Beyond “How Many?”
– The Dawn of Referring Expression Counting

2025. 08. 14.

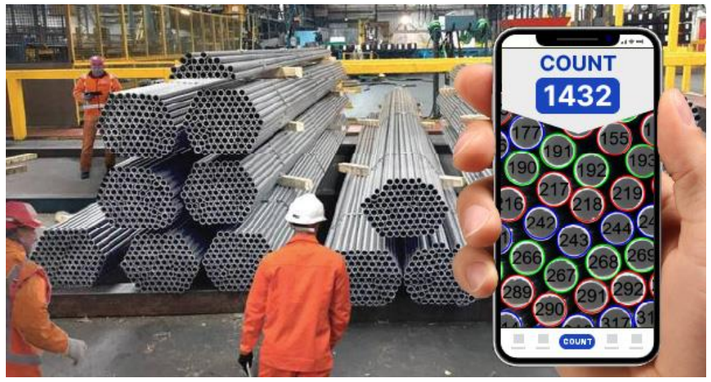
Presented By
강민석

Outline

- Object Counting: Few-shot and Zero-shot methods
- Referring Expression Counting
  - CVPR 2024
- Exploring Contextual Attribute Density in Referring Expression Counting
  - CVPR 2025

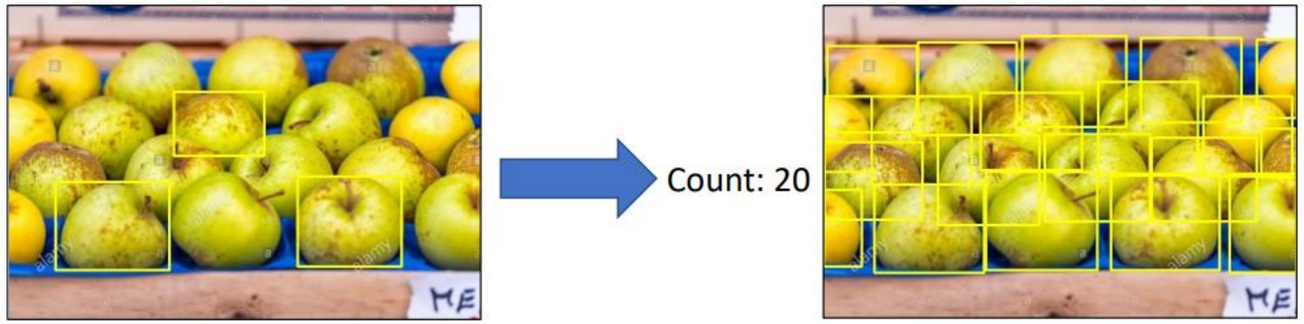
Object Counting: Few-Shot and Zero-Shot Methods

- Object Counting
  - 컴퓨터 비전의 중요한 작업 중 하나로, 이미지 또는 비디오에서 특정 클래스의 객체 수를 세는 것을 목표로 함
    - 사람, 자동차, 나무 등 특정 객체의 수를 파악하는 것이 필요한 다양한 상황에서 사용됨
  - 적용 사례
    - 상품 계수 및 불량품 탐지
    - 생산 부품 재고 관리
    - 불량률 모니터링 및 품질 향상
  - 접근 방법론
    - 일부 방법론은 먼저 객체 검출(object detection)을 수행하여 각 객체를 분리하고 이를 세는 방법을 사용
    - 다른 방법론은 객체의 밀도를 예측하고, 이를 사용하여 전체 객체 수를 추정하는 방법을 사용



Object Counting: Few-Shot and Zero-Shot Methods

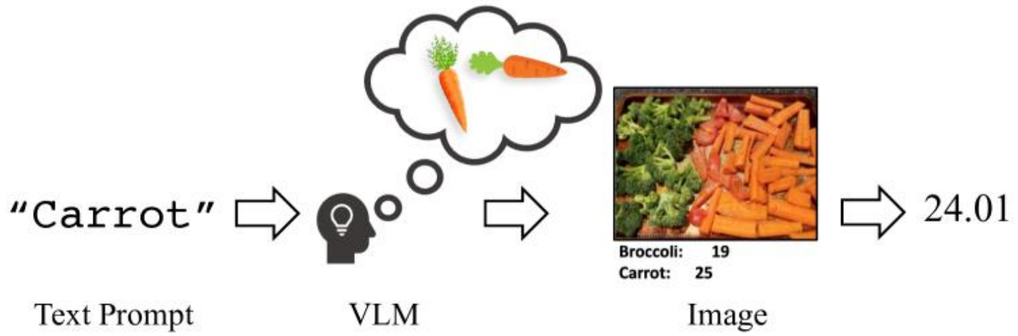
- Few-shot object counting의 정의
  - Few-shot object counting은 소수의 support 이미지로부터 제시된 exemplar 객체의 시각적 특징을 학습하여, query 이미지 내 동일한 객체의 개수를 예측하는 과제임
    - 아래 예시 그림의 경우, exemplar 객체의 bounding box가 3개이므로 3-shot임
  - Support 이미지에 제공된 exemplar 객체를 기준으로 query 이미지 내 유사한 객체의 위치와 수량을 추정함



< Few-shot object counting 예시 >

Object Counting: Few-Shot and Zero-Shot Methods

- Zero-shot object counting의 정의
  - Zero-shot object counting은 사전에 학습되지 않은 새로운 객체를 별도의 샘플 이미지(exemplar) 없이 텍스트 설명만으로 인식하고 수를 세는 기술
    - 아래 예시 그림의 경우, “carrot”이 text prompt로 적용되어, 이미지 내에서 계수를 진행함
  - 일반적으로 CLIP과 같은 Vision-Language Model (VLM)을 적용함

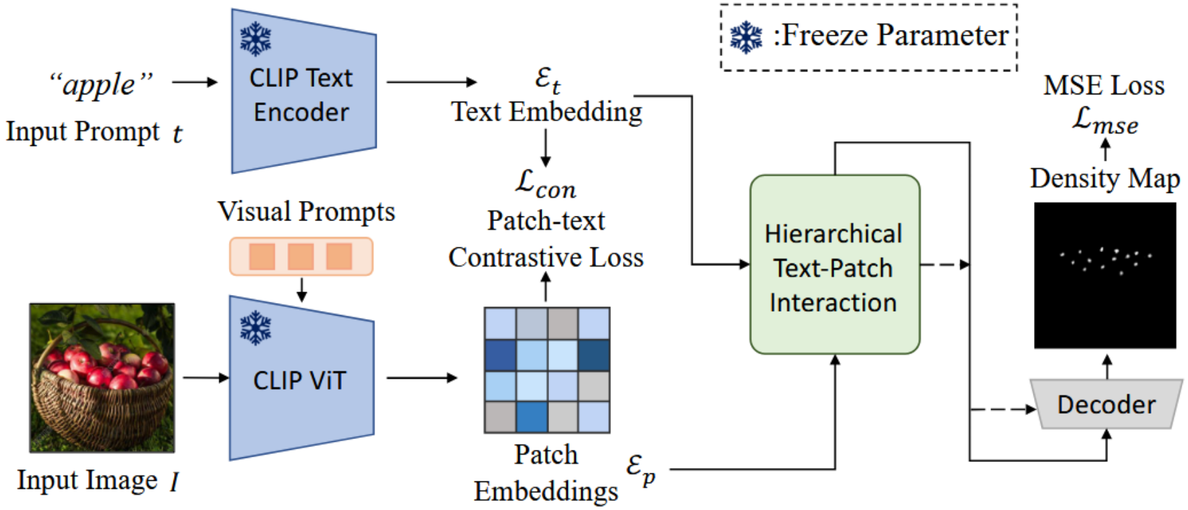


< Zero-shot object counting 예시 >

1) Jiang, Ruixiang, Lingbo Liu, and Changwen Chen. "Clip-count: Towards text-guided zero-shot object counting." Proceedings of the 31st ACM International Conference on Multimedia. 2023.
2) Jia, Menglin, et al. "Visual prompt tuning." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
3) Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." International Journal of Computer Vision 130.9 (2022): 2337-2348.

Object Counting: Few-Shot and Zero-Shot Methods

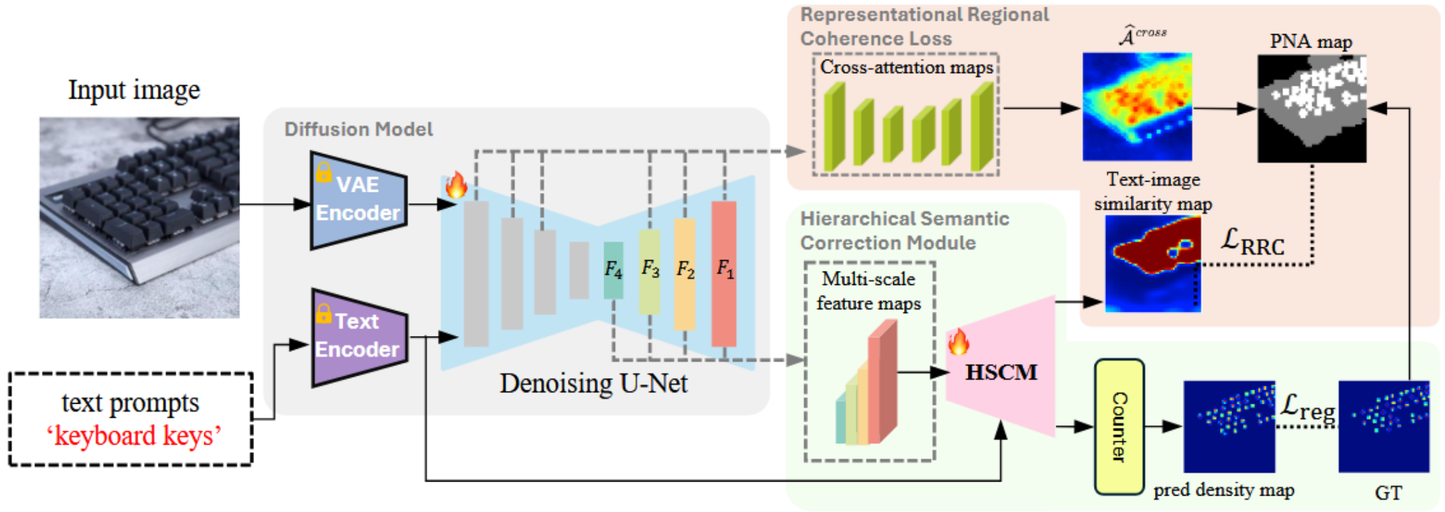
- 기존 zero-shot object counting 방법론 소개 – CLIP-count[1)]
  - CLIP에서 추출한 text 및 image 특징을 입력으로 하여, cross-modal attention을 통해 의미 기반 counting을 수행함
  - Patch feature와 text embedding 간의 상호작용을 위한 Feature Interaction Module (FIM)을 도입하여 의미적으로 관련된 시각 영역을 강조함
  - VPT (Visual Prompt Tuning)[2)], CoOp (Class-specific Prompt Tuning)[3)] 등 프롬프트 학습 방식으로 텍스트의 표현력을 개선
  - Counting에 영향을 미치는 패치 영역을 식별하고 강조하기 위해, cross attention 기반 Spatial Layout Map (SLM)을 활용하여 밀도 맵을 보정함



< CLIP-Count의 전체 architecture >

Object Counting: Few-Shot and Zero-Shot Methods

- 기존 zero-shot object counting 방법론 소개 – T2ICount[1)]
  - 텍스트 프롬프트 (ex. "keyboard keys")를 통해 특정 클래스의 의미적 단서를 제공하고, 이를 기반으로 counting을 보조함
  - Diffusion 기반 데이터 증강 모듈과 함께, VAE 및 Transformer 텍스트 인코더 등을 활용하여 다양하고 강건한 이미지-텍스트 조합을 생성함
  - Semantic Correction Module 및 Text-image Regional Coherence Loss 등 다양한 방법론을 활용하여 멀티모달 정합성을 높이고, 밀도 맵의 일관성을 확보함

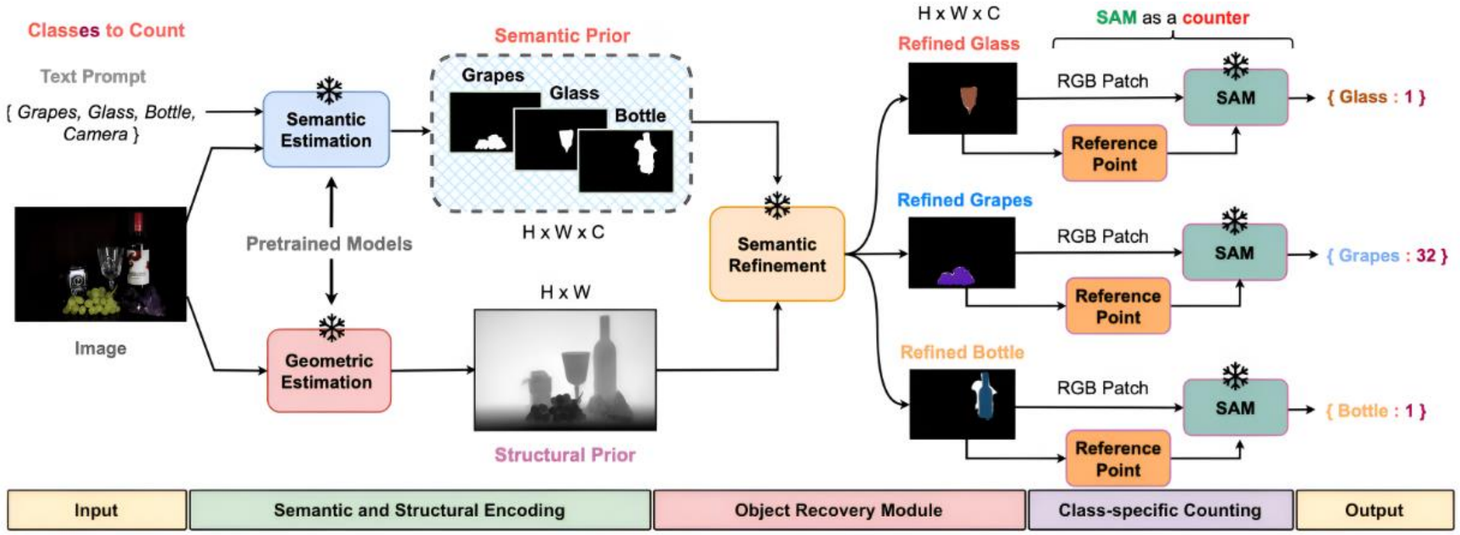


< T2ICount의 전체 architecture >

1) Mondal, Anindya, et al. "Omnicount: Multi-label object counting with semantic-geometric priors." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 18. 2025.

Object Counting: Few-Shot and Zero-Shot Methods

- 기존 zero-shot object counting 방법론 소개 – OmniCount[1)]
  - 다양한 데이터 표현 (점, 바운딩 박스, 질의 응답 등)을 통합할 수 있는 통합 멀티모달 counting 프레임워크를 제안함
  - 이미지 내 복수 클래스에 대한 counting 수행을 위해 Text Query Embedding과 Region-level Feature Matching 기법을 활용함
  - Image, Point, Box, Text를 모두 포함하는 4가지 supervision signal을 활용하여 멀티 supervision 기반의 견고한 학습 구조 구성



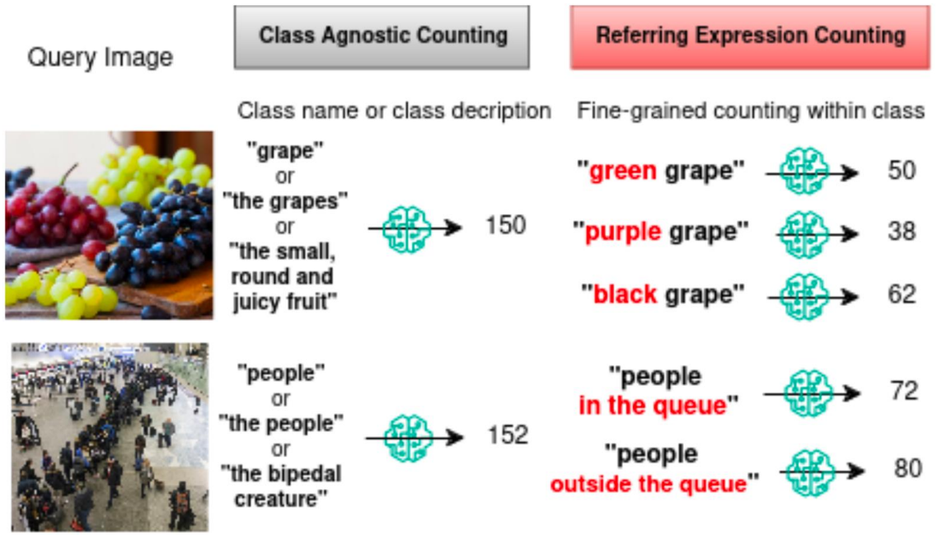
< OmniCount의 전체 architecture >

- **Referring Expression Counting [CVPR 2024]**

- **Exploring Contextual Attribute Density in Referring Expression Counting [CVPR 2025]**

Referring Expression Counting

- 기존의 class-agnostic counting 기술의 한계
  - 기존의 객체 counting은 클래스 단위(e.g., “사람”, “포도”)로만 수행 가능했음
    - “줄 서 있는 사람”이나 “초록색 포도”와 같이 동일 클래스 내의 세분화된 속성을 구분하여 counting하는 것은 불가능함
  - 이는 실제 산업 현장에서 요구하는 구체적인 정량 분석에 한계를 보임
- 새로운 task를 제안 – Referring Expression Counting (REC)
  - 사용자가 자연어(referring expression)로 제시하는 구체적인 속성을 이해하고 해당 객체만 세는 task를 제안
  - “왼쪽 그릇에 있는 빨간 사과”와 같이 동일 클래스 내 객체를 속성에 따라 구분하여 세는 것이 가능해짐



1) Dai, Siyang, Jun Liu, and Ngai-Man Cheung. "Referring expression counting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
2) Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3394–3403, 2021.

Referring Expression Counting

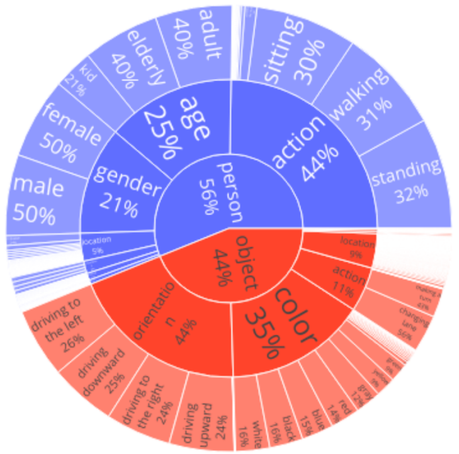
- 신규 벤치마크 데이터셋 구축 – REC-8K
  - REC task는 밀집된 객체 환경에서 다양한 속성을 구분해야 하는 어려움이 있어, 이를 평가하기 위한 새로운 벤치마크 필요
    - 이를 위해 REC task를 위한 전용 벤치마크 데이터셋인 REC-8K를 직접 구축함
  - FSC-147[2] 등 기존 counting 데이터셋과 Pixabay 같은 사진 공유 웹사이트 등 다양한 소스에서 이미지를 수집
    - 이미지에 보이는 객체 속성을 기반으로 새로운 Referring Expression (RE) 라벨을 생성함
    - 객체 중앙 또는 사람 머리 중앙에 점을 찍는 point annotation 방식을 사용했음

1) Dai, Siyang, Jun Liu, and Ngai-Man Cheung. "Referring expression counting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
2) Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3394–3403, 2021.

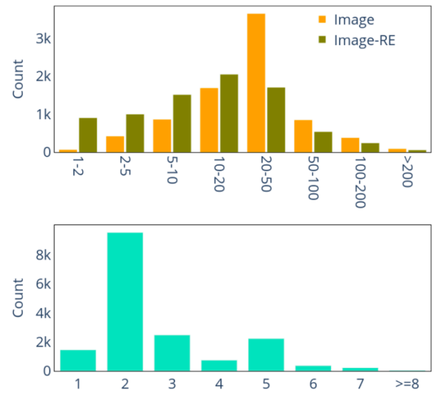
Referring Expression Counting

- 신규 벤치마크 데이터셋 구축 – REC-8K
  - REC-8K의 통계
    - 전체 규모: 총 8,011개의 이미지와 286,621개의 point annotation을 포함
    - 샘플 단위: (Image, RE (Referring Expression)) 쌍을 하나의 학습 샘플로 취급하며, 총 17,122개의 (Image, RE) 쌍으로 구성
    - 객체 수: 이미지 당 최소 1개, 평균 36개, 최대 1,028개의 객체가 포함됨
      - (Image, RE) 쌍 당 타겟 객체 수는 최소 1개, 평균 17개, 최대 1,004개
  - 데이터 독립성: 분할된 set 간에 일부 RE가 공유되기도 하지만(train-val 80개, train-test 79개) , 각 set에 포함된 이미지는 서로 중복되지 않음

Dataset	Text Anno.	Classes	REs	No. of images	Avg. points
JHU-Crowd [39]	×	1	-	4372	345
NWPU [48]	×	1	-	5109	417
UCF CC 50 [10]	×	1	-	50	1279
CARPK [9]	×	1	-	1448	62
FSC-147 [36]	×	147	-	6135	56
REC-8K (ours)	✓	-	**1182**	**8011**	**36**



(a) Percentages of person & object categories, attribute types and attributes.



(b) Histogram of points per image, per Image-RE pair (up) and words per RE (down).

1) Dai, Siyang, Jun Liu, and Ngai-Man Cheung. "Referring expression counting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
2) Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3394–3403, 2021.

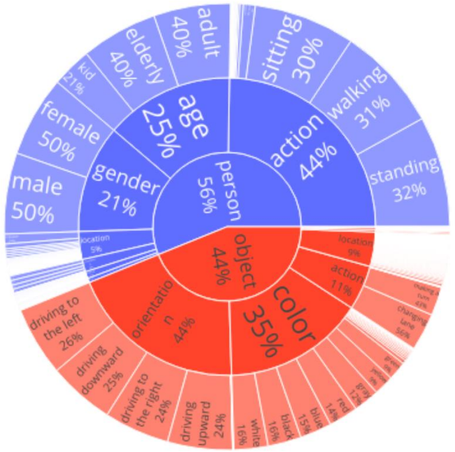
Referring Expression Counting

- 신규 벤치마크 데이터셋 구축 – REC-8K

JSON Raw Data Headers

Save Copy Collapse All Expand All (slow) Filter JSON


```
▼ 0000-fsc147-1106.jpg:
  ▼ red pen:
      class:                "pen"
      attribute:            "red"
    ▼ points:
        ▶ 0:                [ 155.36, 48.55 ]
        ▶ 1:                [ 361.23, 80.54 ]
        ▶ 2:                [ 357.85, 179.2 ]
        ▶ 3:                [ 263.4, 64.0 JS: 64 ]
        ▶ 4:                [ 211.98, 217.61 ]
        ▶ 5:                [ 184.61, 221.41 ]
        ▶ 6:                [ 149.63, 226.73 ]
        ▶ 7:                [ 289.35, 73.04 ]
        ▶ 8:                [ 334.92, 29.05 ]
        ▶ 9:                [ 204.38, 185.28 ]
        ▶ 10:               [ 77.45, 201.28 ]
        ▶ 11:               [ 338.07, 83.1 ]
        ▶ 12:               [ 312.65, 78.86 ]
        ▶ 13:               [ 239.46, 180.12 ]
        ▶ 14:               [ 245.65, 211.08 ]
      type:                 "color"
  ▶ purple pen:             { class: "pen", attribute: "purple", type: "color", … }
  ▶ green pen:              { class: "pen", attribute: "green", type: "color", … }
  ▶ yellow pen:             { class: "pen", attribute: "yellow", type: "color", … }
▶ 0001-fsc147-1306.jpg:     { "black shoe": {…}, "blue shoe": {…}, "white shoe": {…} }
▶ 0002-fsc147-1386.jpg:     { "pill in the blister pack": {…}, "blue pill": {…}, "yellow pill": {…}, … }
```

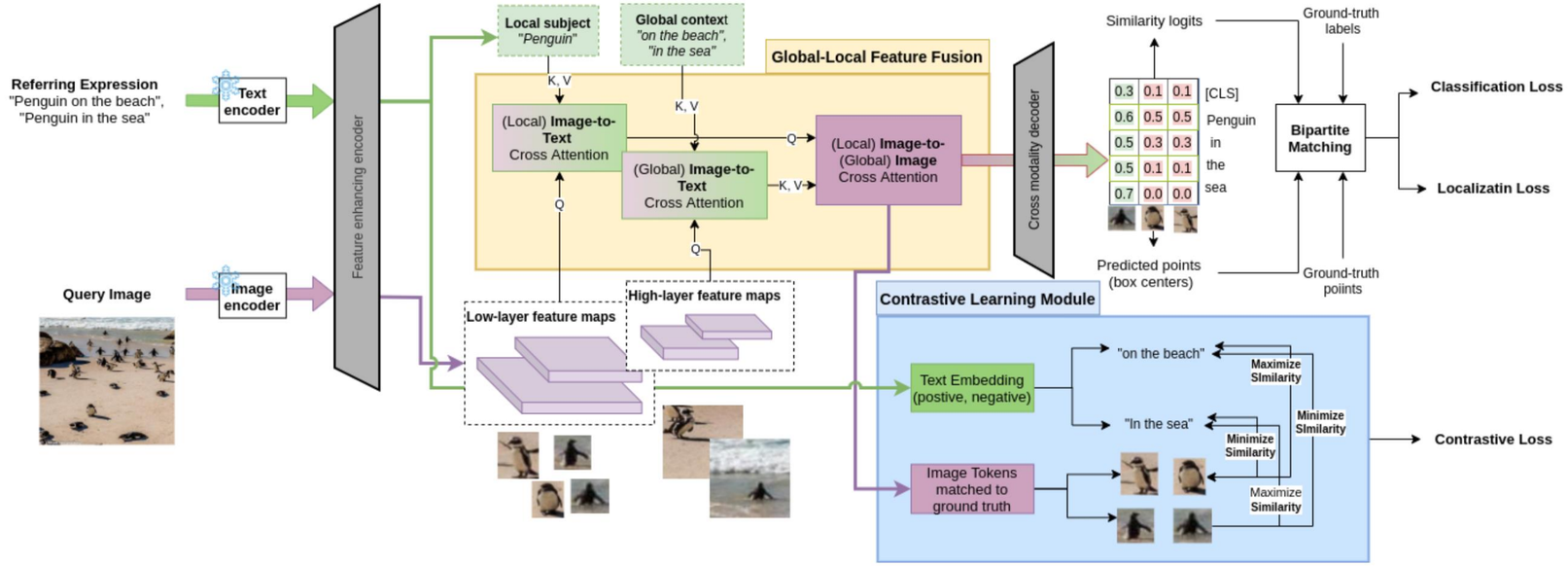



(a) Percentages of person & object categories, attribute types and attributes.

1) Dai, Siyang, Jun Liu, and Ngai-Man Cheung. "Referring expression counting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Referring Expression Counting

- REC task를 위해 새롭게 제안한 모델 – GroundingREC
  - 강력한 오픈셋(open-set) 객체 탐지기인 GroundingDino[2]를 base model로 채택하여 GroundingREC를 개발
    - 기존의 bounding box 예측 대신, 객체의 중심 point를 예측하도록 모델을 수정하여 counting task에 맞게 조정
  - Global-Local Feature Fusion 모듈
    - "왼쪽 선반 위"와 같은 관계적 속성(relational attribute)을 더 잘 이해하기 위해 제안함
  - Contrastive Learning 모듈
    - 같은 클래스 내의 다른 속성을 명확히 구분하기 위해 Contrastive Learning 모듈을 도입함



Referring Expression Counting

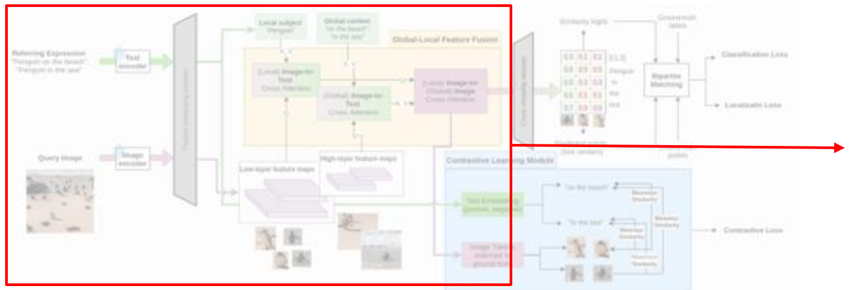
- REC task를 위해 새롭게 제안한 모델 – GroundingREC
  - Global and Local Feature Fusion Module
    - 이 논문에서 RE label의 attributes들은 2개의 category로 나뉨
      - Local Attributes: color, material, age, gender와 같이 이미지의 local crop으로부터 알 수 있는 정보
      - Relational Attributes: location이나 relative size와 같이 context와 이미지의 전반적인 이해를 필요로 하는 정보
    - Base model (GroundingDINO)는 text input의 각각의 object에만 집중하도록 학습되어 있으므로 local attribute에 집중됨
      - 우리는 이미지의 global한 정보를 필요로 하므로 relational attribute를 모델이 잘 파악하도록 설계됨
    - 우선, input referring expression을 두 파트로 나눔
      - **Local subject** part: object with any local attribute
      - **Global context** part: relational attribute

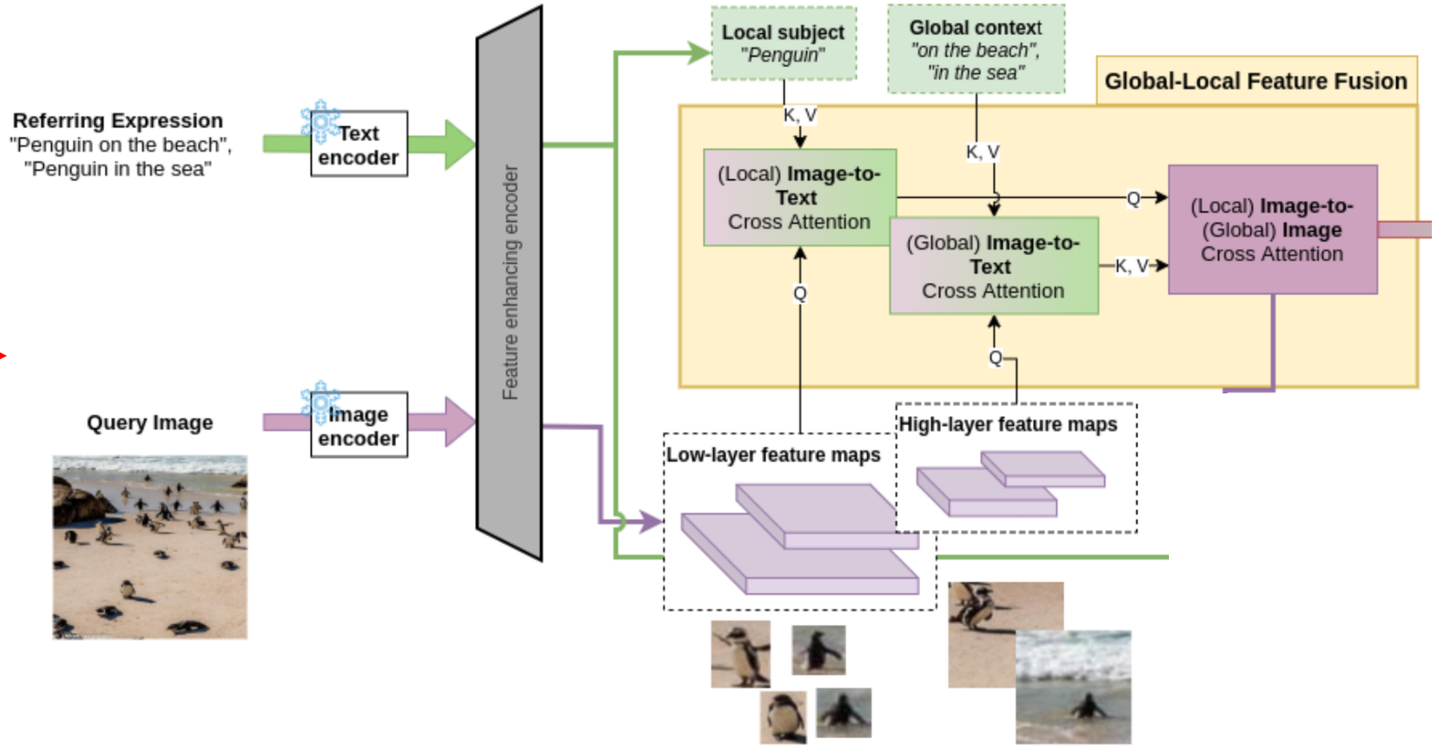
"Red apple in the left bowl."

"Penguin on the beach."

Referring Expression Counting

- REC task를 위해 새롭게 제안한 모델 – GroundingREC
  - Global and Local Feature Fusion Module
    - Image encoder의 다른 layer에서 뽑은 image feature들은 query image 내에서 서로 다른 receptive field를 가지고 있음
      - *Low-layer* image tokens는 local subject에, *high-layer* image tokens는 global context에 집중하도록 구현함
    - 첫째, local/global Image-to-Text Cross Attention을 통해 feature fusion를 수행함
    - 둘째, text 정보가 주입된 low-layer feature와 high-layer feature를 다시 한번 fusion하여 "이 지역적 특징의 전역적 맥락은 무엇인가?"를 학습함

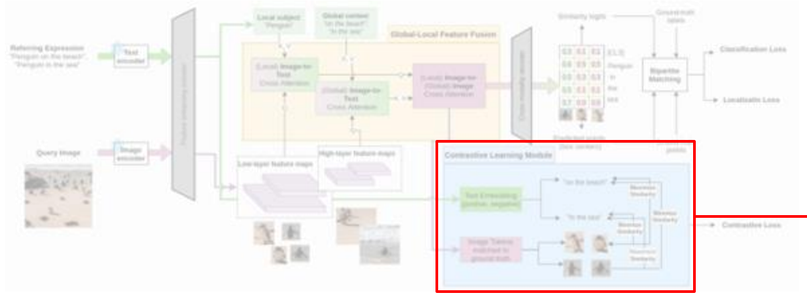


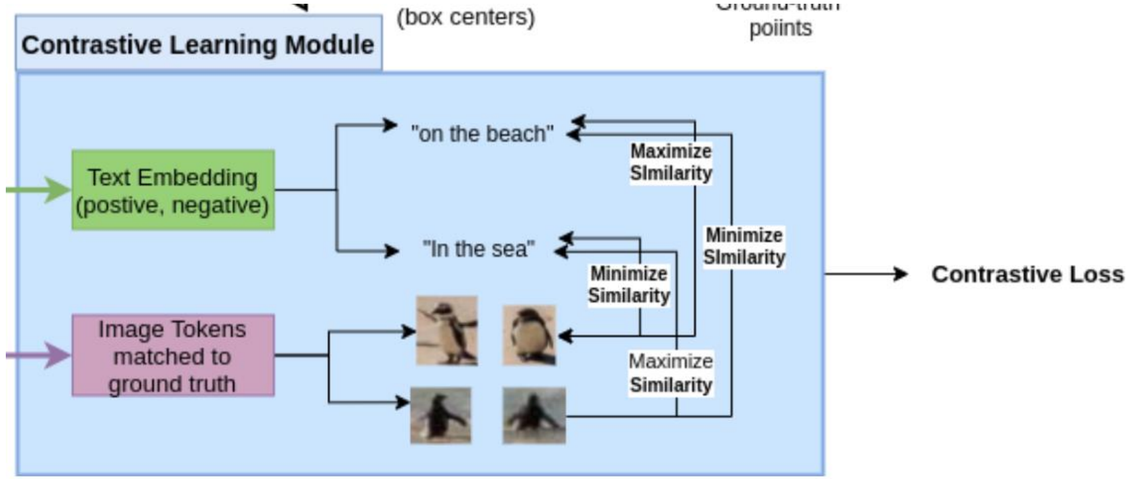


Referring Expression Counting

- REC task를 위해 새롭게 제안한 모델 – GroundingREC
  - Contrastive Learning Module
    - 같은 class 내에서도 다른 속성 차이를 명확히 구분하도록 학습하는 과정
    - 학습 시 모델에게 하나의 이미지를 보여주면서, 그 이미지에 해당하는 여러 개의 다른 Referring Expression을 한번에 사용
    - 모델은 이미지 내에서 각 RE에 해당하는 객체들의 위치를 예측하고, 이 위치와 실제 정답 위치를 Bipartite Matching 알고리즘으로 연결함
      - Positive sample: 해당 이미지 토큰과 동일한 속성의 text 정보가 긍정 샘플이 됨
      - Negative sample: 동시에 입력되었던 다른 속성들은 부정 샘플이 됨

$$\mathcal{L}_{contrast} = -\frac{1}{K}\sum_{k=1}^{K}[\log(s_{ik}) + \sum_{j\neq i}\log(1-s_{jk})]$$



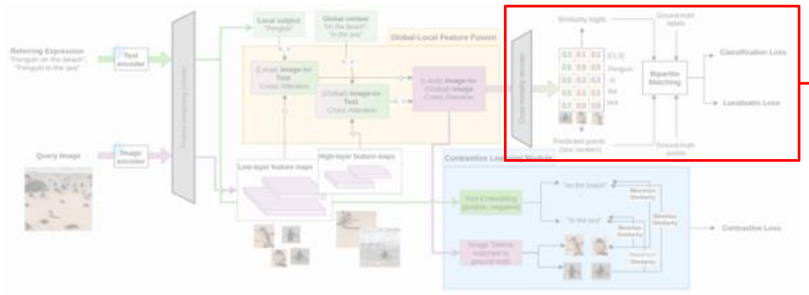


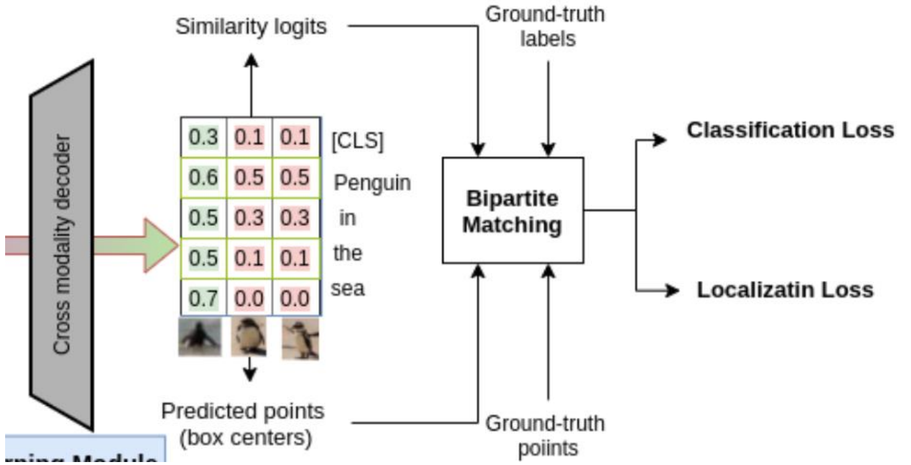
Referring Expression Counting

- REC task를 위해 새롭게 제안한 모델 – GroundingREC
  - 최종 loss function: $\mathcal{L} = \mathcal{L}_{loc} + \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{contrast}$
    - DETR-like object detector를 REC에도 적용하여, Bipartite Matching 알고리즘으로 predicted points를 GT points에 매칭함
    - Localization loss $\mathcal{L}_{loc}$: predicted point와 GT point 간의 L1 distance를 이용한 point regression loss

      $\mathcal{L}_{loc} = \frac{1}{K} \sum_{k=1}^{K} \|\hat{p}_k - p_k\|$
    - Cross-entropy classification loss $\mathcal{L}_{cls}$: 모델이 이미지에서 찾아낸 하나의 predicted image token이 주어진 text token과 얼마나 잘 부합하는지를 정량적으로 측정

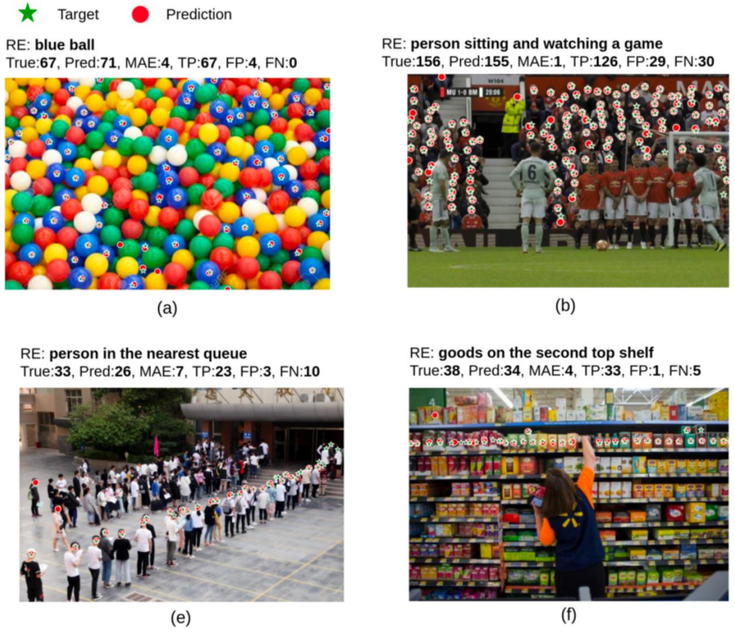
      $\mathcal{L}_{cls} = \frac{1}{K} \sum_{k=1}^{K} \left[-\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log \hat{y}_i^k + (1 - y_i) \log(1 - \hat{y}_i^k) \right] \right]$





Referring Expression Counting

- 실험 결과 및 성능
  - 제안 모델 GroundingREC는 REC-8K 벤치마크에서 기존 방법들 대비 SOTA(State-of-the-art) 성능을 달성했음
    - 아래의 qualitative results는 구체적인 text prompt에도 정확한 계수 성능을 보여줌
  - 기존의 class-agnostic counting 벤치마크인 FSC-147에서도 zero-shot 설정으로 SOTA 성능을 달성하며 모델의 높은 일반화 성능을 증명함



< Qualitative Results >

Method	Setting	Val set		Test set	
		MAE	RMSE	MAE	RMSE
ZSC [54]	zero-shot	26.93	88.63	22.09	115.17
CounTX [2]	zero-shot	17.10	65.61	15.88	106.29
TFOC [62] (text prompt)	zero-shot	47.21	127	24.79	137.15
GDino [21] (w/o finetune)	zero-shot	51.11	101.28	54.40	92.36
GDino [21] (w/ finetune)	zero-shot	10.32	55.54	10.82	104.00
GroundingREC (ours)	zero-shot	**10.06**	**58.62**	**10.12**	**107.19**

< 제안한 REC-8K 데이터셋에서의 zero-shot 성능 >

Method	Backbone	Finetuning	Val set					Test set				
			MAE↓	RMSE↓	Prec↑	Rec↑	F1↑	MAE↓	RMSE↓	Prec↑	Rec↑	F1↑
Mean	-	-	14.28	27.75	-	-	-	13.75	25.91	-	-	-
ZSC [54]	ResNet-50	✓	14.84	31.30	-	-	-	14.93	29.72	-	-	-
ZSC [54]	Swin-T	✓	12.96	26.74	-	-	-	13.00	29.07	-	-	-
TFOC [62]	ViT-B	-	16.08	31.61	0.30	0.07	0.12	17.27	32.68	0.23	0.07	0.11
CounTX [2]	ViT-B-16	✓	11.88	27.04	-	-	-	11.84	25.62	-	-	-
GroundingDino [21]	Swin-T	×	11.77	28.6	0.57	0.25	0.34	11.71	26.97	0.59	0.25	0.35
GroundingDino [21]	Swin-T	✓	9.03	21.98	0.56	0.76	0.65	8.88	21.95	0.59	0.76	0.66
GroundingREC (ours)	Swin-T	✓	**6.80**	**18.13**	**0.65**	**0.71**	**0.68**	**6.50**	**19.79**	**0.67**	**0.72**	**0.69**

< FSC-147 데이터셋에서의 zero-shot 성능 >

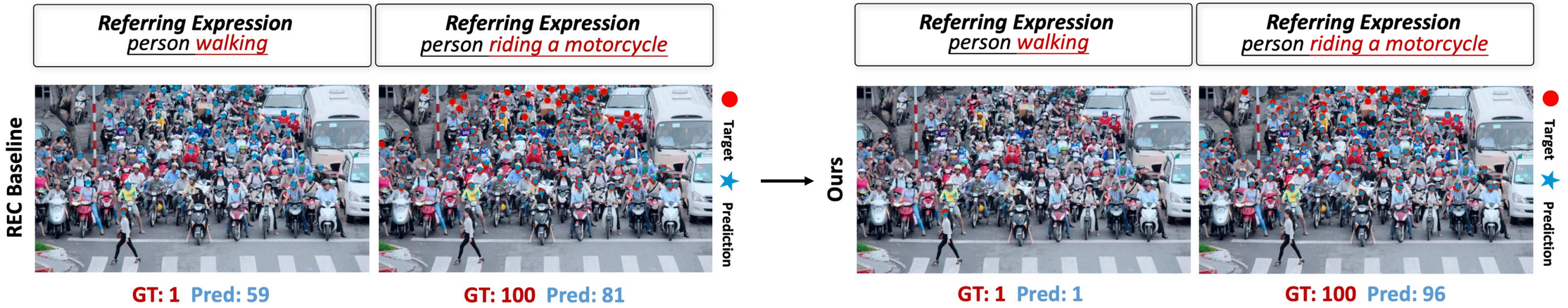
- Referring Expression Counting [CVPR 2024]

- **Exploring Contextual Attribute Density in Referring Expression Counting [CVPR 2025]**

1) Wang, Zhicheng, et al. "Exploring Contextual Attribute Density in Referring Expression Counting." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

Exploring Contextual Attribute Density in Referring Expression Counting

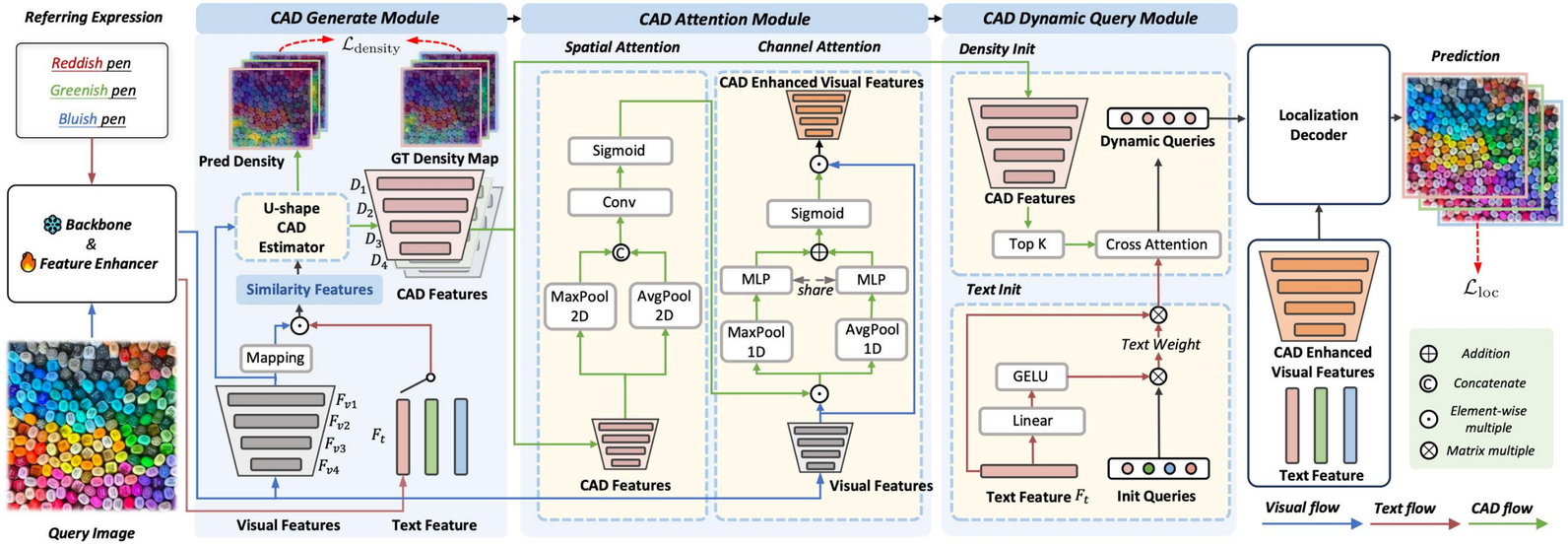
- 기존 REC 기술의 한계
  - REC는 세분화된 속성을 이해하고 counting하는 것을 목표로 하지만, 기존 방법론인 GroundingREC은 속성 정보와 시각적 패턴을 정확히 정렬하는 데 어려움을 겪음
    - Over-counting: 모델이 '걷고 있는'과 같은 세분화된 속성 대신 '사람'이라는 클래스 정보에 과도하게 집중하여, 엉뚱한 객체까지 카운트에 포함시키는 문제가 발생
    - Under-counting: 객체가 가려지거나 크기 변화가 심할 경우, 주어진 속성을 가진 객체를 놓치는 문제가 발생함
- 기존 counting 연구에서 visual density의 중요성은 입증되었으나, REC에서는 "문맥적 속성 밀도 (Contextual Attribute Density, CAD)" 개념이 간과되었음
  - 본 논문은 CAD를 "특정 세분화된 속성이 이미지 영역 내에 얼마나 밀집되어 있는가" 로 정의함



< 기존 GroundingREC (왼쪽)의 over/under-counting 문제 및 해당 논문 (오른쪽)의 개선 결과 >

Exploring Contextual Attribute Density in Referring Expression Counting

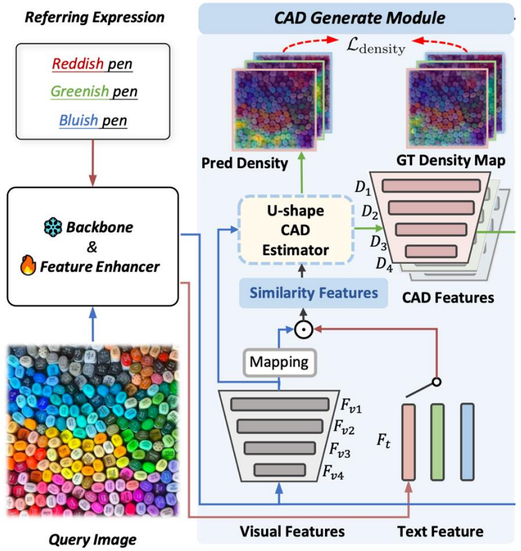
- 새롭게 제안한 프레임워크 – CAD-GD (Contextual Attribute Density Aware Grounding DINO)
  - CAD Generate Module: U-Net 형태의 CAD Estimator를 통해 텍스트와 다중 스케일 이미지 특징을 융합하여, 특정 속성이 밀집된 영역을 나타내는 CAD 특징을 생성함
  - CAD Attention Module: 공간 및 채널 어텐션을 활용해 생성된 CAD 특징을 기존 이미지 특징에 주입함으로써, 특정 속성 영역을 강조하고 속성 간 구별 능력을 향상
  - CAD Dynamic Query Module: 고정된 쿼리 대신 텍스트와 CAD 특징을 결합하여 쿼리를 동적으로 초기화함으로써, 입력 표현에 따라 쿼리 특징이 유연하게 변화하고 명확하게 구별되도록 함



< CAD-GD의 전체 프레임워크 >

Exploring Contextual Attribute Density in Referring Expression Counting

- CAD-GD – CAD Generate Module
  - 이 모듈의 목표는 입력된 텍스트 표현에 해당하는 CAD feature를 생성하는 것임
    - 먼저, visual feature와 text feature 간의 유사도를 계산하여 text와 관련된 이미지 영역에 집중하도록 유도함
    - Similarity feature와 visual feature를 U-Net 구조의 CADE에 입력하여 다양한 scale의 CAD feature와 density map 출력
    - 해당 density map을 GT density map과의 L2 loss를 계산하여 지도 학습 방식으로 진행함
      - GT density map은 referring expression별 GT points에 sigma 15의 Gaussian kernel을 입혀서 제작함



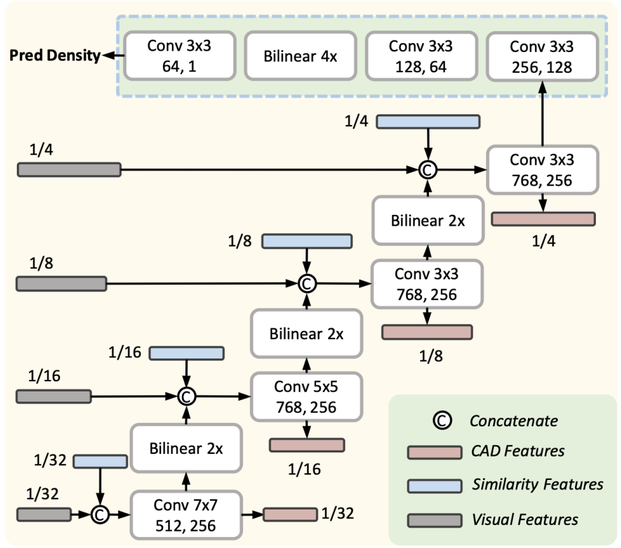
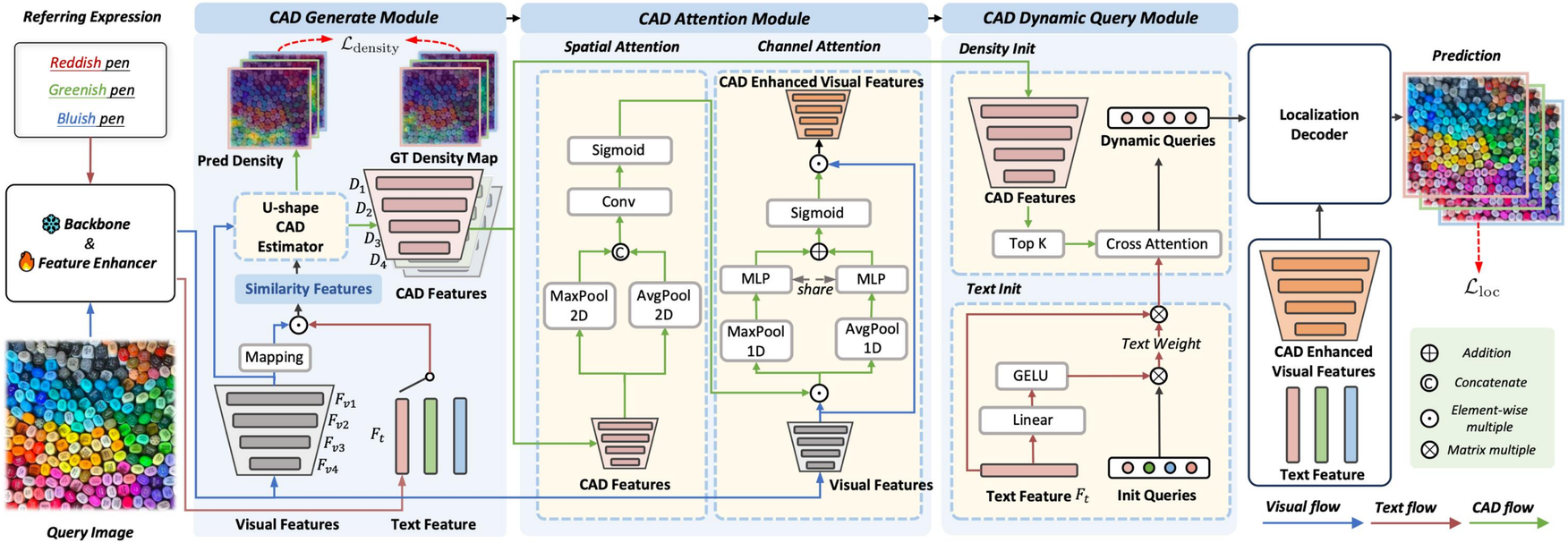


Figure 6. **U-shape CAD Estimator.** The visual features and the similarity features are sent into the U-shape CAD Estimator to obtain the CAD features. Each convolution block, in which the left and right dimensions are input and output dimensions separately, is followed by a ReLU as the activation function.

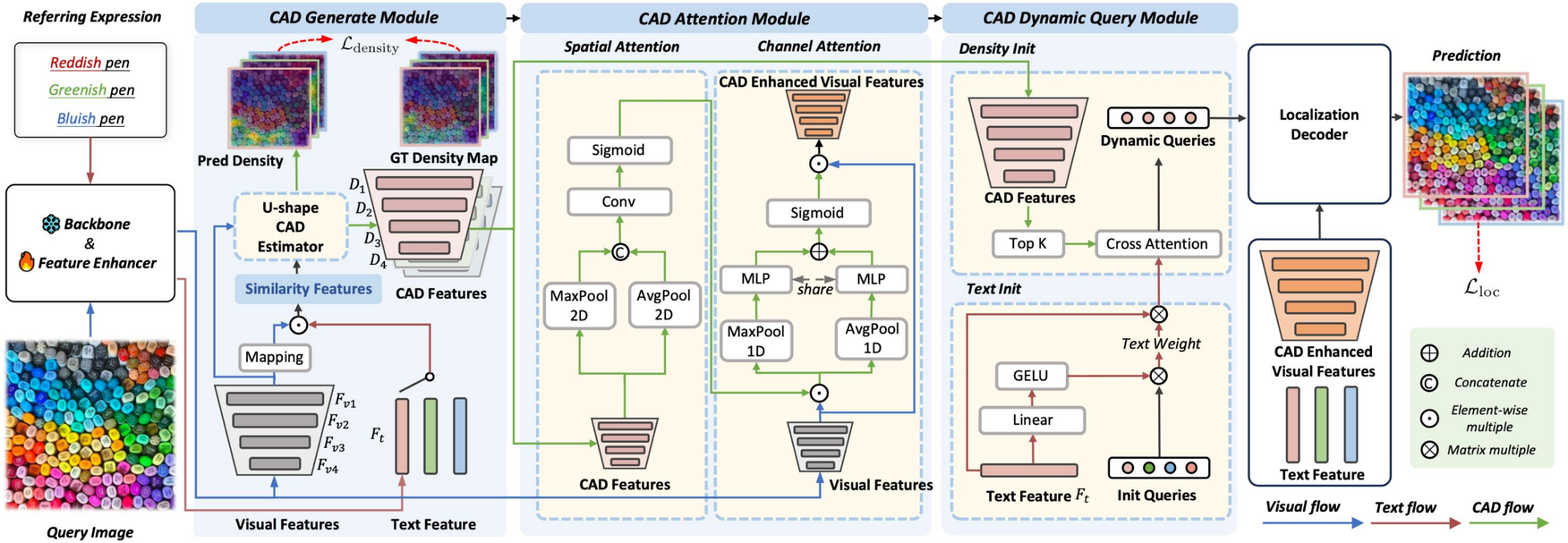
Exploring Contextual Attribute Density in Referring Expression Counting

- CAD-GD – CAD Attention Module
  - Spatial 및 channel attention을 활용해 생성된 CAD feature을 기존 visual feaeture에 주입함으로써, 특정 속성 영역을 강조하고 속성 간 구별 능력을 향상
    - Spatial Attention: 입력된 text에 해당하는 객체가 이미지의 어느 영역(위치)에 있을지 CAD feature를 통해 파악함
      - 이 정보를 바탕으로, 해당 영역의 신호는 강하게 만들고, 관련 없는 배경의 신호는 약하게 만듦
    - Channel Attention: 같은 class의 객체 내에서도 미세한 attribute 차이를 구별하기 위해, visual feature의 여러 채널 중에서 미세한 속성 차이를 구분하는 데 중요한 채널에 더 높은 가중치를 부여함



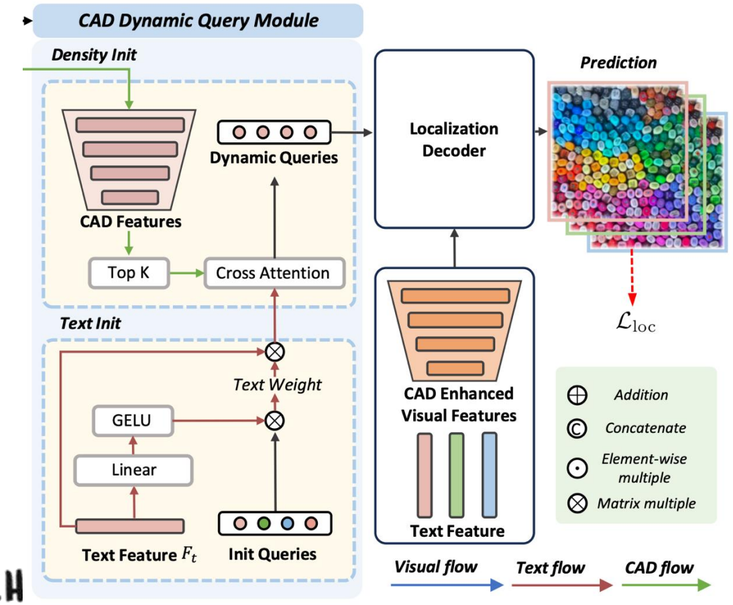
Exploring Contextual Attribute Density in Referring Expression Counting

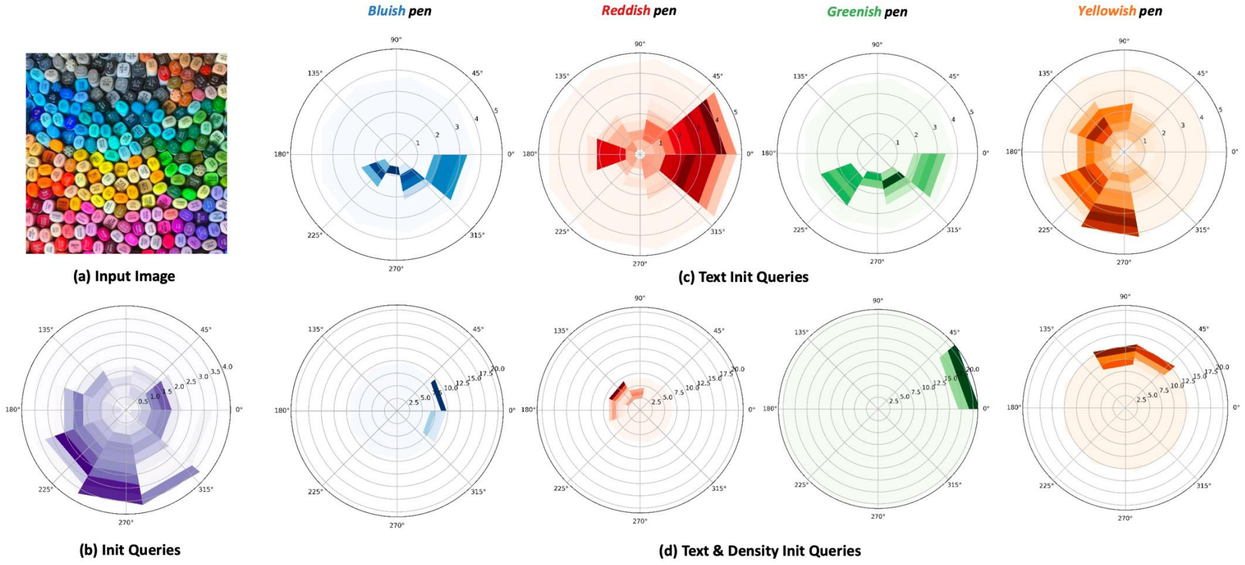
- CAD-GD – CAD Dynamic Query Module
  - 위치 정보와 content feature로 구성된 query는 localization decoder의 또 다른 필수적인 요소임
  - 초기 query 위치를 정하기 위해, text feature와 가장 높은 유사도 점수를 갖는 top-K의 visual feature 위치를 선택하는 GroundingDINO의 전략을 사용함
    - GroundingDINO와 같은 static query embedding 방식이 REC의 이상적인 초기화 방식이 아니라고 주장
    - CAD feature와 text feature를 모두 활용한 dynamic query를 제안함



Exploring Contextual Attribute Density in Referring Expression Counting

- CAD-GD – CAD Dynamic Query Module
  - Text Init: 아무 정보가 없는 초기 힌트(query)에 입력된 텍스트의 의미를 불어넣어주는 초기화 단계
    - Text feature를 학습 가능한 행렬과 곱하고 GELU를 통과시켜 동적 text 특징을 만든 후 text feature와 결합하여 text-dynamic init query $\dot{Q} = (Q \times (F_t \times M)^\top) \times F_t$를 만듦
  - Density Init: 여전히 이미지 정보를 전혀 모르는 query에 실제 이미지 정보를 불어넣어 완성하는 단계
    - 우선, text feature와 가장 높은 similarity score를 갖는 top-K개의 query position의 CAD feature D_K를 얻음
    - Text init을 거친 query와 CAD features D_K를 cross attention을 통해 융합함

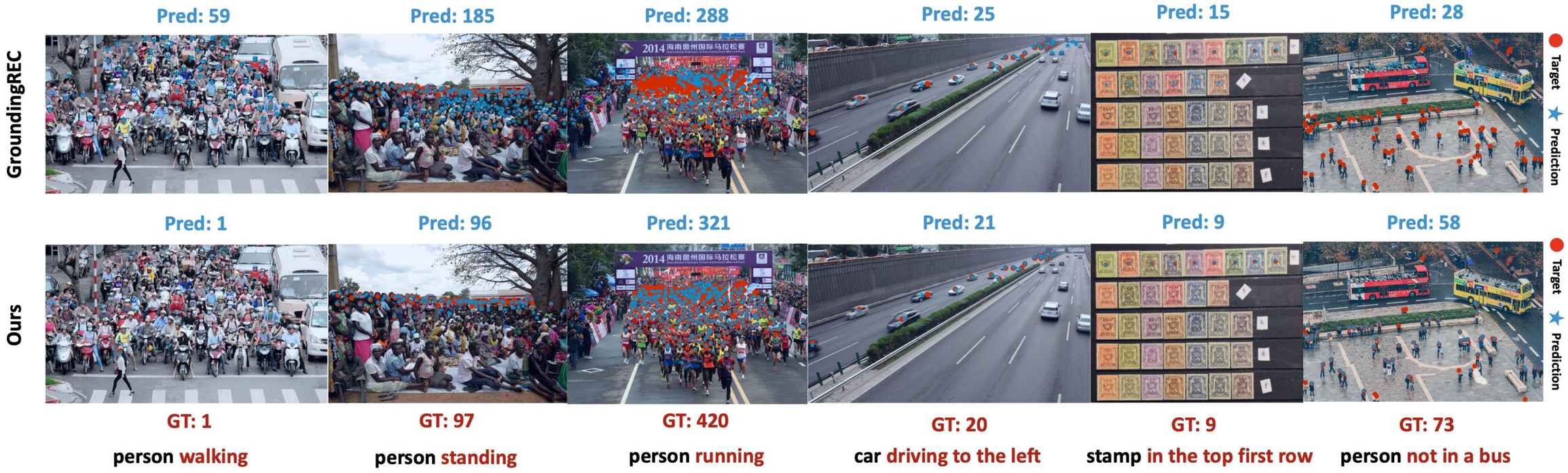




< Init query, text-init query, text&density-init query를 t-SNE로 polar coordinate에 시각화 >

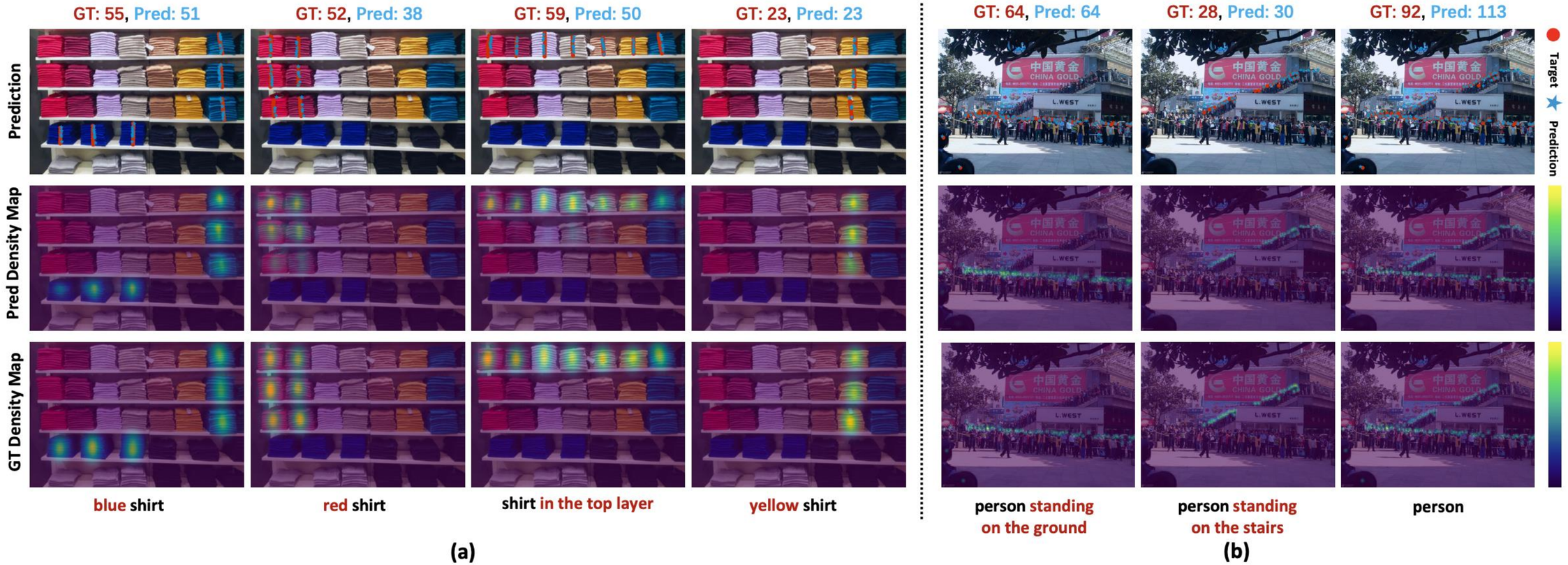
Exploring Contextual Attribute Density in Referring Expression Counting

- 실험 결과
  - Qualitative Results



Exploring Contextual Attribute Density in Referring Expression Counting

- 실험 결과
  - Contextual Attribute Density (CAD) map 시각화



Exploring Contextual Attribute Density in Referring Expression Counting

- 실험 결과
  - 주요 데이터셋: REC-8K (REC 전용), FSC-147 및 CARPK (일반화 성능 검증)
  - REC-8K 데이터셋 실험 결과 (왼쪽)
    - CAD-GD는 기존 SOTA 모델인 GroundingREC 대비 counting 에러를 30% 이상 크게 감소시켰음
  - 일반화 성능 실험 결과 (오른쪽)
    - Zero-shot 설정으로 평가한 FSC-147 및 CARPK 데이터셋에서도 기존 SOTA 모델들을 능가하는 성능을 보임

Method	Backbone	Val set					Test set				
		MAE ↓	RMSE ↓	Prec ↑	Rec ↑	F1 ↑	MAE ↓	RMSE ↓	Prec ↑	Rec ↑	F1↑
Mean	-	14.28	27.75	-	-	-	13.75	25.91	-	-	-
ZSC [38]	Res-50	14.84	31.30	-	-	-	14.93	29.72	-	-	-
ZSC [38]	Swin-T	12.96	26.74	-	-	-	13.00	29.07	-	-	-
CounTX [2]	ViT-B	11.88	27.04	-	-	-	11.84	25.62	-	-	-
GroundingDINO [23]	Swin-T	9.03	21.98	0.56	0.76	0.65	8.88	21.95	0.59	0.76	0.66
GroundingREC [8]	Swin-T	6.80	18.13	0.65	0.71	0.68	6.50	19.79	0.67	0.72	0.69
CAD-GD (ours)	Swin-T	5.43	15.01	0.68	0.72	0.70	5.29	17.08	0.71	0.73	0.72
CAD-GD† (ours)	Swin-T	4.58	13.24	0.68	0.71	0.70	4.59	14.68	0.72	0.70	0.71
GroundingREC* [8]	Swin-B	5.66	15.24	0.66	**0.77**	0.71	5.42	18.47	0.71	0.69	0.70
CAD-GD (ours)	Swin-B	4.83	13.52	0.74	0.76	**0.75**	4.94	14.65	0.75	**0.77**	**0.76**
CAD-GD† (ours)	Swin-B	**4.23**	**13.14**	**0.76**	0.70	0.73	**4.34**	**12.93**	**0.77**	0.71	0.74

Method	Val		Test	
	MAE	RMSE	MAE	RMSE
CounTR [22]	13.13	49.83	11.95	91.23
LOCA [35]	10.24	32.56	10.79	56.97
CACViT [37]	10.63	37.95	9.13	48.96
DAVE [29]	8.91	28.08	8.66	32.36
CountGD [37]	7.10	26.08	5.74	24.09
Patch-selection [38]	26.93	88.63	22.09	115.17
CLIP-count [17]	18.79	61.18	17.78	106.62
VLCounter [18]	18.06	65.13	17.05	106.16
CounTX [2]	17.10	65.61	15.88	106.29
DAVE [29]	15.48	52.57	14.90	103.42
GroundingREC [8]	10.06	58.62	10.12	107.19
CountGD [3]	12.14	47.51	12.98	98.35
CAD-GD (ours)	9.30	40.96	10.35	86.88

감사합니다