BiLSTM-based VAE-GAN for Predicting Future Road States in Autonomous Driving

Donghyun Kim
Division of Electrical Engineering
Hanyang University
Ansan, South Korea
kissw@hanyang.ac.kr

Jaerock Kwon
Electrical and Computer Engineering
University of Michigan-Dearborn
Dearborn, United States of America
jrkwon@umich.edu

Haewoon Nam
Division of Electrical Engineering
Hanyang University
Ansan, South Korea
hnam@hanyang.ac.kr

***Abstract*—The ability to accurately predict future road conditions is essential for the advancement of autonomous driving systems. This study introduces a BiLSTM-based VAE-GAN framework that leverages both temporal and spatial information to generate high-quality future road images. The proposed architecture combines the reconstruction capabilities of Variational Auto-Encoders (VAEs) with the adversarial training of Generative Adversarial Networks (GANs), while incorporating Bidirectional Long Short-Term Memory (BiLSTM) to effectively capture temporal dependencies in sequential driving data. To train the model, diverse datasets were collected from the CARLA simulation environment, encompassing various road conditions and vehicle states. The training process minimizes reconstruction loss, KL divergence, and adversarial loss, enabling the generation of visually consistent and semantically accurate future road images. Quantitative evaluations using PSNR and MSE metrics demonstrate the model's ability to outperform conventional VAE-based approaches, achieving high structural similarity and low reconstruction errors. The results highlight the potential of the proposed framework to enhance decision-making and lane-keeping performance in autonomous vehicles. By predicting future road states with high fidelity, the BiLSTM-based VAE-GAN framework lays the groundwork for integrating generative models into real-world autonomous driving applications, contributing to safer and more reliable driving systems.**

***Index Terms*—Deep learning, Autonomous driving, Generative adversarial networks**

I. Introduction

The ability to anticipate and predict future road conditions is a crucial component in the development of autonomous driving systems. Generating accurate and high-quality future road images based on current environmental observations and vehicle control parameters is essential for reliable decision-making, lane-keeping, and collision avoidance [1]. Variational Auto-Encoders (VAEs) have been widely employed for such generative tasks due to their capability to encode complex data distributions into a meaningful latent space [2]. However, conventional VAE-based models often struggle to capture the temporal dependencies inherent in sequential driving scenarios, leading to a lack of detail and consistency in generated images.

To address these limitations, this study introduces a Bidirectional Long Short-Term Memory (BiLSTM)-based VAE-GAN framework that combines the reconstruction capabilities of Variational Auto-Encoders (VAEs) with the adversarial training approach of Generative Adversarial Networks (GANs). The incorporation of BiLSTM into the VAE-GAN structure allows the model to effectively capture temporal dependencies and spatial patterns in sequential road data. Inspired by previous works leveraging VAE-GANs for image generation tasks [3] and extending LSTM-based generative models [4], this study builds upon these advancements to address the unique challenges of sequential data in autonomous driving scenarios. BiLSTMs are particularly suited for this task as they process information in both forward and backward directions, enabling a more comprehensive understanding of temporal context compared to unidirectional LSTM networks [5]. This bidirectional approach ensures that the model learns not only future dependencies but also the influence of past events on current and future road conditions.

Using a simulation environment provided by CARLA [6], we generated diverse datasets under various road conditions and vehicle states. These datasets encompass a wide range of scenarios, including different steering angles, vehicle speeds, and road geometries, which are essential for training a robust model. By integrating temporal and spatial information into the latent space, the proposed BiLSTM-based VAE-GAN effectively predicts future road trajectories with enhanced structural and semantic fidelity.

The proposed BiLSTM-based VAE-GAN model addresses the quality and consistency issues of traditional VAE-generated images while providing a robust tool for predicting steering angles and improving lane-keeping performance in autonomous vehicles. This research not only highlights the advantages of incorporating BiLSTM for sequential data modeling but also lays the groundwork for integrating advanced generative models into practical autonomous driving applications.

II. Methodology

In this section, we describe the setup of the CARLA simulation environment for experiments, the network structure of the BiLSTM-based VAE-GAN model, and its training process. The CARLA environment is used to acquire realistic data, and the BiLSTM-based VAE-GAN structure is employed to train the model to generate future road images that can be utilized in autonomous driving systems.

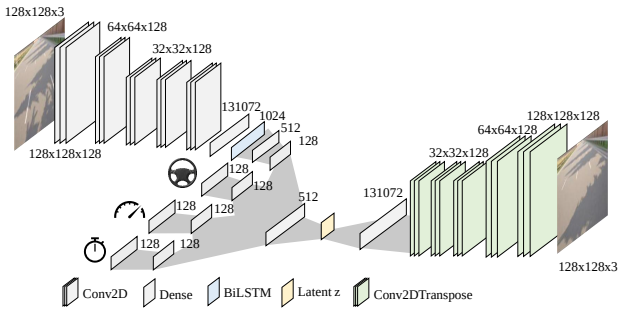

Fig. 1. The BiLSTM-based VAE-GAN generator is designed to take road images and vehicle control values as inputs and generate future road images.

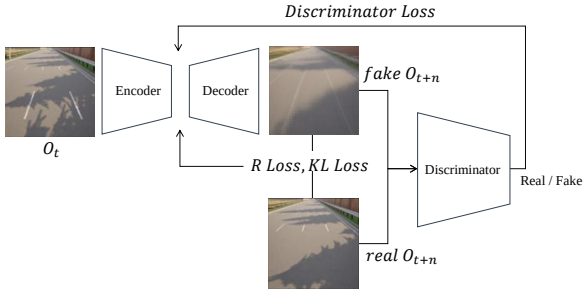

Fig. 2. The architecture of the proposed BiLSTM-based VAE-GAN model. The encoder compresses the input image O_t into a latent space, while the decoder generates a future road image fake O_{t+n}. The discriminator evaluates the quality of the generated image by comparing it to the real future image real O_{t+n}. The overall training optimizes reconstruction loss (R Loss), KL divergence (KL Loss), and discriminator loss to improve the realism of the generated images.

A. Network Architecture

The entire network comprises a generator, consisting of an encoder and decoder based on the VAE architecture, and a discriminator for distinguishing between real and generated images. The generator architecture, as depicted in Fig. 1, includes an encoder that compresses the input into a latent representation and a decoder that reconstructs the future road image from the latent space. The discriminator, shown in Fig. 2, is trained to differentiate between real images and those generated by the model, thereby improving the quality and realism of the generated outputs.

- Encoder: Compresses input data, including images (128 × 128), steering angle, velocity, and time, into a latent space. A BiLSTM is employed to effectively learn spatiotemporal features.
- Decoder: Performs an inverse transformation from the latent space to generate future road images. ConvTranspose2D layers are utilized to reconstruct high-resolution images.
- Discriminator: As part of the GAN structure, it distinguishes between generated and real images, improving the quality of generated images by aligning their distributions.

B. Training Configuration

The training process aimed to simultaneously minimize the losses of the VAE and GAN. The total loss function of the BiLSTM-based VAE-GAN model is a weighted sum of three components: reconstruction loss (R_{Loss}), KL divergence (KL_{Loss}), and discriminator loss (D_{Loss}). These components are combined as:

$$\mathcal{L} = \alpha R_{\text{Loss}} + \beta KL_{\text{Loss}} + \gamma D_{\text{Loss}} \tag{1}$$

where α, β, and γ are weighting factors for each component. For this study, the weighting factors were set as:

- $\alpha = H \times W \times 100$, where $H = W = 128$, corresponding to the image resolution.
- $\beta = 1$, ensuring stable latent space regularization.
- $\gamma = 100$, emphasizing the importance of adversarial training for improving image quality.

The models were trained for 500 epochs with a batch size of 6, and the Adam optimizer was used with a learning rate of 5×10^{-5}. These configurations ensured a fair and controlled setup for comparing the proposed BiLSTM-based VAE-GAN with the conventional VAE model.

1) Reconstruction Loss (R_{Loss}): This loss minimizes the difference between the input and the reconstructed data, ensuring the generated future road images are accurate and realistic. The reconstruction loss is computed as:

$$R_{\text{Loss}} = \frac{1}{N} \sum_{i=1}^{N} \|\text{real } O_{t+n}^{(i)} - \text{fake } O_{t+n}^{(i)}\|^2 \tag{2}$$

2) KL Divergence (KL_{Loss}): This term regularizes the latent space by minimizing the divergence between the latent variable distribution and a standard Gaussian distribution:

$$KL_{\text{Loss}} = -\frac{1}{2} \sum_{j=1}^{d} \left(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2\right) \tag{3}$$

where μ_j and σ_j are the mean and standard deviation of the latent variables, respectively.

3) Discriminator Loss (D_{Loss}): As part of the GAN structure, this loss improves the quality of the generated images by distinguishing between real and fake images:

$$D_{\text{Loss}} = -\mathbb{E}[\log D(\text{real } O_{t+n})] - \mathbb{E}[\log(1 - D(\text{fake } O_{t+n}))] \tag{4}$$

where $D(\cdot)$ denotes the discriminator's output.

C. Experiment Setup

Data was collected using the CARLA simulator in various road environments for training purposes. The dataset consists of the current road image, steering angle, speed, time, and future road image captured during driving. A total of 31K data samples were used for training.

TABLE I
COMPARISON OF PSNR AND MSE PERFORMANCE BETWEEN VAE AND BILSTM-BASED VAE-GAN

Model	PSNR	MSE
VAE	32.96	0.00060
BiLSTM-based VAE-GAN	**33.71**	**0.00049**

III. EXPERIMENTAL RESULTS

This section presents the evaluation of the proposed BiLSTM-based VAE-GAN model through quantitative metrics and a comparative analysis with a conventional VAE model. By analyzing the generated future road images, we assess the effectiveness of the model in terms of reconstruction accuracy and pixel-wise similarity.

A. Evaluation Metrics

To evaluate the performance of the proposed BiLSTM-based VAE-GAN model, we compared the generated future road images with their corresponding ground truth images using Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE) as evaluation metrics. These metrics were chosen to quantitatively assess both the perceptual similarity and pixel-wise accuracy between the generated and original images.

B. Quantitative Evaluation

The PSNR scores and MSE values, as summarized in Table I, demonstrate the effectiveness of the proposed BiLSTM-based VAE-GAN model. The VAE-GAN achieved an average PSNR of 33.71 and an MSE of 0.00049, outperforming the conventional VAE model, which achieved a PSNR of 32.96 and an MSE of 0.00060. These results highlight the ability of the VAE-GAN to generate future road images with higher fidelity and reduced pixel-wise reconstruction error compared to the VAE.

The improvement in PSNR indicates that the VAE-GAN effectively reduces noise and distortion in the generated images, resulting in a closer resemblance to the ground truth images. The lower MSE further confirms the model's capability to preserve fine-grained details and structural consistency in the predicted images.

The experimental results validate that the proposed BiLSTM-based VAE-GAN model outperforms the traditional VAE-based approach in both PSNR and MSE metrics. The incorporation of adversarial training using a discriminator enhances the quality of the generated images, making the VAE-GAN a robust solution for autonomous driving tasks. These findings suggest that the proposed model is well-suited for future road image prediction, providing critical input for reliable decision-making in autonomous systems.

IV. CONCLUSION

In this study, we proposed a BiLSTM-based VAE-GAN framework for generating high-quality future road images based on current road conditions and vehicle control inputs. By incorporating adversarial training, the proposed model effectively addresses the limitations of conventional VAE models, particularly the lack of detail in generated images.

Furthermore, the ability of the model to generate diverse future scenarios by varying control inputs highlights its potential for real-world applications, particularly in autonomous driving systems. By predicting future road states with high fidelity, the model can contribute to reliable lane-keeping and collision-avoidance systems, enhancing overall driving safety.

Future work will focus on further optimizing the model to handle more complex road conditions and integrating the system with reinforcement learning frameworks for end-to-end autonomous driving solutions.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (RS-2024-00428465) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] Y. Zhou, H. Dong, and A. El Saddik, "Deep learning in next-frame prediction: A benchmark review," *IEEE Access*, vol. 8, pp. 69 273–69 283, 2020.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.

[3] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1558–1566.

[4] Z. Niu, K. Yu, and X. Wu, "LSTM-based VAE-GAN for time-series anomaly detection," *Sensors*, vol. 20, no. 13, p. 3738, Jul. 2020.

[5] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5, pp. 602–610, 2005.

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