# 언어 기반 가이드를 통한 제한된 시야 환경에서의 도로 주행 환경 인식 성능 향상 연구

신연우 [1], 전은솜 [2]
[1] 서울과학기술대학교 글로벌테크노경영학과 학부생
[2] 서울과학기술대학교 컴퓨터공학과 교수

syw3159@seoultech.ac.kr, ejeon6@seoultech.ac.kr

# Enhancing Driving Scene Perception in Limited Field of View via Language Guidance

Yoen Woo Shin [1], Eun Som Jeon [2]
[1]Dept. of GTM, SeoulTech
[2]Dept. of Computer Science and Engineering, SeoulTech

**요 약**

자율주행 시스템의 도로 주행 환경 인식 모델은 사각지대 등 시야가 제한된 상황에서 성능이 저하된다. 본 연구는 제한된 시야 환경의 이미지 의미론적 분석에서 다양한 유형의 언어 가이던스의 효과성을 분석한다.

## 1. 서론

도로환경 인식에서 사각지대 등 제한된 시야 상황은 빈번히 발생하지만, 제약적인 정보로부터 효과적인 인식 성능을 취득하는 것에는 어려움이 있다. 본 연구는 언어적 가이던스를 의미론적 분할 모델에 함께 활용하여 강인한 모델 설계를 목표로 한다. 특히, LLM 을 통해 취득한 언어를 유형화하고 제한된 시야 환경에서 영상 인식 성능 향상에 가장 효과적인 언어 정보 유형을 분석한다.

## 2. 실험

### 2.1 Framework

Yuxiang Ji et al. [1]가 제안한 프레임워크를 기반으로 영상 인식에서의 언어 정보의 유형별 효과성에 대해 분석한다 (그림 1). 주석 $Y$ 에서 클래스별 마스크 $M$ 와 텍스트 프롬프트 $C$ 를 구성하여 조건부 DIFF 모듈 $F_{condiff}$ 에서 잠재변수 $I$ 를 예측한다. 이때 $\phi$ 는 사전학습 된 확산모델이다. 다중 시간-스케일 잠재변수를 학습 가능한 성곱층으로 집계해 분할헤드 $D$ 에 입력하며, 다음 수식을 통해 합성곱층과 분할헤드를 학습한다: $L_{condit} = CD(D(F_{condiff}(X, M, C)), Y)$
비조건부 분기 $F_{uncondiff}$ 를 도입해 추론시 프롬프트 독립성을 유지한다:

$$L_{consis} = \|D(F_{condiff}(X, M, C), D(F_{uncondiff}(X)))\|$$

총 손실은 $L_{total} = \lambda_1 L_{condit} + \lambda_2 L_{consis}$ 로 정의되며, 분기간 가중치 공유로 지식을 증류시킨다.

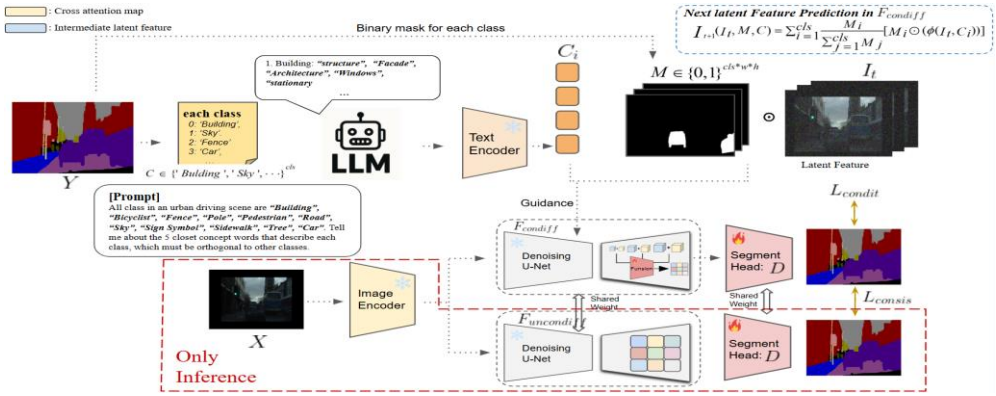
### 2.2 LLM for Prompt

GPT-4 [2]를 활용해 각 클래스마다 4 가지 유형의 프롬프트를 생성하였다. 프롬프트 유형은 아래와 같으며, 생성 결과는 <표 2>와 같다.

- Label: 클래스명 그대로 사용
- Sentence: 클래스에 대한 일반적 문장 표현
- Bag of words: 각 클래스를 표현하는 일반적 단어 표현 4 개와 클래스명 1 개
- Orthogonal concept: 오직 각 클래스만을 표현하는 고유 단어 표현 4 개와 클래스명 1 개

### 2.3 실험설정

NVIDIA GeForce RTX4090(24GB)과 256GB RAM 환경에서 CamVid 데이터셋 [3]으로 실험하였다. CamVid 는 11 개 클래스와 701 개 이미지로 구성된다. 512x384 의 해상도에서 2-fold [4] 교차 검증을 적용한다. Stable diffusion v2-base [5]를 backbone, DAFormer [6]를 head 로 사용하고 배치크기 1, 30K iteration, AdamW 옵티마이저로 학습하였다.



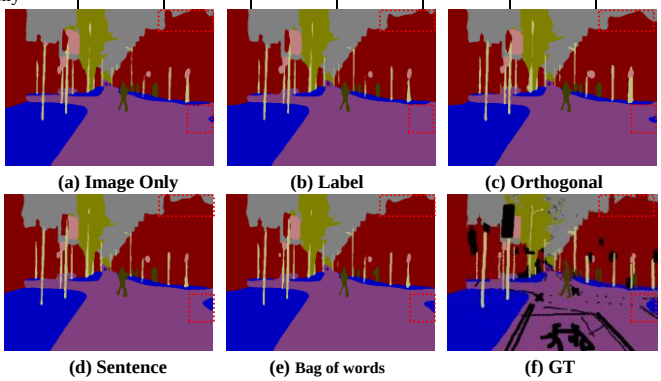

(그림 1) 전체 프레임워크

<표 1> 유형별 프롬프트 생성 결과

| Label | Sentence | Bag of words | Orthogonal Concept |
|---|---|---|---|
| 'Sky' | 'The area above the earth, usually blue during the day with clouds. ' | 'Sky', 'Cloud', 'Blue', 'Sun', 'Daylight' | 'Sky', 'Atmosphere', 'Horizon', 'Clouds', 'Airspace' |
| 'Building' | 'A tall structure made of concrete and glass that houses offices and apartments.' | 'Building', 'Concrete', 'Window', 'Foundation', 'Architecture' | 'Building', 'Facade', 'Structure, 'Windows, 'Architecture' |
| 'Pole' | 'A vertical post used to support electrical wires or street lamps. ' | 'Pole', 'Metal', 'Wire', 'Cable', 'Support' | 'Pole', 'Slender', 'Vertical', 'Cable', 'Metallic' |
| 'Road' | 'A paved surface for vehicles to drive on, often marked with lanes ' | 'Road', 'Drive', 'Path', 'Route', 'Street' | 'Road', 'Asphalt', 'Lane', 'Driveway', 'Trafficway' |
| 'Sidewalk' | 'A paved path beside the road for people to walk safely. ' | 'Sidewalk', 'Pavement', 'Path', 'Walkway', 'Walk' | 'Sidewalk', 'Curb', 'Walkway', 'Pedway', 'Paving' |
| 'Tree' | 'A tall plant with a trunk, branches, and leaves providing shade.' | 'Tree', 'Branch', 'Leaf', 'Forest', 'Wood' | 'Tree', 'Foliage', 'Trunk', 'Branches', 'Canopy' |
| 'Signsymbol' | 'A traffic sign or symbol used to guide or warn drivers and pedes-trian.' | 'Signsymbol', 'Traffic', 'Board', 'Sign', 'Marker' | 'Signsymbol', 'Traffic sign', 'Pictogram', 'Instruction', 'Regulation' |
| 'Fence' | 'A wooden or metal barrier that encloses a yard or property.' | 'Fence', 'Wood', 'Barrier', 'Boundary', 'Post' | 'Fence', 'Pickets', 'Barrier', 'Boundary', 'Perimeter' |
| 'Car' | 'A motor vehicle with four wheels used for transportation.' | 'Car', 'Wheel', 'Door', 'Engine', 'Tire' | 'Car', 'Vehicle', 'Windshield', 'Engine', 'Parking' |
| 'Pedestrian' | 'A person walking on the sidewalk or crossing the street.' | 'Pedestrian', 'Walk', 'Food', 'Step', 'Crossing' | 'Pedestrian', 'Crosswalk', 'Footsteps', 'Walker', 'Striding' |
| 'Bicyclist' | 'A person wearing a helmet riding a bicycle on the street.' | 'Bicyclist', 'Rider', 'Wheels', 'Cycling', 'Path' | 'Bicyclist', 'Helmet', 'Pedals', 'Wheels', 'Cycling' |

<표 2> 프롬프트 유형별 성능지표

| Prompt | Sidewalk | Signsy-mbol | Bicyclist | PA | mPA | FW IoU | mIoU |
|---|---|---|---|---|---|---|---|
| Label | 85.70 | 35.63 | 79.71 | 93.21 | 78.39 | 87.63 | 71.60 |
| Sentence | 86.51 | 38.76 | 78.59 | **93.23** | 78.76 | 87.69 | **71.70** |
| B of Words | 86.83 | 36.49 | 78.17 | 93.32 | 78.16 | **87.83** | 71.49 |
| Orthogonal | 86.10 | 38.40 | 72.96 | 93.21 | **78.86** | 87.66 | 71.04 |
| Image Only | 85.87 | 35.65 | 75.91 | 93.17 | 78.30 | 87.56 | 71.03 |



**(a) Image Only** **(b) Label** **(c) Orthogonal**

**(d) Sentence** **(e) Bag of words** **(f) GT**

(그림 2) 프롬프트 유형별 예측 시각화

## 3. 결과 및 결론

실험 결과(<표 1>, (그림 2)), 대개 문장 유형이 가장 뛰어난 성능을 보였으며, Orthogonal 유형은 이미지만 사용한 경우와 유사한 성능을 나타냈다. (그림 2)의 도보영역과 같이 가려진 부분에서 Sentence 와 Bag of words 에서 더 효과적임을 확인했다.

Sentence 의 다양한 객체와 술어가 여러 클래스에 어텐션을 형성하여 편향된 어텐션이 발생한다. 이러한 편향된 어텐션이 객체 간 관계 정보를 포함하고, 노이즈 예측 시에 반영되어 제한된 시각 정보 상에서 더 높은 성능을 달성한 것으로 해석된다. 반면 Orthogonal 은 각 클래스에만 집중된 어텐션을 보여 클래스 간 관계성을 활용하지 못하므로, 제한된 시각 환경에서 성능저하가 발생한것으로 분석된다.

## 참고문헌


[1] Yuxiang Ji, Boyong He, Chenyuan Qu, Zhuoyue Tan, Chuan Qin, and Liaoni Wu, "Diffusion Features to Bridge Domain Gap for Semantic Segmentation," in *ICASSP*, Hyderabad, India, 2025. pp. 1-5

[2] OpenAI, "GPT-4 Technical Report," *arXiv:2303.08774*, 2023.

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla "Segmentation and recognition using structure from motion point clouds," *in ECCV*, Marseille, France, 2008, pp. 44–57.

[4] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, Montreal, Canada, 1995, pp. 1137-1145.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[6] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," *in CVPR*, 2022, pp. 9924–9935.