Journal of KIIT. Vol. 24, No. 7, pp. 149-159, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2026.24.7.149

# 다단계 프롬프트와 그리드 오버레이를 활용한 VLM 기반 밀집 차량 검출

최 권 택*

# Dense Vehicle Detection using Multi-Stage Prompting and Grid Overlay with Vision-Language Models

Kwon-Taeg Choi*


## 요 약

최근 Vision Language Model은 추가 학습 없이 텍스트 프롬프트만으로 객체 검출이 가능하여 높은 유연성을 제공한다. 그러나 주차장과 같이 동일 클래스 객체가 밀집된 환경에서는 시각 인코더의 공간 정보 압축으로 인해 인접 객체가 하나의 바운딩 박스로 병합되거나 일부 객체가 누락되는 과소 검출 문제가 발생한다. 본 연구는 이러한 문제를 해결하기 위해 모델 수정이나 재학습 없이, 입력 이미지의 장면 특성을 분석하여 밀집 객체의 개별 검출을 유도하는 다단계 프롬프트의 자동 생성 기법과, 격자 형태의 가상 라인을 원본 이미지에 합성하여 밀집 영역을 시각적으로 분할하는 그리드 오버레이 기법을 제안한다. 3개 공개 데이터셋에 대한 실험 결과, 제안 방법은 baseline 대비 재현율 25.3%, F1-score 17.9% 향상시켰으며, 특히 밀집도가 높은 데이터셋일수록 개선 폭이 큰 것으로 나타나 밀집 차량 검출에서의 효과를 입증하였다.

## Abstract

Recent Vision-Language Models (VLMs) offer high flexibility by enabling object detection through text prompts alone without additional training. However, in dense environments such as parking lots, where objects of the same class are closely packed, VLMs suffer from under-detection in which adjacent objects are merged into a single bounding box or some objects are missed entirely, due to spatial information compression in the visual encoder. To address this problem, this study proposes a multi-stage prompting method that analyzes scene characteristics to automatically generate prompts guiding the individual detection of densely packed objects, together with a grid overlay method that composites virtual grid lines onto the original image to visually partition dense regions, without any model modification or additional training. Experiments on three public datasets demonstrate that the proposed method improves recall by 25.3 percentage points and F1-score by 17.9 percentage points over the baseline, with larger gains observed on more densely packed datasets, confirming its effectiveness in dense vehicle detection.


Keywords

vision-language model, dense object detection, car detection, prompt engineering

* 강남대학교 ICT융합공학부 교수
- ORCID: http://orcid.org/0000-0001-5331-321X

· Received: Apr. 07, 2026, Revised: May 13, 2026, Accepted: May 16, 2026
· Corresponding Author: Kwon-Taeg Choi
Dept. of ICT Convergence Engineering, Kangnam University, Yongin, Korea
Tel.: +82-31-280-3660, Email: kwontaeg.choi@kangnam.ac.kr

## Ⅰ. 서 론

객체 검출은 컴퓨터 비전 분야에서 이미지 내 특정 클래스의 객체 위치와 범위를 식별하는 핵심 문제이다. 이 분야는 CNN(Convolutional Neural Network) 기반 모델의 발전을 통해 빠르게 성장해 왔으며, Regions CNN 계열의 2단계 검출기로부터 시작하여 YOLO(You Only Look Once)와 SSD(Single Shot MultiBox Detector) 등의 1단계 검출 방식, 그리고 DETR(DEtection TRansformer)과 같은 Transformer 기반 검출 방식으로 발전해왔다[1]-[4]. 특히 차량 검출은 자율주행, 교통 모니터링, 주차장 관리 등 다양한 응용 분야에서 중요한 역할을 한다. 이 중에서도 주차장 영상은 동일 클래스 객체가 높은 밀도로 반복적으로 등장하고 객체 간 간격이 매우 좁은 특성을 가지며, 정밀한 경계 구분이 요구되는 대표적인 밀집 객체 검출 환경에 해당한다.

그러나 전통적인 객체 검출 방식은 대규모 라벨링 데이터를 기반으로 학습되어야 하며, 새로운 도메인에 적용하기 위해서는 추가적인 데이터 수집과 재학습이 필요하다는 한계를 가진다. 특히 주차장과 같이 카메라 시점과 환경 조건이 다양한 경우, 데이터 구축 비용이 증가하는 문제가 발생할 수 있다.

최근 대규모 사전학습 기반의 VLM(Vision Language Model)이 등장하면서, 이미지와 텍스트를 통합적으로 이해하는 새로운 패러다임이 형성되었다[5]. 이러한 VLM은 텍스트 프롬프트만으로 이미지 묘사, 시각적 질의응답, OCR(Optical Character Recognition), 객체 검출 등 다양한 시각적 과제를 수행할 수 있다[6]-[8]. VLM은 zero-shot 방식으로 동작하여 추가 학습 없이 새로운 환경에 즉시 적용 가능해 높은 유연성을 제공하며, 데이터 수집 및 라벨링 비용을 크게 절감할 수 있는 장점을 가진다.

그러나 VLM은 이미지를 시각 인코더를 통해 고정된 수의 시각 토큰으로 변환한 후 이를 언어 모델이 처리하는 구조적 특성으로 인해, 원본 이미지의 공간적 세부 정보가 압축되는 한계를 가진다[9][10]. 이러한 특성은 특히 동일 클래스 객체가 밀집된 환경에서 더욱 두드러지며, 객체 간 경계 정보가 소실되는 문제가 발생한다. 실제로 주차장과 같이 차량이 밀집된 환경에서는 인접한 차량들이 하나의 객체로 병합되거나, 일부 차량이 검출되지 않는 과소 검출 현상이 나타날 수 있다.

반면, 전통적인 객체 검출 방식은 다양한 해상도의 특징 맵을 생성하고, 업샘플링과 다운샘플링을 통해 의미 정보와 공간 정보를 양방향으로 결합함으로써 객체의 위치와 경계를 보다 정밀하게 유지한다. 또한, 미리 정의된 다양한 크기와 비율의 anchor 박스를 기반으로 객체를 탐지하고 이를 보정하는 방식으로 다양한 스케일의 객체를 효과적으로 처리하는 연구도 수행되었다[11]. 여기서는 밀집 환경에서 발생하는 전경-배경 클래스 불균형 문제를 완화하기 위한 손실 함수도 제안되었다. 이러한 구조적 차이로 인해, 전통적 검출기는 밀집 환경에서도 개별 객체를 비교적 정확하게 분리하는 반면, VLM은 장면 수준 이해에 최적화되어 있어 정밀한 객체 분리 성능이 상대적으로 낮을 수 있다.

이로 인해 VLM을 주차장 영상과 같은 밀집 차량 환경에 적용할 경우, 인접한 차량을 하나의 큰 바운딩 박스로 묶어 검출하는 병합 검출, 일부 차량만 검출되는 과소 검출, 심지어 이미지 전체를 하나의 객체로 인식하는 오류가 발생할 수 있다.

한편, 자율 주행이나 주차 관리 시스템 등 실제 운용 시스템의 작업 특화 검출기는 대규모 도메인 특화 데이터에 대한 지도 학습을 전제로 하며, 최근에는 가려짐이나 주차면 인식 같은 더 고도화된 과제로 연구가 확장되고 있다. 본 연구의 목적은 이러한 시스템을 대체하거나 그 성능을 개선하기 보다는 사전학습 VLM의 밀집 객체 검출 한계를 정량 분석하고 입력 단계 제어만으로 이를 완화할 수 있는지를 검증하는 데 연구 목적을 둔다. 따라서 제안하는 방법은 데이터 수집 및 추가 학습이 어려운 도메인이나 소규모 응용 프로토타입 단계에서 활용 가능하다.

본 연구에서는 VLM을 활용한 주차장 영상 내 차량 검출 성능 향상을 목표로 한다. 이를 위해 모델 구조를 변경하거나 추가 학습을 수행하는 대신, 입력 단계에서 VLM의 주의를 지역적 문맥에 집중시키는 접근 방식을 제안한다. 구체적으로, 다단계 프롬프트를 통해 입력 이미지의 장면을 먼저 해석하고, 이를 기반으로 밀집 객체의 개별 검출을 유도하는 프롬프트를 자동 생성한다. 또한, 원본 이미지

에 격자 형태의 선을 합성하는 그리드 오버레이를 활용한 시각적 단서를 제공하여 밀집 영역을 분할하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 선행 연구를 검토하고, 3장에서는 다단계 프롬프트 생성과 그리드 오버레이 기반의 제안 방법을 기술한다. 4장에서는 3개의 공개 데이터셋에 대한 실험 결과를 제시하며, 5장에서는 결론 및 향후 연구 방향을 논의한다.

## II. 관련 연구

### 2.1 Vision language model 기반 객체 검출

VLM은 대규모 이미지-텍스트 쌍 데이터로 사전 학습되어 이미지와 텍스트를 통합적으로 이해하는 모델이다. 초기 VLM인 CLIP은 이미지-텍스트 대조 학습을 통해 시각적 표현과 언어적 표현을 공유 임베딩 공간에 정렬하는 방식을 제안하였다[5]. 이후 Flamingo[6], BLIP-2[7]은 LLM(Large Language Model)과 시각 인코더를 결합하여, 이미지에 대한 자유 형식의 텍스트 출력을 생성하는 VLM으로 발전하였다. 이러한 모델들은 이미지 묘사, VQA, OCR 등 다양한 시각적 과제에서 범용적인 성능을 보이며, 단일 모델로 복수의 문제를 처리할 수 있다는 장점을 가진다.

VLM의 객체 검출 능력은 visual grounding 과제를 통해 구현된다. Kosmos-2[12]는 이미지 내 객체의 위치를 텍스트와 함께 출력하는 grounded language-image 사전학습을 제안하였으며, Qwen-VL[8]은 바운딩박스 좌표를 정규화된 텍스트 토큰으로 출력하는 방식을 통해 zero-shot 객체 검출을 지원한다[8]. Grounding DINO(DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detectio)[13]는 텍스트 프롬프트와 이미지 특징을 결합한 open-set 검출기로서 높은 검출 성능을 달성하였으나, 이는 검출에 특화된 아키텍처를 사용한다는 점에서 범용 VLM과는 구분된다.

특히 Qwen2.5-VL은 동적 해상도 처리와 개선된 시각 인코더를 통해 다양한 크기의 이미지를 효과적으로 처리할 수 있다. 이러한 기능은 별도의 검출 전용 모듈이나 추가 학습 없이 텍스트 프롬프트만으로 객체 위치를 추정할 수 있다는 점에서 기존 검출 모델 대비 유연성이 높다. 그러나 VLM은 auto-regressive 방식으로 좌표를 순차적으로 생성하므로, 출력 토큰 수의 제한으로 인해 다수의 객체를 완전히 열거하는 데 한계가 있다. 이는 특히 밀집 장면에서 검출 누락의 직접적 원인이 될 수 있다.

기존 VLM 기반 객체 검출 연구는 주로 소수의 다양한 클래스 객체가 분포하는 일반적인 장면 해석에 초점이 맞추어져 있다. 밀집된 동일 클래스 객체, 예를 들어 주차장의 다수 차량이나 군중 장면과 같은 시나리오에서 VLM의 검출 성능이 어떻게 변화하는지에 대한 분석은 충분히 연구되어 있지 않다. 특히 차량 밀집도가 증가함에 따라 검출 성능이 어떤 패턴으로 저하되는지, 그 원인이 무엇인지에 대한 정량적 연구는 아직 충분히 이루어지지 않았다. 이에 본 연구에서는 VLM 기반 차량 검출에서 밀집도에 따른 검출 성능 저하를 정량적으로 분석하고 이를 개선하는 방법을 제안한다.

### 2.2 시각적 프롬프트(Visual prompt)

시각적 프롬프트는 입력 이미지를 수정하여 VLM의 시각적 이해를 유도하는 기법으로, 텍스트 프롬프트만으로 전달하기 어려운 공간적 정보를 시각적 단서를 통해 보완하는 접근이다[14][15]. 이미지에 마커, 화살표, 원, 번호 등의 시각적 요소를 오버레이하여 특정 영역을 명시적으로 지칭함으로써, 모델이 관심 영역을 보다 정확하게 인식하도록 유도할 수 있다.

대표적으로 SoM(Set-of-Mark) 프롬프팅[14]은 이미지 내 각 영역에 번호가 부여된 마커를 삽입하여 VLM이 특정 객체를 정확하게 참조하도록 하는 방법이며, Visual Chain-of-Thought 프롬프팅[15]은 시각적 단서를 단계적으로 결합하여 VLM의 추론 성능을 향상시키는 방식이다. 이러한 연구들은 시각적 단서의 추가가 VLM의 visual grounding 성능을 개선하는 데 효과적임을 보여준다.

한편, VLM의 성능은 텍스트 프롬프트 설계에도

크게 의존한다. 기존 연구에서는 고정된 프롬프트 템플릿을 사용하거나 단계적 지시를 통해 모델의 추론을 유도하는 방식이 활용되었으며, 검출 대상의 특성이나 출력 형식을 명시적으로 기술하는 것이 성능 향상에 기여하는 것으로 알려졌다. 그러나 이러한 접근은 이미지 간 장면 구성이나 객체 밀집도와 같은 차이를 반영하지 못하는 한계를 가지고 있다.

특히 기존 시각적 프롬프트 연구는 소수의 객체를 지칭하거나 특정 영역을 강조하는 데 초점을 두고 있으며, 다수의 동일 클래스 객체가 밀집된 환경에서 객체를 공간적으로 분리하여 개별 검출을 유도하는 문제는 충분히 다루지 않았다. 또한 SoM 방식은 사전 영역 분할 결과에 의존한다는 점에서 추가적인 모델이나 전처리 과정이 필요하다는 한계를 가지고 있다.

본 연구는 이러한 한계를 극복하기 위해, 텍스트 프롬프트와 시각적 단서를 결합한 입력 제어 방식을 제안한다. 입력 이미지의 장면을 분석하여 프롬프트를 적응적으로 생성하고, 동시에 그리드 기반 시각적 단서를 제공함으로써 밀집된 객체를 공간적으로 분리하여 검출을 유도한다. 이는 기존 고정 프롬프트 기반 접근과 달리, 장면 특성에 따라 동적으로 입력을 구성한다는 점에서 차별화된다.

## III. 제안 방법

VLM 기반 밀집 차량 검출에서 나타나는 낮은 재현율 문제를 해결하기 위해, 본 연구에서는 그리드 오버레이 기반 프롬프트 방법을 제안한다. 이 방법은 추가 학습이나 파인튜닝 없이, 입력 이미지에 시각적 단서를 부여하고 장면에 적응적인 프롬프트를 구성함으로써 VLM의 밀집 객체 분별력을 향상시키는 것을 목표로 한다.

VLM은 밀집된 차량 영역에서 개별 객체를 명확히 구분하지 못하고 하나의 큰 바운딩박스로 병합하여 출력하는 경향이 있다. 단순히 "Detect all cars"와 같은 직접적인 지시만으로는 이러한 문제를 해결하기 어렵다. 선행 연구에서도 프롬프트의 구성 방식이 VLM의 출력 품질에 영향을 미치는 것으로 알려져 있다[16]. 따라서 밀집 환경에서도 개별 객체를 분리하여 검출하도록 유도하는 정교한 프롬프트 설계가 필요하다.

제안하는 방법에서는 입력 이미지 $I$에 대해 $n\times n$ 그리드를 오버레이한 이미지 $G_n(I)$를 생성하고, 밀집 객체를 잘 분리할 수 있도록 구성된 프롬프트 $T_{multi}$를 VLM에 입력하여 객체 검출을 수행한다. 이를 식 (1)로 표현할 수 있다.

$$B_{position} = F_{VLM}(G_n(I), T_{multi}) \tag{1}$$

여기서 $F_{VLM}(I,T)$은 이미지 I와 프롬프트 T를 입력으로하는 VLM 기반 객체 검출 함수이다.

식 (1)의 전체 처리 과정은 그림 1에 도식화하였다. 제안 방법은 세 개의 프롬프트 문맥과 시각적 전처리로 구성된다. 장면 묘사(Scene description context)에서는 원본 이미지 I로부터 장면 설명 프롬프트를 생성한다. 이때 VLM 입력 프롬프트의 핵심 요소는 차량 시점, 차량 배치 패턴, 배경 요소, 밀집도이며, 이들 항목을 포함한 2문장의 장면 묘사를 요청한다. 분리 지시(Anti-cluster context)는 그리드 공간 안내, 셀 단위 독립 검출, 검출 세부 지침의 세 가지 요소로 구성되며, 이를 통해 밀집된 객체의 개별 분리를 유도한다. 마지막 검출 지시(Detection context)는 객체 검출 위치를 출력할 수 있도록 올바른 json 출력과 예제로 구성되어 있다. 시각적 전처리에서는 원본 이미지 I 위에 $n\times n$ 격자 라인을 알파 블렌딩으로 합성하여 $G_n(I)$를 생성한다. 마지막으로 $G_n(I)$와 $T_{multi}$를 VLM에 함께 입력하여 바운딩 박스 집합 $B_{position}$을 얻는다. 3.1절에서는 프롬프트 $T_{multi}$, 3.2절에서는 그리드 오버레이 함수 $G_n(I)$ 각각 설명한다.

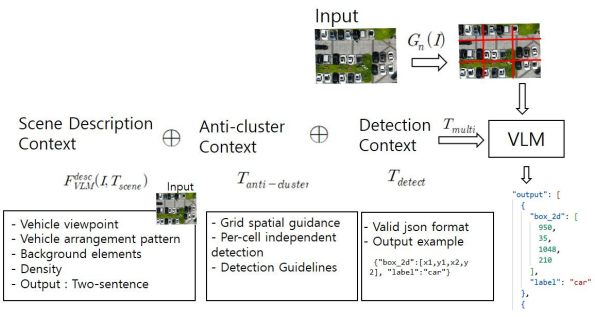


그림 1. 단계별 데이터 흐름과 제안 방법의 전체 파이프라인

Fig. 1. Step-by-step data flow and overall pipeline of the proposed method

### 3.1 다단계 프롬프트(Multi-stage prompt)

식 (1)에서 프롬프트 $T_{multi}$는 입력 이미지의 장면 정보를 반영하여 이미지 마다 개별적으로 생성되며, 장면 문맥, 밀집 객체 분리 지시, 검출 명령의 세 가지 요소를 결합해 식 (2)처럼 표현된다.

$$T_{multi} = (F_{VLM}^{desc}(I, T_{scene}) \oplus T_{anti-cluster} \oplus T_{detect} \quad (2)$$

여기서 $F_{VLM}^{desc}()$는 이미지에 대한 장면 설명을 생성하는 함수이며, $\oplus$는 텍스트 결합 연산을 의미한다. $T_{scene}$은 전체 이미지에 대한 설명을 유도하는 지시문, $T_{anti-cluster}$는 밀집 객체의 개별 검출을 유도하는 분리 지시문, $T_{detect}$는 검출 대상 및 출력 형식을 지정하는 검출 명령 프롬프트이다.

#### 3.1.1 장면 설명 유도 지시문

검출에 앞서, VLM을 이용하여 입력 이미지의 장면 상황을 분석할 필요가 있다. 이는 LLM 기반 추론에서 문맥 정보를 사전에 추출하여 후속 추론의 일관성을 높일 수 있다. 이를 통해 VLM의 장면 이해 능력을 검출 프롬프트 구성에 직접 활용한다.

생성된 장면 기술로부터 검출에 유의미한 문맥 정보를 추출하여 문맥 프롬프트를 구성한다. 예를 들어, “a crowded parking lot with cars parked in tight rows”와 같은 설명에서 밀집 상태와 배치 패턴을 추출하여 프롬프트에 반영한다. 이 과정은 이미지마다 개별적으로 수행되기 때문에 다양한 촬영 조건과 차량 배치에 유연하게 대응할 수 있다.

#### 3.1.2 밀집 객체 분리 지시문

$T_{anti-cluster}$는 밀집 환경에서 발생하는 객체 병합 문제를 완화하기 위해, 인접한 차량을 개별 객체로 분리하여 검출하도록 VLM을 유도하는 프롬프트이다. 이 프롬프트는 다음의 세 가지 요소로 구성된다. 첫째, 인접한 차량을 하나의 바운딩 박스로 통합하지 않도록 하는 지시이다. 둘째, 밀집된 영역에서도 각 차량을 독립적인 객체로 검출하도록 강조하는 것이다. 마지막으로, 이미지 경계에 부분적으로 포함된 차량까지 검출 대상에 포함하도록 명시하는 것이다. 이러한 구성은 VLM의 장면 수준 인식 경향을 개별 객체 수준의 정밀한 검출로 전환하는 역할을 수행한다.

#### 3.1.3 검출 명령 및 출력 형식 정의 지시문

$T_{\text{detect}}$는 검출 대상과 출력 형식을 명확히 지정하기 위한 프롬프트이다. 이는 검출 대상 클래스, 바운딩 박스 좌표 형식 등을 포함한다. 출력 형식을 명확히 정의함으로써 VLM이 자유 형식 텍스트 대신 구조화된 검출 결과를 생성하도록 유도할 수 있으며, 후속 처리 과정의 안정성을 확보할 수 있다.

식 (2)의 최종 프롬프트 $T_{multi}$는 장면 문맥, 밀집 객체 분리 지시, 검출 명령을 결합하여 생성된다. 이러한 다단계 구조를 통해, 단일 지시문으로는 달성하기 어려운 적응적이고 정밀한 검출 제어가 가능하다. 특히 제안하는 프롬프트는 그리드 오버레이 이미지 $G_n(I)$에 적용되어, 시각적 단서와 언어적 지시가 결합된 형태로 VLM의 검출 성능을 높일 수 있도록 구성했다.

### 3.2 그리드 오버레이(Grid overlay)

VLM은 밀집된 차량 영역에서 개별 객체를 구분하지 못하고 전체 영역을 하나의 바운딩 박스로 출력하는 경향이 있으며, 이는 VLM의 시각 인코더가 밀집 영역 내 객체 경계를 충분히 분리하지 못하는 데에서 기인한다. 텍스트 프롬프트만으로는 이러한 시각적 인코딩 한계를 완전히 극복하기 어렵다. 기존 연구에서는 이미지를 물리적으로 분할하여 소형 객체의 검출률을 향상시키는 방법이 제안되었다[17]. 본 연구에서는 이미지 분할 대신, 원본 이미지 위에 규칙적인 격자 형태의 시각적 참조 구조를 합성하는 그리드 오버레이 방법을 제안한다. 그리드 라인은 밀집 영역을 시각적으로 분할함으로써 VLM이 지역적 영역에 주의를 집중하도록 유도하는 역할을 한다.

먼저, 입력 이미지 $I$에 대해 $n \times n$ 그리드에서 각 픽셀이 그리드 라인에 속하는지를 나타내는 이진 마스크 $M(x,y)$를 식 (3)과 같이 정의한다.

$$M(x,y) = I((x,y) \in \Omega_{grid}) \tag{3}$$

여기서 $\Omega_{grid}$는 그리드 라인이 차지하는 픽셀 영역의 집합이다. 따라서 특정 픽셀이 $\Omega_{grid}$에 속하면 1을 그렇지 않으면 0이 된다. 그리드의 폭은 너무 작으면 인식이 안되고, 너무 크면 노이즈로 인식될 수 있어 실험적으로 3으로 결정했다.

식 (1)에서 그리드 오버레이 이미지 $G_n(I)$는 마스크 $M(x,y)$와 투명도 계수 $\alpha$를 이용한 알파 블렌딩을 통해 생성된다. 구체적으로, 그리드 라인 영역에서는 지정된 색상 벡터 $c_{line}$과 원본 이미지 $I(x,y)$를 혼합하고, 그 외 영역에서는 원본 이미지를 그대로 유지한다. 즉 $M(x,y)=1$에서는 그리드 색상과 원본 이미지를 알파 블렌딩한 값을 사용하고, $M(x,y)=0$에서는 원본 이미지 값을 그대로 유지한다. 이 과정은 식 (4)와 같이 정의된다.

$$\begin{aligned} G_n(x,y) = M(x,y) \cdot \mathrm{G_{grid}(x,y)} \\ + (1 - M(x,y)) \cdot \mathrm{I(x,y)} \end{aligned} \tag{4}$$

여기서 $G_n(x,y)$는 그리드 오버레이가 적용된 최종 이미지의 픽셀 값이며, 식 (1)의 $G_n(I)$를 픽셀 단위로 정의한 것이다. $G_{grid}(x,y)$는 그리드 라인 영역에서의 픽셀 값으로, 식 (5)와 같이 정의된다.

$$G_{grid}(x,y) = \alpha \mathrm{c_{line}} + (1-\alpha)\mathrm{I(x,y)} \tag{5}$$

여기서 $\alpha \in [0,1]$는 그리드 색상 $c_{line}$과 원본 이미지 $I(x,y)$ 간의 혼합 비율을 조절하는 투명도 계수이다. $\alpha=1$인 경우 그리드 라인은 완전히 불투명하여 원본 정보를 가리게 되며, $\alpha=0$인 경우 그리드 라인이 보이지 않게 된다. 본 연구에서는 시각적 단서를 제공하면서도 객체 정보를 과도하게 훼손하지 않도록 $\alpha$값을 실험적으로 0.2로 설정했다.

또한, 그리드 라인의 색상 $c_{line}$은 입력 이미지의 색상 분포를 고려하여 결정된다. $c_{line}$은 이미지의 대표 색상들과 가장 잘 구별되는 색상으로 정의되어야 한다. 이를 위해 입력 이미지에 k-means 클러스터링을 적용하여 대표 색상 집합 $\mu_1, \mu_2, \ldots, \mu_K$를 추출하고, 사전 정의된 후보 색상 집합 C 중에서 대표 색상들과의 CIELAB(International Commission on Illumination L*a*b* Color Space) 색상 거리가 최대가 되는 색상을 식 (6)을 사용해 선택한다.

$$c_{line} = \arg max_{c \in C} \min_k \Delta \mathrm{E(c,\mu_k)} \tag{6}$$

식 (6)의 $\Delta E(c,\mathrm{u_k})$는 CIELAB 색공간에서의 유클리드 거리로, 식 (7)과 같이 정의된다.

$$\Delta E(c,\mu_k) = \sqrt{(\mathrm{L_c - L_k^2}) + (\mathrm{a_c - a_k^2}) + (\mathrm{b_c - b_k^2})} \tag{7}$$

여기서 $(L_c, a_c, b_c)$는 후보 색상 c의 CIELAB 좌표이며, $(L_k, a_k, b_k)$는 클러스터 중심 $\mu_k$ CIELAB 좌표이다. L은 밝기, a와 b는 각각 녹-적, 청-황 색상 축을 나타낸다. 이러한 방식은 이미지 내 주요 색상과의 대비를 최대화하여, 그리드 라인이 객체와 시각적으로 구분되도록 한다.

그리드 오버레이 생성에 필요한 하이퍼 파라미터를 요약하면 그리드 폭, 그리드 크기, 색상 그리고 투명도이다. 색상은 식 (6)을 사용해 계산된다. 그리드 폭과 투명도는 실험적으로 3, 0.2로 설정했다. 이들 값은 성능에 큰 영향을 미치지 않는다.

$G_n(x,y)$에서 그리드 크기 $n$은 하이퍼 파라미터로서 검출 성능에 직접적인 영향을 미친다. $n$이 작은 경우($n$=2)에는 그리드 셀이 커져 밀집 영역 분할 효과가 제한적이지만, 그리드 라인이 객체를 가릴 확률이 낮아 오검출이 증가하지 않는 장점이 있다. 반면 $n$이 큰 경우($n$=8)에는 밀집 영역을 세밀하게 분할할 수 있어 재현율 향상에 유리하나, 그리드 라인이 객체와 중첩될 가능성이 증가하여 VLM의 인식 혼동을 유발할 수 있다. 또한 $n$이 증가할수록 그리드 라인이 차지하는 전체 픽셀 비율이 증가하여 원본 이미지 정보의 손실이 커지므로, 최적의 $n$ 값은 실험적으로 결정되어야 한다.

제안 방법의 전체 과정을 정리하면, 입력 이미지 $I$에 대해 먼저 VLM 장면 해석을 통해 문맥 정보

를 추출하고, anti-clustering 지시문 및 검출 지시와 결합하여 프롬프트를 생성한다. 이후 원본 이미지와 그리드 오버레이 이미지를 각각 VLM에 입력하여 검출을 수행한다. 이를 통해 밀집 영상에서도 객체 검출에 대한 재현율을 높일 수 있다.

## Ⅳ. 실험 및 결과 분석

### 4.1 실험 데이터 세트

제안 방법의 일반화 성능을 검증하기 위해 3개의 항공/위성 차량 데이터셋[18]-[20]을 사용하였다. DB1은 150장의 이미지에 총 2,874대의 차량(이미지당 평균 19.2대, 최소 4대, 최대 62대)을 포함하며, DB2는 299장에 5,683대(평균 19.0대, 최소 1대, 최대 93대), DB3은 31장에 717대(평균 23.9대, 최소 6대, 최대 47대)를 포함한다. 각 이미지에 대해 수작업 라벨링하여 Ground Truth를 구성하였다. 전체 데이터셋은 총 480장, 9,274대의 차량으로 구성되어 있으며, 데이터셋 간 촬영 조건과 해상도가 상이하여 다양한 환경에서의 성능 검증이 가능하다. 표 1은 3개의 데이터셋에 대한 구성을 보여주고, 그림 2는 각 데이터셋의 대표 샘플 이미지를 보여준다. 이들로부터 촬영 시점과 차량 밀집도의 차이를 확인할 수 있다.

표 1. 데이터셋 구성(이미지 수, 차량 수, 해상도 레벨 분포)
Table 1. Dataset composition(number of images, vehicles, and resolution level distribution)

| Dataset | DB1 | DB2 | DB3 |
|---|---|---|---|
| #Images | 150 | 299 | 31 |
| #Vehicles | 2874 | 5683 | 717 |
| Avg. Veh. | 19.2 | 19.0 | 23.9 |
| Min. Veh. | 4 | 1 | 6 |
| Max. Veh. | 62 | 93 | 47 |
| L1 | 1 | 0 | 4 |
| L2 | 29 | 0 | 5 |
| L3 | 120 | 7 | 12 |
| L4 | 0 | 292 | 9 |

Qwen2.5-VL 모델은 입력 이미지의 총 픽셀 수에 따라 시각 토큰의 수가 결정되며, 이는 모델이 처리하는 공간 정보의 해상도에 직접적인 영향을 미친다[21]. 이러한 특성을 고려하여, 본 실험에서는 이미지의 총 픽셀 수를 기준으로 4단계 해상도 레벨로 분류하였다. L1은 250,000 픽셀 미만, L2는 250,000 이상 500,000 미만, L3은 500,000 이상 1,003,520 이하, L4는 1,003,520 초과로 구분한다. DB1은 L2와 L3 수준의 이미지가 대부분(149장)이며, DB2는 L4 수준의 고해상도 이미지가 292장으로 전체의 97.7%를 차지한다. DB3은 L1부터 L4까지 다양한 해상도 분포를 보여 해상도 다양성을 반영하며, 이를 통해 해상도에 따른 검출 성능의 변화도 간접적으로 확인할 수 있다.

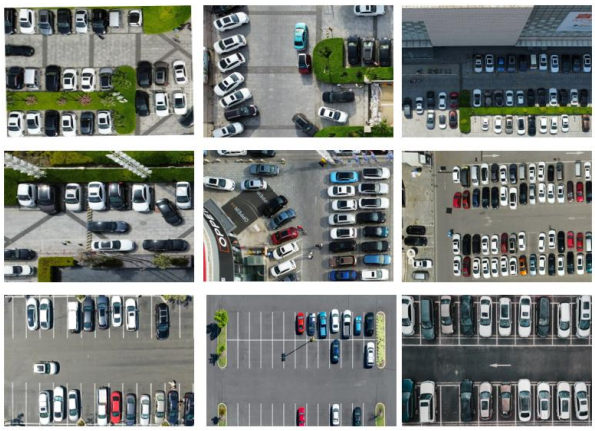

그림 2. 각 데이터셋의 샘플 이미지(좌: DB1, 중: DB2, 우: DB3)
Fig. 2. Sample images from each dataset(left: DB1, center: DB2, right: DB3)

### 4.2 VLM 기반 차량 검출 성능

본 실험은 NVIDIA RTX 5090 GPU(32GB) 환경에서 수행되었으며, VLM 모델로는 Qwen2.5-VL-3B를 사용하였다. Qwen2.5-VL은 7B, 32B, 72B 등 더 큰 규모의 모델도 제공하지만, 7B 모델부터 메모리 부족으로 안정적인 추론이 불가능하여 3B 모델을 채택하였다.

Baseline 실험은 원본 이미지를 VLM에 직접 입력하여 차량을 검출하는 방식으로 수행하였다. 프롬프트는 기본적인 검출 지시문("Detect all cars in the image")만을 사용하며, 그리드 오버레이나 다단계 프롬프트 생성은 적용하지 않았다. 이를 통해 VLM

의 기본적인 차량 검출 능력과 밀집 영역에서의 성능 한계를 정량적으로 확인하고자 하였다.

본 연구에서는 객체 검출 성능을 정량적으로 평가하기 위해 정밀도, 재현율, F1-score 세 가지 지표를 사용하였다. 본 연구는 밀집 영상의 과소 검출 문제 완화를 위해 재현율을 핵심 평가 지표로 삼되, 오검출 증가를 함께 모니터링하기 위해 정밀도와 F1-score를 보조 지표로 사용하였다.

검출 박스와 GT(Ground Truth) 박스 간의 매칭 여부는 PASCAL VOC(PASCAL Visual Object Classes), MS COCO(Microsoft Common Objects in Contex) 등 객체 검출 분야의 표준 평가 프로토콜[22]에 따라 IoU 0.5 이상을 매칭 임계값으로 채택하였다. 또한, GT 차량 수를 기준으로 이미지를 5개 밀도 구간(1 - 10, 11 - 20, 21 - 30, 31 - 50, 51대 이상)으로 분류하고, 각 구간에서 baseline과 제안 방법의 재현율 변화를 분석하여 밀집도 증가에 따른 성능 변화 패턴을 정량적으로 검증하였다.

Baseline의 전체 성능은 표 2와 같이 3개 데이터셋 합산 기준 정밀도 0.925, 재현율 0.429, F1 0.586으로 나타났다. 정밀도는 높은 수준을 유지하였으나 재현율이 매우 낮아, VLM이 검출한 차량은 대부분 정확하지만 실제 존재하는 차량의 절반 이상을 놓치는 심각한 과소 검출 문제가 확인되었다.

데이터셋별로는 DB1의 재현율 0.516, DB2의 재현율 0.370, DB3의 재현율 0.542를 기록하였다. DB2는 밀집 이미지 비율이 가장 높아 재현율이 가장 낮은 수준을 보였다.

표 2. Baseline과 제안 방법의 데이터셋별 성능 비교 (정밀도, 재현율, F1)
Table 2. Performance comparison between Baseline and Proposed method by dataset (Precision, Recall, F1)

| | Method | Precision | Recall | F1 |
|---|---|---|---|---|
| D1 | Baseline | 0.955 | 0.516 | 0.670 |
| | Proposed | 0.936 | 0.753 | 0.835 |
| D2 | Baseline | 0.903 | 0.370 | 0.525 |
| | Proposed | 0.833 | 0.633 | 0.720 |
| D3 | Baseline | 0.937 | 0.542 | 0.687 |
| | Proposed | 0.889 | 0.772 | 0.826 |
| Total | Baseline | 0.925 | 0.429 | 0.586 |
| | Proposed | 0.871 | 0.681 | 0.765 |

그림 3은 Baseline에서 나타나는 대표적인 검출 실패 사례를 보여준다. 밀집된 차량 영역에서 인접한 차량이 하나의 바운딩박스로 병합되거나, 다수의 차량이 검출에서 누락되는 과소 검출 현상을 확인할 수 있다.

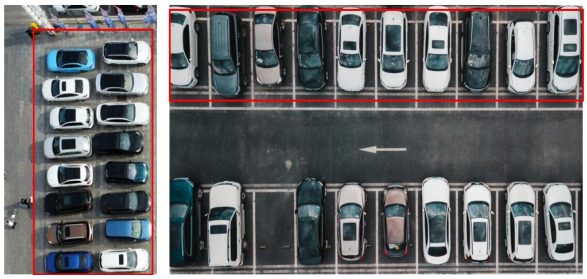

그림 3. Baseline에서의 검출 실패 사례: 병합 검출(좌)과 과소 검출(우)
Fig. 3. Detection failure cases in Baseline: merged detection (left) and under-detection (right)

차량 밀도에 따른 제안 방법의 효과성을 분석하기 위해, GT 차량 수를 1~10, 11~20, 21~30, 31~50, 51+ 구간으로 분류하여 baseline 방법과 제안 방법의 구간별 재현율 측정하여 그림 4의 그래프로 나타내었다.

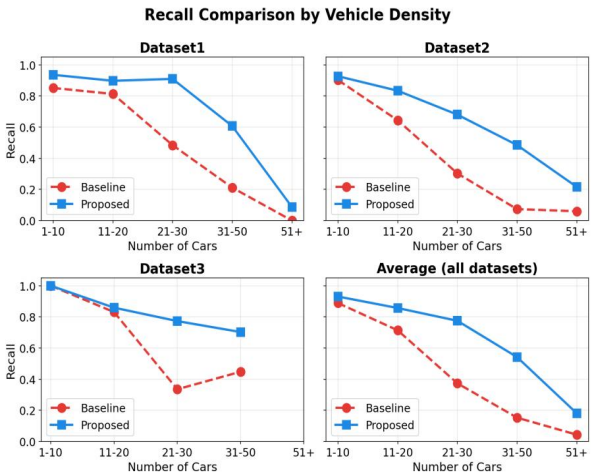


그림 4. 차량 밀도 구간별 재현율 비교
Fig. 4. Recall comparison across vehicle density ranges

그림 4의 그래프 결과를 분석하면, DB1에서는 저밀도 구간(1 - 10, 11 - 20)의 격차가 각각 8.5%, 8.4%로 작지만, 21 - 30 구간에서 42.6%, 31 - 50 구간에서 39.5%로 격차가 크게 확대된다. 특히 51+

구간에서 baseline은 완전히 실패하는 반면 제안 방법은 8.7%의 부분 검출을 유지한다. DB2의 1 – 10 구간에서 두 방법이 거의 일치(격차 2.5%)하나, 31 – 50 구간에서 baseline 7.3%, 제안 방법 48.5%로 41.2%의 가장 큰 격차를 보인다. 이는 밀집 이미지 비중이 높은 데이터셋에서 제안 방법의 효과를 보여준다. DB3은 1 – 10 구간에서 두 방법 모두 재현율 100%로 동일하나, 21 – 30 구간에서 43.8%의 격차가 발생하였다. 실험에 사용된 세 개의 데이터셋 모두에서 두 곡선의 격차가 차량 수 증가에 따라 확대된다는 점을 확인할 수 있다.

### 4.3 제안 방법의 성능 평가

표 2에서 확인할 수 있듯이, 제안 방법을 적용한 결과, 전체 성능은 정밀도 0.871, 재현율 0.681, F1-score 0.765로 나타났다. Baseline 대비 재현율이 0.429에서 0.681로 약 25% 향상되었다. 정밀도는 0.925에서 0.871으로 약 5% 하락하였으나, 이는 그리드 이미지에서의 추가 검출로 인한 FP 증가에 기인하며 재현율 25% 향상 대비 정밀도 5% 하락은 상대적으로 작은 수준이다. 정밀도와 재현율을 모두 고려한 F1-score의 경우 제안하는 방법이 0.586에서 0.765로 17.9% 개선되었다.

데이터셋별로는 DB1에서 재현율이 0.516에서 0.753으로 23.7%, DB2에서 0.370에서 0.633으로 26.3%, DB3에서 0.542에서 0.772로 23.0% 향상되었다. 특히 Baseline에서 가장 낮은 재현율을 보였던 DB2에서 약 26.3%의 가장 큰 개선이 나타났으며, 이는 밀집도가 높은 이미지가 많은 데이터셋에서 제안 방법이 효과적임을 확인할 수 있다.

제안 방법의 구성 요소가 성능에 어떻게 기여하는지 검증하기 위해, 장면 컨텍스트 $F_{VLM}^{desc}()$와 시각적 컨텍스트 $T_{anti-duster}$를 각각 단독으로 적용한 실험을 수행해 표 3에 나타내었다. 두 컨텍스트는 단독 적용 시에도 baseline 대비 모든 데이터셋에서 F1-score를 향상시켰으며, DB1에서는 장면 컨텍스트(+13.9%)가, DB2, DB3에서는 시각적 컨텍스트(+17.6%, +11.1%)가 더 큰 기여를 보였다. 두 컨텍스트를 결합한 제안 방법은 모든 데이터셋에서 단일 적용보다 높은 성능을 달성하며(DB1 +16.4%, DB2 +19.4%, DB3 +13.9%), 이는 두 메커니즘이 각각 다른 실패 유형에 대해 상호 보완적임을 알 수 있다.

표 3. 제안 방법의 구성 요소별 F1-score 기여도 분석
Table 3. Component-wise F1-score contribution of the proposed method

| | | DB1 | DB2 | DB3 |
|---|---|---|---|---|
| baseline | I | 0.671 | 0.525 | 0.687 |
| + $F_{VLM}^{desc}()$ | I | 0.810 | 0.634 | 0.710 |
| + $T_{anti-duster}$ | $G_n(I)$ | 0.792 | 0.702 | 0.799 |
| proposed | $G_n(I)$ | 0.835 | 0.720 | 0.826 |

그리드 크기 n에 따른 성능 비교 결과를 표 4에 제시한다. n=2에서 F1 0.778로 가장 높은 성능을 기록하였으며, n이 증가할수록 재현율은 소폭 상승하나 정밀도 하락이 더 크게 작용하여 F1-score가 감소하는 경향을 보인다. 이는 그리드 라인이 객체와 중첩될 가능성이 증가하여 오검출이 늘어나기 때문이다.

표 4. 그리드 크기 n에 따른 검출 성능 비교
Table 4. Detection performance comparison by grid size n

| Grid | Precision | Recall | F1 |
|---|---|---|---|
| 2×2 | 0.842 | 0.723 | 0.778 |
| 3×3 | 0.861 | 0.687 | 0.764 |
| 4×4 | 0.798 | 0.719 | 0.757 |
| 8×8 | 0.734 | 0.722 | 0.728 |

## V. 결 론

본 연구는 VLM을 활용한 밀집 차량 검출 환경에서 시각 인코더의 공간 정보 압축으로 인해 인접 객체가 병합되거나 누락되는 과소 검출 문제를 정량적으로 분석하고, 이를 완화하기 위한 입력 단계의 제어 방법을 제안하였다. 본 연구는 작업 특화 학습 검출기 수준의 성능 개선보다는 사전학습된 범용 VLM의 zero-shot 검출 능력 자체를 검증하고 개선하는 데 목적을 둔다. 제안 방법은 모델 구조 변경이나 추가 학습 없이, 입력 이미지의 장면 특성을 VLM 자체로 해석하여 밀집 객체의 검출을 유도하는 다단계 프롬프트를 자동 생성하고, 원본 이미지 위에 격자 형태의 시각적 참조 구조를 알파 블

렌딩으로 합성하는 그리드 오버레이를 결합하여 밀집 영역에 대한 VLM의 지역적 주의를 강화한다.

3개의 공개 항공·위성 차량 데이터셋(총 480장, 9,274대)에 대한 실험에서 제안 방법은 baseline 대비 재현율 약 25%, F1-score 약 18%의 향상을 달성하였다. 항공·위성 영상은 작은 객체 크기, 높은 밀도, 약한 시각적 단서가 결합되어 현재 VLM이 충분한 성능을 보이지 못하는 영역이며, 본 연구는 이 환경에서 입력 단계 제어만으로 의미 있는 개선이 가능함을 보였다.

다만, 제안 방법은 다단계 추론으로 인한 추론 시간 증가와 GPU 메모리 제약으로 본 실험에서 Qwen2.5-VL-3B 모델을 사용한 한계가 있다. 향후 연구에서는 적응적 그리드 분할과 단일 추론 기반 경량화로 추론 효율을 개선하고, 차량 외 다양한 밀집 객체 클래스 및 지상 카메라 기반 응용 데이터셋으로 검증을 확장한다. 또한 본 연구의 객체 검출 범위를 넘어 가려짐 처리와 주차면 단위 인식 등 고차원 시각 이해 과제로 연구를 확장하고, 고성능 GPU 환경 확보 후 Qwen3-VL을 비롯한 최신 VLM에서의 검증과 알고리즘 고도화를 수행하고자 한다.

## References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, pp. 580-587, Jun. 2014. https://doi.org/10.1109/CVPR.2014.81.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Adv. Neural Inf. Process. Syst. (NeurIPS), Montréal CANADA, Vol. 28, Dec. 2015.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, pp. 779-788, Jun. 2016. https://doi.org/10.1109/CVPR.2016.91.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers", Eur. Conf. Comput. Vis. (ECCV), Online, pp. 213-229, Aug. 2020. https://doi.org/10.48550/arXiv.2005.12872.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision", Int. Conf. Mach. Learn. (ICML), Online, Vol. 139, pp. 8748-8763, Jul. 2021.

[6] J.-B. Alayrac, et al., "Flamingo: A Visual Language Model for Few-Shot Learning", Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, Louisiana, USA, Vol. 35, pp. 23716-23736, Dec. 2022.

[7] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", Int. Conf. Mach. Learn. (ICML), Honolulu, Hawaiim, USA, Vol. 202, pp. 19730-19742, Jul. 2023.

[8] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond", arXiv preprint arXiv:2308.12966, pp. 1-30, Aug. 2023. https://doi.org/10.48550/arXiv.2308.12966.

[9] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision Transformers Need Registers", arXiv preprint arXiv:2309.16588, pp. 1-28, Sep. 2023. https://doi.org/10.48550/arXiv.2309.16588.

[10] M. Nikouei, B. Baroutian, S. Nabavi, F. Taraghi, A. Aghaei, A. Sajedi, and M. E. Moghaddam, "Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications", Intell. Syst. Appl., Vol. 27, Art. no. 200561, Sep. 2025. https://doi.org/10.1016/j.iswa.

2025.200561.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection", Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, pp. 2980-2988, Oct. 2017. https://doi.org/10.1109/ICCV.2017.324.

[12] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding Multimodal Large Language Models to the World", arXiv preprint arXiv:2306.14824, pp. 1-20, Jun. 2023. https://doi.org/10.48550/arXiv.2306.14824.

[13] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang,C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO:Marrying DINO with Grounded Pre-Training for Open-Set Object Detection", Eur. Conf. Comput. Vis. (ECCV), Milan, Italy, pp. 38-55, Sep.-Oct. 2024. https://doi.org/10.48550/arXiv.2303.05499.

[14] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V", arXiv preprint arXiv:2310.11441, pp. 1-20, Oct. 2023. https://doi.org/10.48550/arXiv.2310.11441.

[15] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, Z. Sun, D. Gutfreund, and C. Gan, "Visual Chain-of-Thought Prompting for Knowledge-based Visual Reasoning", Proc. AAAI Conf. Artif. Intell. (AAAI), Vancouver, Canada, Vol. 38, No. 2, pp. 1254-1262, Mar. 2024. https://doi.org/10.1609/aaai.v38i2.27888.

[16] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning", Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, Louisiana, USA, Vol. 36, pp. 34892-34916, Dec. 2023.

[17] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection", Proc. IEEE Int. Conf. Image Process. (ICIP), Bordeaux, France, pp. 966-970, Oct. 2022. https://doi.org/10.1109/ICIP46576.2022.9897990.

[18] J. Wang, Y. Wang, X. Zhang, L. Li, X. Wang, S. Lv, R. Zhu, and T. Li, "PRPSV: Parking Efficiency and Reservation Service Optimization Based on Parking Space View", IEEE Trans. Intell. Transp. Syst., Vol. 26, No. 8, pp. 11696-11711, Aug. 2025. https://doi.org/10.1109/TITS.2025.3574305.

[19] Raunge, "Aerial View Car Detection for YOLOv5", Kaggle, 2022. https://www.kaggle.com/datasets/braunge/aerial-view-car-detection-for-yolov5. [accessed: Jul. 22, 2026]

[20] UniqueData, "Parking Space Detection Dataset", Hugging Face, 2024. https://huggingface.co/datasets/UniqueData/parking-space-detection-dataset. [accessed: Jul. 22, 2026]

[21] S. Wang, Y. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, H. Liu, J. Lin, J. Wen, J. Zhou, and J. Yang, "Qwen2.5-VL Technical Report", arXiv preprint arXiv:2502.13923, pp. 1-41, Feb. 2025. https://doi.org/10.48550/arXiv.2502.13923.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, pp. 740-755, Sep. 2014. https://doi.org/10.1007/978-3-319-10602-1_48.

## 저자소개

최 권 택 (Kwon-Taeg Choi)

2006년 2월 : 연세대학교 컴퓨터공학과(공학석사)
2011년 2월 : 연세대학교 컴퓨터공학과(공학박사)
2016년 3월 ~ 현재 : 강남대학교 소프트웨어응용학부 교수
관심분야 : 가상현실, 증강현실, 모바일컴퓨팅, 기계학습, HCI