https://doi.org/10.7840/kics.2025.50.10.1532

ViT 기반 미래 도로 이미지 예측: VLM을 활용한 평가

김 동 현•, 권 재 락*, 남 해 운°

ViT-Based Future Road Image Prediction: Evaluation via VLM

Donghyun Kim•, Jaerock Kwon*, Haewoon Nam°

요 약

본 논문은 미래 주행 상황을 효과적으로 예측하기 위해 Vision Transformer (ViT) 기반 미래 도로 이미지 예측 모델을 제안한다. 제안하는 ViT 모델 구조는 입력 이미지를 패치 단위로 처리하고, 어텐션 메커니즘을 통해 전역적인 시각 정보를 효율적으로 학습할 뿐만 아니라, 제어 입력과의 통합적 처리로 시각-제어 간 상관관계를 효과적으로 반영할 수 있는 장점을 가진다. 생성된 이미지의 화질 및 설명 유사도 측면에서 성능을 비교한 결과, 제안하는 모델은 기준 모델보다 선명한 이미지를 생성하였으며, Vision-Language Model (VLM)을 활용한 설명 평가에서도 높은 의미 유사도를 나타냈다. 이는 ViT 구조가 미래 예측에 효과적일 뿐 아니라, 설명 정보를 활용한 자율주행 제어 연계에도 유용함을 시사한다.

Key Words : Autonomous Driving, Vision-Language Model, Semantic Evaluation, Vision Transformer

ABSTRACT

This paper proposes a Vision Transformer (ViT)-based model for predicting future driving scenes. The proposed ViT architecture processes input images as patches and leverages the attention mechanism to efficiently learn global visual information, while also integrating control inputs to effectively capture correlations between visual context and driving actions. Experimental results show that the ViT-based model generates sharper images than the baseline and achieves higher semantic similarity in explanation evaluations using a Vision-Language Model (VLM). These results suggest that the ViT architecture is effective not only for future prediction but also for explainable autonomous driving control.

Ⅰ. 서 론

자율주행 시스템에서는 주행 안정성과 판단력을 높이기 위해 미래 상황을 예측하는 내부 모델 기반 구조가 주목받고 있으며, 미래 예측 기반의 자율주행 연구에서는 시계열 이미지와 제어 입력을 활용하여 미래 프레임이나 제어신호를 생성하는 모델 구조가 활발히 연구되고 있다[1-3]. 이는 이미지와 제어 정보를 통합하여, 잠재 공간에서 미래 상태 모델링을 통해 미래 프레임을 효과적으로 생성하고 제어 결정에 대한 해석 가능성을 제공한다. 선행 연구에서 우리는 전방 도로 이미지와 차량 제어 데이터를 통합하여 미래 주행 상황을 예측하기 위한 Bidirectional Long Short-Term Memory (BiLSTM), Variational Autoencoder (VAE), Generative Adversarial Network (GAN) 기반 이미지 생성 모델을 제안한 바 있다[4]. 본 논문에서는 미래 도로 이미지 예측 모델을 Vision Transformer (ViT)기반 구조로 확장하여 미래 예측 성능과 해석 가능성을 비교 분석한다. 또한, Vision-Language Model (VLM)과 이미지 캡셔닝 모델을 활용하여 예측된 이미지에 대한 설명 정보를 생성하고, 이를 정량적 및 정성적 관점에서 평가함으로써 설명 기반 자율주행 제어 연계 가능성을 탐색한다.

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업(IITP-2025-RS-2023-00258639)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2022R1A2C1011862)

• First Author : (ORCID:0000-0002-6626-3706) Hanyang University Department of Electrical and Electronic Engineering, kissw@hanyang.ac.kr, 학생(박사), 학생회원

° Corresponding Author : (ORCID:0000-0001-9847-7023) Hanyang University Department of Electrical Engineering, hnam@hanyang.ac.kr, 정교수, 정회원

* (ORCID:0000-0002-5687-6998) University of Michigan-Dearborn Electrical and Computer Engineering, jrkwon@umich.edu, 부교수

논문번호 : 202505-116-A-LU, Received May 16, 2025; Revised May 27, 2025; Accepted May 27, 2025

Ⅱ. 본 론

본 논문에서는 미래 도로 이미지를 예측하기 위한 ViT 기반 모델을 제안하고 기준 모델과 이미지 품질 및 설명 평가 비교를 진행하였다.

2.1 미래 예측 모델 구조

2.1.1 제안 모델

제안하는 ViT 기반 모델은 그림 1과 같이 6개의 레이어로 구성된 인코더 및 디코더 스택으로 구성되어 있으며, 각 레이어의 패치 임베딩 벡터 길이는 256, 어텐션 헤드는 8로 설정하였다. 입력 이미지는 64x64 해상도의 RGB 이미지로, 8x8 크기의 패치 단위로 분할되어 총 64개의 시각 토큰으로 변환된다. 제어 입력은 두 개의 선형 계층과 ReLU 활성화를 포함한 인코더를 통해 시각 임베딩과 동일한 차원으로 정규화되며, 이후 64개의 패치 수에 맞춰 반복되어 시각 임베딩과 일대일 대응되도록 구성된다. 이후 두 임베딩을 결합하여 인코더에 입력하고, 인코더는 이미지의 전역적인 시각적 패치 간 관계를 통합한다. 디코더는 학습 가능한 쿼리 토큰을 입력으로 받아, 인코더의 출력을 참조하며 미래 프레임의 시각 정보를 생성한다. 이러한 구조는 이미지 내 패치 간 전역적 시각 정보를 병렬적으로 학습할 수 있어, 예측 정확도와 학습 효율 측면 모두에서 이점을 가진다.

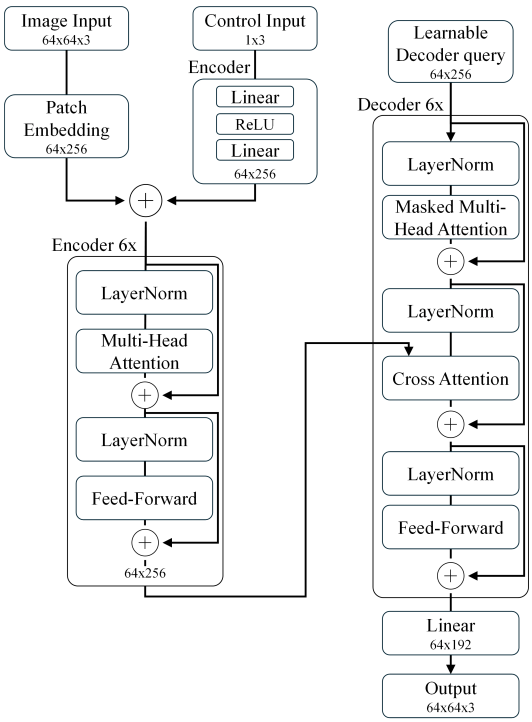

그림 1. ViT 기반 미래 도로 이미지 예측 모델 구조
Fig. 1. ViT-based future road image prediction model architecture

2.1.2 기준 모델

기준 모델은 BiLSTM, VAE, GAN을 결합한 형태이다[4]. BiLSTM은 과거 시점의 정보를 양방향으로 처리하여 시간적 맥락을 반영할 수 있으며, 일부 연구에서는 Convolutional LSTM 대비 시계열 예측 성능이 우수하다는 연구 결과를 보고한 바 있다[5]. 이 구조는 단일 LSTM, VAE, 또는 GAN에 비해 시간적 정보의 요약, 불확실성 표현, 그리고 고해상도 이미지 복원의 측면에서 상호 보완적인 강점을 가진다. 한편, 최근 주목받는 디퓨전 기반 생성 모델은 고품질 이미지 생성에는 효과적이지만, 모델의 복잡성으로 인해 학습이 어렵고 추론 속도가 느리다는 단점이 있어 실시간 제어가 필요한 자율주행 모델에 적용하기에는 어려움이 있다. 반면에 기준 모델 구조는 비교적 가벼운 연산 구조를 바탕으로 빠른 추론이 가능하여 자율주행과 같은 실시간 예측이 필요한 작업에 적합한 특성을 가진다.

2.2 데이터셋 및 실험 환경

본 연구에서는 미래 도로 이미지 예측을 위한 학습 및 평가를 위해 CARLA 시뮬레이터의 Town02 에서 차량을 직접 주행하여 데이터를 수집하였다. 차량의 주행 시점에서 촬영된 전방 도로 이미지와 함께 차량 제어 데이터(조향각, 속도, 시간)를 동기화하여 저장하였다. 입력으로 사용되는 현재 이미지와 제어 값을 기반으로, 0.1~0.6초 후의 미래 이미지를 예측하도록 학습되었으며, 전체 데이터는 73,000개로 구성되어 학습에는 전체 데이터의 85%를 사용하고, 나머지 15%는 테스트 및 성능 비교에 활용하였다. 또한, 생성된 미래 이미지에 대한 해석 가능성을 평가하기 위해, 각 이미지에 대응하는 Ground truth (GT) 설명을 수작업으로 라벨링 하였다. 라벨은 차량의 차선 위치 상태(중앙, 왼쪽 치우침, 탈선 등)와 도로 유형(직선, 곡선, 교차로)으로 조합된 문장이며, 1,000개의 데이터로 구성되어 있다. 해당 데이터를 기반으로 VLM 및 이미지 캡셔닝 모델을 파인튜닝하였다.

2.3 미래 예측 결과 평가 방법

성능을 평가하기 위해 생성된 미래 이미지의 두 가지 관점에서 분석을 수행하였다. 첫 번째는 생성된 이미지와 실제 이미지 간의 유사도를 측정하는 이미지 품질 평가이며, 두 번째는 생성된 이미지를 VLM에 입력하

였을 때 출력된 자연어 설명이 GT와 얼마나 유사한지를 측정하는 설명 평가이다. 이미지 품질 평가는 픽셀 수준의 유사도와 구조적 일관성을 수치화하는 대표적인 정량 지표인 Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM)을 기준으로 수행되며, 이 두 지표를 통해 기준 모델과 ViT 기반 모델의 이미지 생성 성능을 비교하였다. 설명 평가에서는 VLM 모델인 BLIP[6]과 ViT 기반의 경량화된 이미지 캡셔닝 모델 (TinyIC)을 파인튜닝하여 사용하였다. 이후, 각 내부 모델로부터 생성된 예측 이미지에 대해 자연어 설명을 생성한 후, GT 설명과의 코사인 유사도를 산출하여 평가하였다.

2.4 실험 결과 및 분석

성능 평가 결과는 표 1과 같이, ViT 기반 모델이 기준 모델에 비해 PSNR 및 SSIM 지표에서 더 우수한 수치를 보였으며, 그림 2의 시각적 비교에서도 선명도와 구조적 정밀도 측면에서 더 뛰어난 이미지 생성 성능을 보였다. 또한 자연어 설명 평가 결과에서도 제안하는 모델이 예측한 이미지를 통해 생성된 설명이 GT 설명과 더 높은 유사도를 나타내었다. 그림 3과 같이, 도로에서 차량이 차선 중앙에 있는 경우에도 기준 모델이 생성한 이미지를 VLM에 입력하였을 때, "vehicle deviated far right" 또는 "vehicle collided or off the road" 등과 같이 잘못된 설명을 생성했지만, 제안한 모델이 생성한 이미지를 사용하였을 땐 차량 위치와 도로 형태

표 1. 생성된 미래 도로 이미지 품질 및 설명 유사도 비교
Table 1. Comparison of generated future road image quality and explanation similarity

Model	PSNR (dB)	SSIM	Cosine Similarity	
			BLIP	TinyIC
Baseline	26.43	0.7785	0.928	0.933
ViT	**28.05**	**0.8418**	**0.93**	**0.947**

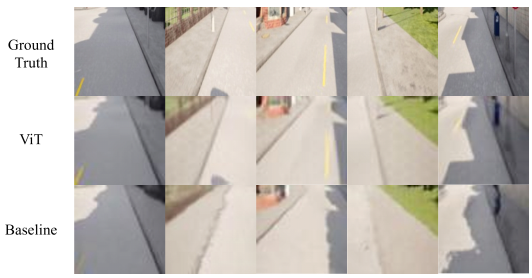

그림 2. 미래 도로 이미지 예측 결과의 시각적 비교
Fig. 2. Qualitative comparison of future road image prediction results

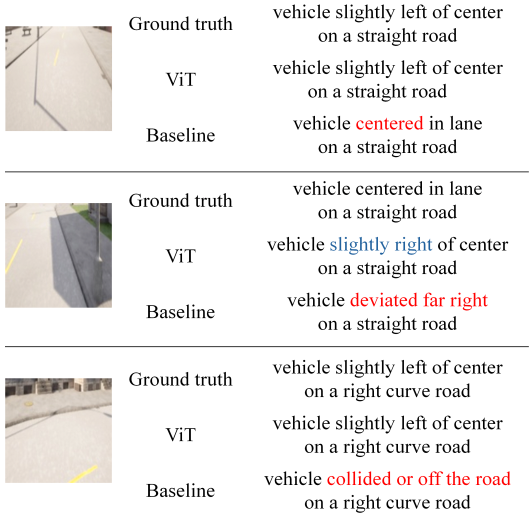

그림 3. 예측 이미지의 자연어 설명 비교
Fig. 3. Comparison of explanation from predicted images

를 반영한 설명을 제공하였다. 이러한 결과는 기준 모델이 VAE 구조의 본질적인 한계로 인해 출력 이미지가 흐릿하게 생성되기 때문이다. VAE는 잠재 공간에서의 확률적 샘플링과 복원 과정에서 평균적인 형태를 생성하려는 특성에 의해 세부 경계나 객체 형태를 정확하게 표현하기 어렵다. 이에 따라 VLM이 이미지 내 객체나 도로 구조를 명확하게 인식하지 못해, 설명 평가 성능이 저하되는 원인이 된다. 반면 제안된 모델은 입력 이미지를 패치 단위로 분할하고 어텐션을 통해 전역적인 시각적 관계를 효과적으로 학습함으로써, 전체 장면에서 도로 경계, 차량 위치, 구조물 등의 요소를 정밀하게 반영할 수 있다.

Ⅲ. 결 론

본 논문에서는 미래 도로 이미지 예측을 위해 ViT 기반 모델을 제안하고 기준 모델과 이미지의 품질 및 의미 기반 설명 성능을 비교하였다. 실험 결과, ViT 기반 모델은 기준 모델보다 더 높은 이미지 품질을 보였으며, 설명 생성에서도 GT와 더 높은 유사성을 나타냈다. 이러한 결과는 ViT 구조가 어텐션 메커니즘을 통해 전역적인 시각 정보를 효율적으로 학습하며, 시각-제어 간 상관관계를 효과적으로 반영함으로써 기존 구조 대비 우수한 이미지 표현 성능을 제공함을 시사한다. 향후 연구에서는 VLM을 주행 데이터에 정밀하게 최적화하고, 생성된 설명 정보를 활용하여 제어 판단까지 연계하는 구조로 확장하는 것을 목표로 한다.

References

[1] X. Wang, et al., “Drivedreamer: Towards real-world-drive world models for autonomous driving,” in *Proc. ECCV*, pp. 55-72, 2025.

[2] D. Jeong and B. Jeon, “Development of a near-future vehicle speed prediction technology using generative deep learning model,” *Trans. KSAE*, vol. 29, no. 7, pp. 629-637, Jul. 2021.

[3] F. Jia, et al., “Adriver-i: A general world model for autonomous driving,” *arXiv preprint arXiv:2311.13549*, 2023.

[4] D. Kim, et al., “Bilstm-based VAE-GAN for predicting future road states in autonomous driving,” in *Proc. ICAIIC*, pp. 905-907, Fukuoka, Japan, Feb. 2025.

[5] Arwansyah, et al., “Deep sequence models for time series data: A comparative study and parameter fine tuning approach,” in *Int. Conf. EECSI*, pp. 703-709, 2024.

[6] J. Li, et al., “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proc. ICML*, pp. 12888-12900, 2022.