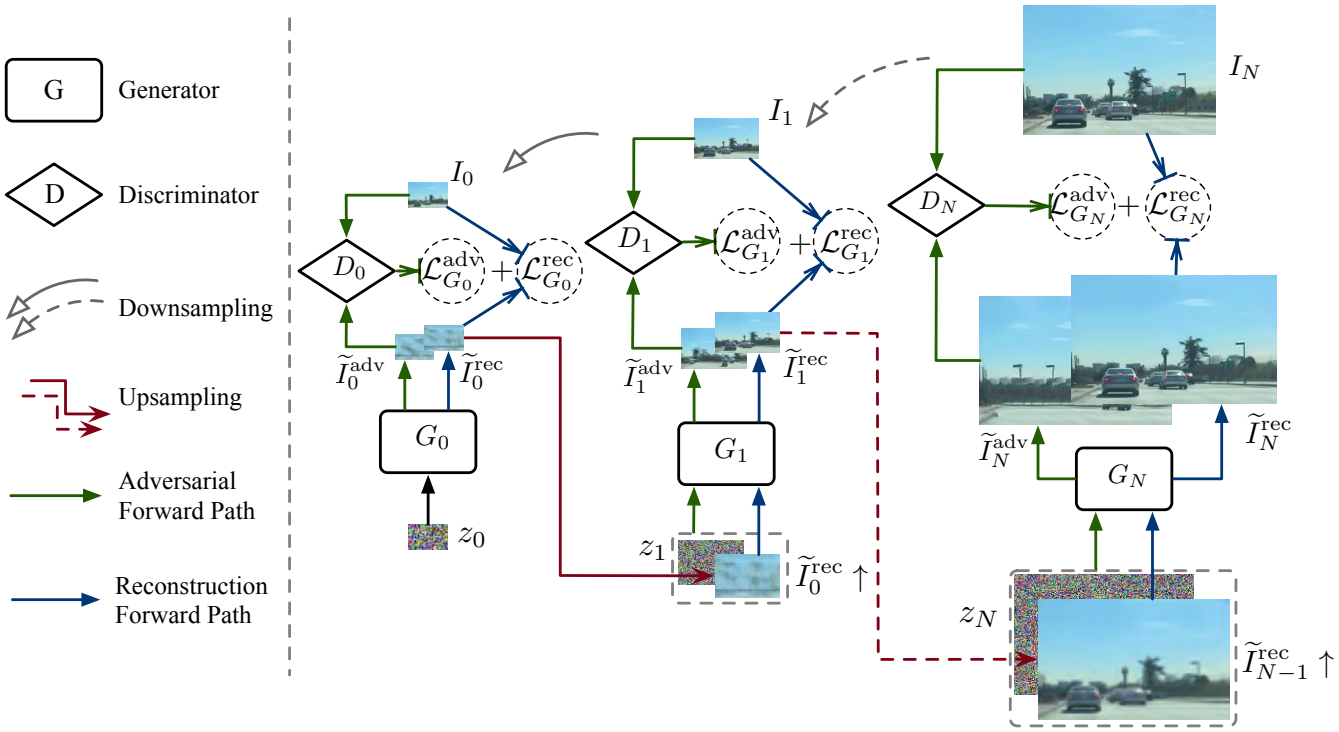

Fig. 2. Illustration of model architecture and training process. The architecture of the model is a pyramid of GANs from coarse to fine. I_i represents the real images of different scales obtained after the training sample is downsampled with configured scale factors. These real images are applied to adversarial training and reconstruction training of the generator network in each layer of the pyramid. z_i and $\widetilde{I}_{i-1}^{\text{rec}} \uparrow$ denote the noisy inputs and the inputs from the generator network of the upper layer G_{i-1}, respectively. In addition, $\widetilde{I}_i^{\text{adv}}$ and $\widetilde{I}_i^{\text{rec}}$ denote the adversarial and reconstruction results of the generator in each layer. Where $\widetilde{I}_{i-1}^{\text{rec}} \uparrow$ is the upsampling result of $\widetilde{I}_{i-1}^{\text{rec}}$ with the same scale factor as I_i to I_{i-1}. (Best viewed in color.)

controllable generation of images with various scene contents. Currently, many researchers have proposed new ideas and methods in the fields of image inpainting [34] and image augmentation [35], which inspire us to think and practice for controllable generation of scene contents in complex scenes.

DVI [36] can synthesize regions temporarily obscured by traffic agents in traffic videos using depth or point cloud information. As a result, the erasure of dynamic traffic agents such as vehicles and pedestrians in the traffic video can be achieved. Zhang et al. [37] present a method to add virtual pedestrians to real images and control the generation of diversified images by modifying the position and posture of the virtual pedestrians. Vobecký et al. [38] propose another method to insert real pedestrian images to the real images. Also, the different arrangement of person keypoints can generate images of pedestrians with different postures, thereby controlling the postures of added pedestrians. AADS [10] builds a huge data augmentation system of autonomous driving by fusing kinds of information such as traffic flow, point cloud, depth, and semantic annotation, which can erase or insert vehicles in traffic videos.

III. Methodology

Style transfer methods based on deep learning such as CycleGAN [39] often requires a large number of target style images as training set, which is not feasible in some scarce scenes. And 3D-based content generation methods require additional information such as depth or point cloud, which is obtained at an additional cost and expense. In addition, there is a lack of a unified way to both transfer the style of images and generate scene content. Inspired by recent works [40], [41] on super resolution and texture expansion trained with a single natural image, this paper proposes a novel method to augment and modify traffic images in terms of both style and content with only a single training image.

A. Model Architecture and Training

The network proposed in this paper has a pyramidal architecture consisting of several generative adversarial networks. Similar GAN-based pyramid architectures have been studied and explored in several works [41], [42], and they are applied to generate realistic traffic images in this paper.

As shown in Fig. 2, the GAN networks located at each layer of the pyramid have similar structures and different scales of inputs and outputs. To easily illustrate the network, we use L_i to represent the i-th level of the pyramid. L_0 and L_N represent the top and bottom layer of the pyramid, respectively, and each layer from L_0 to L_N corresponds to a coarse-to-fine GAN. Similarly, with a scale factor close to 0.75, the original training sample I_N is downsampled to I_0 layer by layer. Then, I_i is used as the training sample of the GAN in layer L_i. z_i is the noise input of the generator G_i, while $\widetilde{I}_i^{\text{adv}}$ and $\widetilde{I}_i^{\text{rec}}$ are the output of G_i. Both of them have the same size as I_i. $\widetilde{I}_i^{\text{rec}} \uparrow$, as the input to G_{i+1}, is

the upsampling result of $\widetilde{I}_i^{\text{rec}}$. And, it has the same size as z_{i+1}. In addition, a residual structure similar to generator in SinGAN [41] as shown in Eq. 1 is applied to the generators of each layer.

$$\begin{aligned} \text{Out} &= G(\text{Noise},\ \text{Out}_{\text{pre}}) \\ &= \text{CNN}_G(\text{Noise} + \text{Out}_{\text{pre}}) + \text{Out}_{\text{pre}} \end{aligned} \tag{1}$$

Where, Out and Out_{pre} are the output of generator in the current and previous layers. CNN_G represents the convolutional neural network in the generator. Also, in order to accommodate images from more complex traffic scene, a larger network is applied to improve the model capacity.

The training of the whole network starts from layer L_0. And the training loss for the i-th GAN in layer L_i is comprised of an adversarial term and a reconstruction term,

$$\min_{G_i} \max_{D_i} \left[\mathcal{L}_{G_i,D_i}^{\text{adv}} + \alpha \mathcal{L}_{G_i}^{\text{rec}}\right] \tag{2}$$

In the adversarial training process of layer L_i, the input of G_i $(i > 0)$ consists of two parts, which are the random noise input and $\widetilde{I}_{i-1}^{\text{rec}} \uparrow$ from the previous layer L_{i-1}. The generated result is discriminated from the training sample I_i by a discriminator. As shown in Eq. 3, the WGAN-GP loss [43] is used as the adversarial loss to update the network.

$$\begin{aligned} \mathcal{L}_{G_i,D_i}^{\text{adv}} =& E\left[D_i\left(G_i\left(z_i, \widetilde{I}_{i-1}^{\text{rec}} \uparrow\right)\right)\right] - E\left[D_i\left(I_i\right)\right] \\ &- \lambda E\left[\left(\left\|\nabla_{\widetilde{I}_i^{\text{adv}}} D_i\left(\widetilde{I}_i^{\text{adv}}\right)\right\|_2 - 1\right)^2\right] \end{aligned} \tag{3}$$

Where, λ is the weight coefficient of the gradient penalty. In the reconstruction training process of layer L_i, $\widetilde{I}_{i-1}^{\text{rec}} \uparrow$ from the previous layer is the only input to G_i, and $\widetilde{I}_i^{\text{rec}}$ is obtained after the generator. With the reconstruction training, $\widetilde{I}_i^{\text{rec}}$ gradually approximates I_i under the constraint of the structural similarity index (SSIM) loss [44], $\mathcal{L}_{G_i}^{\text{rec}}$ as shown in Eq. 4.

$$\mathcal{L}_{G_i}^{\text{rec}} = 1 - \frac{\left(2\mu_{\widetilde{I}_i^{\text{rec}}}\mu_{I_i} + c_1\right)\left(2\sigma_{\widetilde{I}_i^{\text{rec}} I_i} + c_2\right)}{\left(\mu_{\widetilde{I}_i^{\text{rec}}}^2 + \mu_{I_i}^2 + c_1\right)\left(\sigma_{\widetilde{I}_i^{\text{rec}}}^2 + \sigma_{I_i}^2 + c_2\right)} \tag{4}$$

Where, μ and σ represent the mean and variance of the corresponding images, respectively. In addition, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are constants. Generally, $k_1 = 0.01,\ k_2 = 0.03$, and L is the dynamic range of the pixel values, $2^8 - 1 = 255$ for 8-bit depth images. The final effect of the reconstruction training satisfies Eq. 5. z_0 as the input can reconstruct the original training sample I_N after some series of trained generators.

$$I_N \simeq \widetilde{I}_N^{\text{rec}} = G_N(0,\ ...G_1(0,\ G_0(z_0) \uparrow) \uparrow ...) \tag{5}$$

B. Process of Controllable Style Transfer

In the coarse-scale layers of the pyramid, small changes in the image content can result in dramatic changes in the content of final output image. In contrast, in the fine-scale layers of the pyramid, the generator does not make major changes to the content of the input image, but only to the style of the image.

Therefore, when the network is applied to style transfer, we first train the pyramid network on images with the target style. At this time, the generators in the pyramid can learn the style information of the training sample while reconstructing it. When we directly feed the source image that needs to be transferred to the trained fine-scale generator, the generated image can retain the content information of the source image, but the style of the image is not very obviously transformed. It is because during the previous reconstruction training, the input and output images are similar, and the generators do not learn to make a style transfer. Therefore, we need to further train the fine-scale generator with the source and target images in order to enable the generators to learn to transfer the image style while keeping the image content essentially unchanged.

We select the finer scale generators (G_N or G_{N-1}) for further training. First, the target image is converted into a transition status such as the grayscale or quantized image, which has the same content information as the target image, but a different style transferred by traditional image processing methods. And then, the transition image is fed into the generator to reconstruct the target image. The generator further trained in this way has the ability to transfer the transition image to the target style without changing the content information. In the inference stage, the same methods are applied to convert the source image to a transition state as well. Finally, the transition image is fed into the well-trained generator to transfer the source image to target style.

C. Process of Controllable Contents Generation

After training the pyramid network with a single sample, we fix the parameters of generators. Referring to Eq. 5, a series of generators can gradually reconstruct the image from a specific noise input. During the reconstruction process, we can artificially modify the image fed into one of the coarse-scale generators. And the defects present in the modified image can be gradually improved by the finer-scale generators. In this way, realistic image content can be generated in a controlled manner.

The input image of the coarse-scale generator can be obtained by modifying the original image I_N and downsampling it. The modification of the original image is controllable and can be accomplished automatically by setting specific targets. However, the image with a simple modification is not realistic enough. With the aid of a series of fine-scale generators, it is possible to obtain a realistic image with the content that we design. Also, if you want to keep the unmodified area completely, you can generate a mask while modifying the image. The final target image is obtained by masking out the part of the generated image outside the modified area and then overlapping it with the original image.

IV. Experiments

In this section, we conduct experiments on the BDD100K dataset [11], which is the largest publicly autonomous driving dataset with the most diverse content. Therefore, we can find some traffic images in rare scenes from BDD100K as the

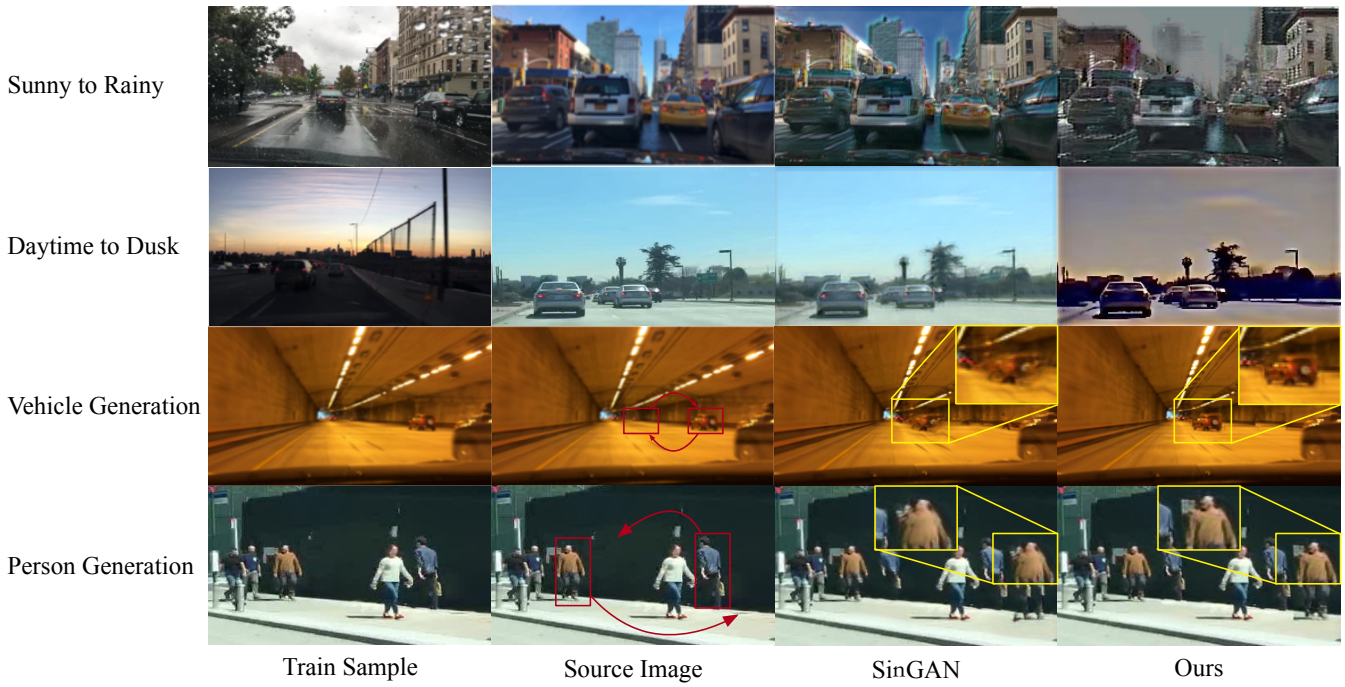

Fig. 3. Results of our proposed method for style transfer and content generation are shown in the figure above. Take the examples of sunny to rainy, daytime to night, and vehicle or pedestrian generation. The proposed method is also compared with SinGAN [41], which also uses a single sample as the training set. (Best viewed in color.)

training set. These scenes are rare in reality, and that's why they are often easily ignored and difficult to identify by the autonomous driving vision system.

Training the network on only a single target style image, we can change the style of the images such as weather and lighting conditions. As shown in the first two rows of Fig. 3, the images from common scenes such as sunny and daytime images are transferred into rainy and dusk images, respectively. In addition, the method proposed in this paper can also perform style transfer between images of various weather and lighting conditions. And the key is that only one sample in the rare style needs to be captured to enable the expansion of a large number of images.

Some traffic scenes, such as tunnels, are relatively rare in image datasets for automated driving. But as long as there is one image that has been captured, we can automatically modify and adjust the image based on the label of lanes, vehicles, pedestrians, etc. The directly modified images are not realistic enough, and our proposed method applied to them results in similar but more realistic images. As shown in the last two rows of Fig. 3, we have artificially modified the positions of vehicles and pedestrians in the real images, respectively, and then realistic images are generated.

Overall, our proposed method provides a novel and excellent solution to generate realistic images from both style and content directions when there is a severe lack of training samples. Also, as the comparative experiments with SinGAN which is also trained on a single sample, shown in Fig. 3, the proposed method can generate more realistic images. The style of images generated by our proposed method is closer to the target style, while the details of the traffic target are restored more realistically and carefully. In contrast, SinGAN trained under the same conditions generates images with little difference in style from the original images, and the generated traffic objects are heavily defaced, destroying the integrity of the objects and the authenticity of the images.

V. Conclusion

We introduce a controllable approach to generate realistic images for parallel vision. In particular, for some rare traffic scenes, there are only few or even a single training image that can be acquired. The proposed method trained on a single traffic image can controllably generate a lot of realistic images through both style transfer and content generation as needed. In addition, it is known from experiments that SST-GAN performs better under the same conditions compared with other methods.

References

[1] J. E. Hoffmann, H. G. Tosso, M. M. D. Santos, J. F. Justo, A. W. Malik, and A. U. Rahman, "Real-time adaptive object detection and tracking for autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 450–459, 2021.

[2] X. Wang, X. Zheng, W. Chen, and F.-Y. Wang, "Visual human-computer interactions for intelligent vehicles and intelligent transportation systems: The state of the art and future directions," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 253–265, 2021.

[3] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Y. Li, "Soft-weighted-average ensemble vehicle detection method based on single-stage and two-stage deep learning models," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 100–109, 2021.

[4] K. Zhao, L. Liu, Y. Meng, H. Liu, and Q. Gu, "3d detection for occluded vehicles from point clouds," *IEEE Intelligent Transportation Systems Magazine*, pp. 2–14, 2021.

[5] P. Lu, C. Cui, S. Xu, H. Peng, and F. Wang, "Super: A novel lane detection system," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 583–593, 2021.

[6] C. Zhou, Y. Liu, Q. Sun, and P. Lasang, "Vehicle detection and disparity estimation using blended stereo images," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 4, pp. 690–698, 2021.

[7] H. Zhang, G. Luo, Y. Tian, K. Wang, H. He, and F.-Y. Wang, "A virtual-real interaction approach to object instance segmentation in traffic scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 863–875, 2021.

[8] W. Zhang, J. Wang, Y. Wang, and F.-Y. Wang, "Parauda: Invariant feature learning with auxiliary synthetic samples for unsupervised domain adaptation," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2022.

[9] P. Cai, Y. Sun, H. Wang, and M. Liu, "Vtgnet: A vision-based trajectory generation network for autonomous vehicles in urban environments," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 419–429, 2021.

[10] W. Li, C. Pan, R. Zhang, J. Ren, Y. Ma, J. Fang, F. Yan, Q. Geng, X. Huang, H. Gong *et al.*, "AADS: Augmented autonomous driving simulation using data-driven algorithms," *Science Robotics*, vol. 4, no. 28, p. eaaw0863, 2019.

[11] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2633–2642.

[12] C. Gou, Y. Wu, K. Wang, K. Wang, F.-Y. Wang, and Q. Ji, "A joint cascaded framework for simultaneous eye detection and eye state estimation," *Pattern Recognition*, vol. 67, pp. 23–31, 2017.

[13] J. Wang, X. Wang, T. Shen, Y. Wang, L. Li, Y. Tian, H. Yu, L. Chen, J. Xin, X. Wu, N. Zheng, and F.-Y. Wang, "Parallel vision for long-tail regularization: Initial results from ivfc autonomous driving testing," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 286–299, 2022.

[14] L. Li, X. Wang, K. Wang, Y. Lin, J. Xin, L. Chen, L. Xu, B. Tian, Y. Ai, J. Wang, D. Cao, Y. Liu, C. Wang, N. Zheng, and F.-Y. Wang, "Parallel testing of vehicle intelligence via virtual-real interaction," *Science Robotics*, vol. 4, no. 28, p. eaaw4106, 2019.

[15] T. Feng, F. Fan, and T. Bednarz, "A review of computer graphics approaches to urban modeling from a machine learning perspective," *Frontiers of Information Technology & Electronic Engineering*, vol. 22, no. 7, pp. 915–925, 2021.

[16] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 746–753.

[17] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 4340–4349.

[18] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[19] X. Li, Y. Wang, K. Wang, L. Yan, and F.-Y. Wang, "The paralleleye-cs dataset: Constructing artificial scenes for evaluating the visual intelligence of intelligent vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 37–42.

[20] M. Schutera, M. Hussein, J. Abhau, R. Mikut, and M. Reischl, "Night-to-day: Online image-to-image translation for object detection within autonomous driving by night," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 480–489, 2021.

[21] K. Wang, C. Gou, N. Zheng, J. M. Rehg, and F.-Y. Wang, "Parallel vision for perception and understanding of complex scenes: methods, framework, and perspectives," *Artificial Intelligence Review*, vol. 48, no. 3, pp. 299–329, 2017.

[22] F.-Y. Wang, "Parallel system methods for management and control of complex systems," *Control and Decision*, vol. 19, no. 5, pp. 485–489, 2004.

[23] S. Wang, J. Housden, T. Bai, H. Liu, J. Back, D. Singh, K. Rhode, Z.-G. Hou, and F.-Y. Wang, "Robotic intra-operative ultrasound: Virtual environments and parallel systems," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 5, pp. 1095–1106, 2021.

[24] T. Liu, H. Wang, B. Tian, Y. Ai, and L. Chen, "Parallel distance: A new paradigm of measurement for parallel driving," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 4, pp. 1169–1178, 2020.

[25] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, and F.-Y. Wang, "Parallel transportation systems: Toward iot-enabled smart urban traffic control and management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4063–4071, 2020.

[26] T. Liu, B. Tian, Y. Ai, and F.-Y. Wang, "Parallel reinforcement learning-based energy efficiency improvement for a cyber-physical system," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 2, pp. 617–626, 2020.

[27] W.-t. You, H. Jiang, Z.-y. Yang, C.-y. Yang, and L.-y. Sun, "Automatic synthesis of advertising images according to a specified style," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 10, pp. 1455–1466, 2020.

[28] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, Sep 2018.

[29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[30] X. Li, K. chang Kou, and B. Zhao, "Weather gan: Multi-domain weather translation using generative adversarial networks," *ArXiv*, vol. abs/2103.05422, 2021.

[31] V. Vinod, K. R. Prabhakar, R. V. Babu, and A. Chakraborty, "Multi-domain conditional image translation: Translating driving datasets from clear-weather to adverse conditions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1571–1582.

[32] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[33] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[34] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, vol. 51, no. 2, pp. 2007–2028, 2020.

[35] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[36] M. Liao, F. Lu, D. Zhou, S. Zhang, W. Li, and R. Yang, "Dvi: Depth guided video inpainting for autonomous driving," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17.

[37] W. Zhang, K. Wang, Y. Liu, Y. Lu, and F.-Y. Wang, "A parallel vision approach to scene-specific pedestrian detection," *Neurocomputing*, vol. 394, pp. 114–126, 2020.

[38] A. Vobeckỳ, D. Hurych, M. Uřičář, P. Pérez, and J. Sivic, "Artificial dummies for urban dataset augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2692–2700.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[40] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3118–3126.

[41] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4570–4580.

[42] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5077–5086.

[43] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[44] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.