Open-World Semantic Segmentation Including Class Similarity

12 Mar 2024

CVPR 2024

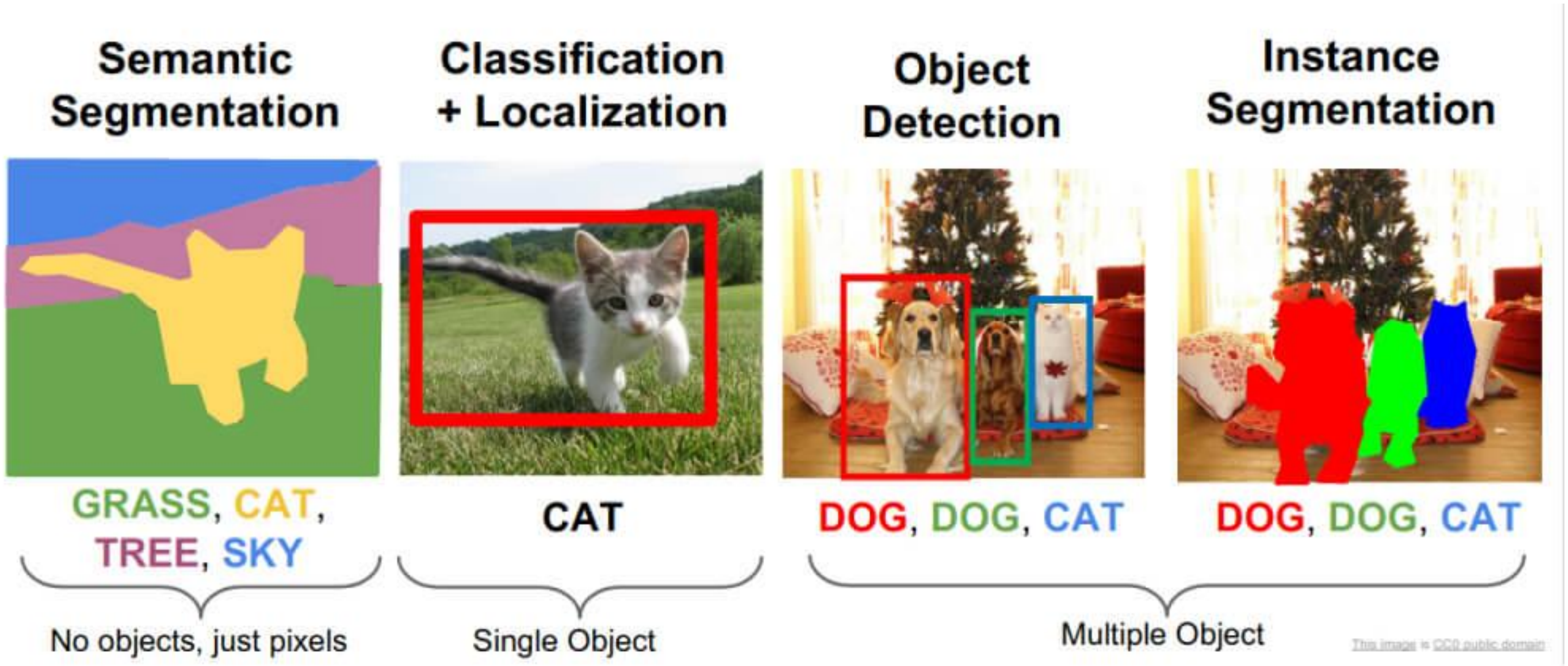
Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, Cyrill Stachniss

Presented by Gwanyong Kim

GSPS

Semantic Segmentation

이미지의 모든 픽셀에 대해 어떤 객체에 속하는지 분류하는 것



-자율 주행 차량과 같이 자율적으로 작동하는 시스템에게데이터 해석은 매우 중요하다.
-학습 중에 보지 못했던 객체가 발생하는 경우데이터 해석에 어려움이 생긴다

Closed-World Prediction’s Problem

-모델의 판단에 overconfident하는 경향이 있음
-알고 있는 class에 속하지 않는 객체를 올바르게 인식하지 못함

Previous Open-World Prediction

-이상 샘플을 버려야 하는 객체로 인식하도록 분류하는 것을 탐색
- -Thresholding the softmax activations
- -Using a background class for tackling unknown samples
- -Using model ensembles

Problem

-Closed-world prediction은 알려지지 않은 샘플에 대해서 소프트맥스에서 피크를 보이며 이를 Overconfident하는 경향을 보임
-이 세상 모든 객체를 학습시키는 것은 비효율적임

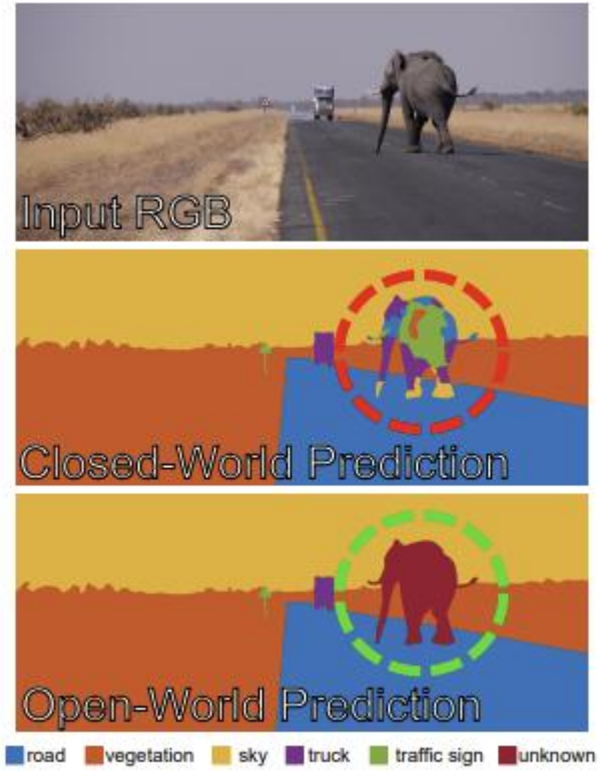


Figure 1. Given an image containing a previously-unseen object (top), closed-world methods for semantic segmentation classify the pixels belonging to that object as one of the known classes (center, red circle). Our goal is to segment the unknown object and identify it as a semantic class different to the previously-known ones (bottom, green circle).

1 Encoder - 2 Decoder

Loss function

1 Encoder - 2 Decoder

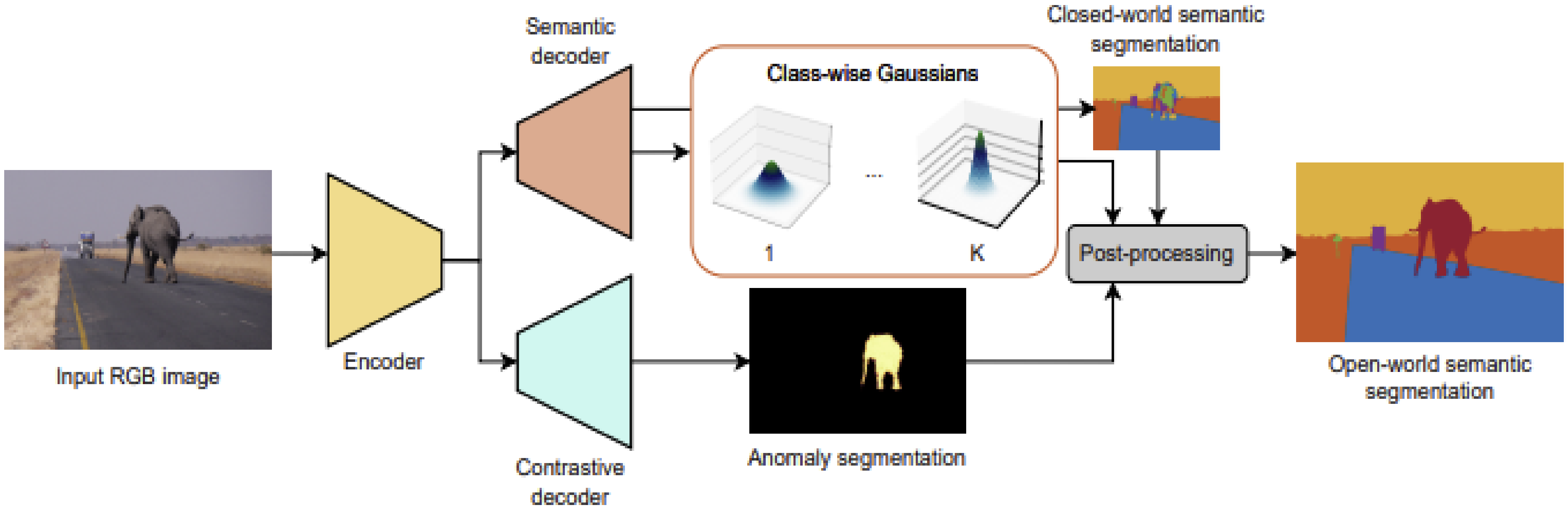


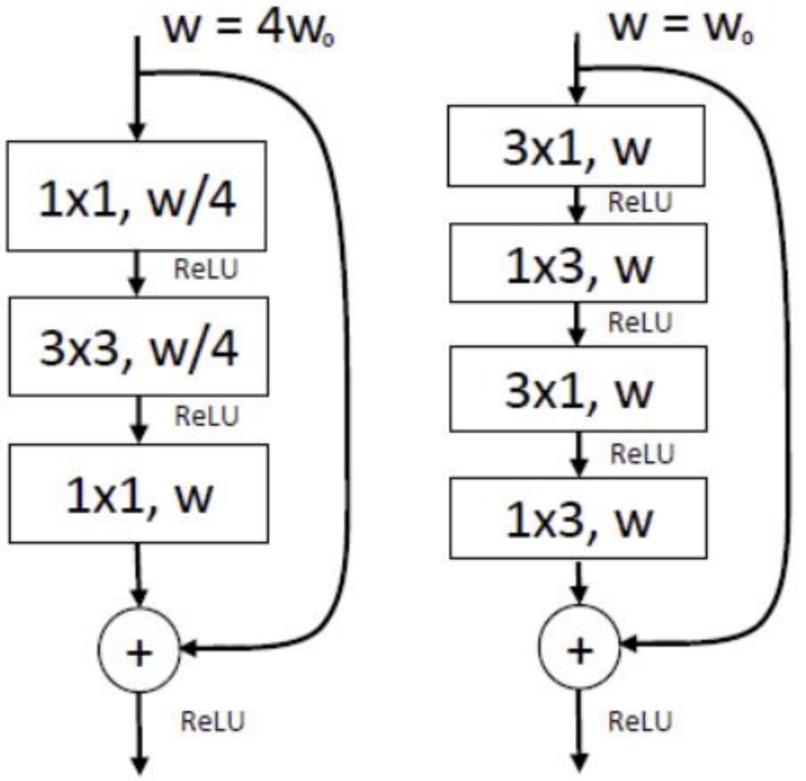
Figure 2. Given an RGB image as input, our network processes it by means of an encoder and two decoders. The semantic decoder produces a closed-world semantic segmentation and a Gaussian model for each known category. The class Gaussian models are built from a learned class descriptor (mean) and the variance of all predictions from it. A 3D example is shown in the image. The contrastive decoder provides an anomaly segmentation output. A post-processing phase finally achieves open-world semantic segmentation.

GSPS

Encoder

-ResNet 34
-기본 ResNet block을 NonBottleNeck-1D Block으로 대체
-3x1, 1x3conv로 대체되어 더 가벼운 구조

-인코더 부분 후에 피라미드 풀링 모듈을 사용하여 ResNet의 제한된 Receptive field를 확장



(b) Bottleneck (c) Non-bottleneck-1D

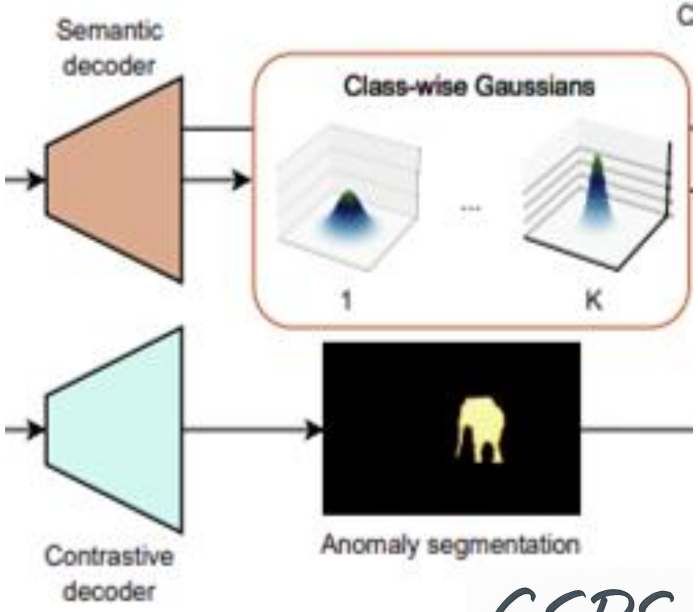
GSPS

Decoder

1) Semantic Decoder
   -기본의 Semagtic segmentation을 수행하는 역할
   -각 클래스에 대해 고유한 특징 벡터를 생성하도록 feature space를 조작
   -known클래스에 대해 정확한 segmentation을 수행하면서 각 픽셀의 pre-softmax feature가 해당 클래스의 descriptor와 유사하도록 만듦.
   -이를 통해 특정 클래스에 할당된 픽셀의 feature 벡터가 그 클래스의 descriptor와 크게 다르면 해당 픽셀을 unknown class로 탐지할 수 있습니다

2) Contrastive Decoder
   -새로운 객체(Unknown class)를 탐지하는 역할
   -Binary Segmentation (Known class - Unknown Class)

-> 두 Decoder의 결과를 후처리하여 최종 Open-world prediction을 얻음



Loss function

Semantic Decoder loss function

$$\mathcal{L}_{\text{sdec}} = w_1 \mathcal{L}_{\text{sem}} + w_2 \mathcal{L}_{\text{feat}}$$

$$\mathcal{L}_{\text{sem}} = -\frac{1}{|\Omega|} \sum_{p \in \Omega} \omega_k \boldsymbol{t}_p^{\top} \log\left(\sigma(\boldsymbol{f}_p)\right),$$

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\Omega|} \sum_{k=1}^{K} \sum_{p \in \Omega_k} \frac{\|\boldsymbol{f}_p - \boldsymbol{\mu}_k^{e-1}\|}{\boldsymbol{\sigma}_k^{e-1}}$$

$$\boldsymbol{\sigma}_k^2 = \frac{1}{|\hat{\Omega}_k|} \sum_{p \in \hat{\Omega}_k} (\boldsymbol{f}_p - \boldsymbol{\mu}_k)^2$$

$$\boldsymbol{\mu}_k = \frac{1}{|\hat{\Omega}_k|} \sum_{p \in \hat{\Omega}_k} \boldsymbol{f}_p$$

Ω : 이미지의 모든 픽셀 집합
Ω_k: GT 클래스도 k, 예측도 클래스 k인 픽셀
t_p: : 픽셀 p의 one hot GT
w_q : 클래스 불균형 보정용 가중치
f_p : 픽셀 p에 대한 pre softmax feature

Loss function

Contrastive Decoder loss function

$$\mathcal{L}_{\text{cdec}} = w_3\,\mathcal{L}_{\text{cont}} + w_4\,\mathcal{L}_{\text{obj}}$$

$$\mathcal{L}_{\text{cont}} = -\sum_{k=1}^{K} \log \frac{\exp\left(\bar{\boldsymbol{f}}_k^{\top} \bar{\boldsymbol{\mu}}_k^{e-1}/\tau\right)}{\sum_{i=1}^{K} \exp\left(\bar{\boldsymbol{f}}_k^{\top} \bar{\boldsymbol{\mu}}_i^{e-1}/\tau\right)}$$

$$\mathcal{L}_{\text{obj}} = \begin{cases} \max\left(\xi - \|\boldsymbol{f}_p\|^2, 0\right) & \text{, if } p \in \mathcal{D}_k \\ \|\boldsymbol{f}_p\|^2 & \text{, otherwise} \end{cases}$$

$$\bar{\boldsymbol{f}}_k = \frac{1}{|\Omega_k|} \sum_{p \in \Omega_k} \boldsymbol{f}_p^d$$

ξ: : hypersphere의 반지름
Ω_k: GT 클래스도 k, 예측도 클래스 k인 픽셀
t_p: : 픽셀 p의 one hot GT
w_q : 클래스 불균형 보정용 가중치
f_{dp} : 픽셀 p에 대한 pre softmax feature
μ_e-1_k: 이전 epoch에서의 클래스 k의 분포의 평균값
Dk: known pixel
T(타우): temperature, 작으면 클래스 구분 강해짐

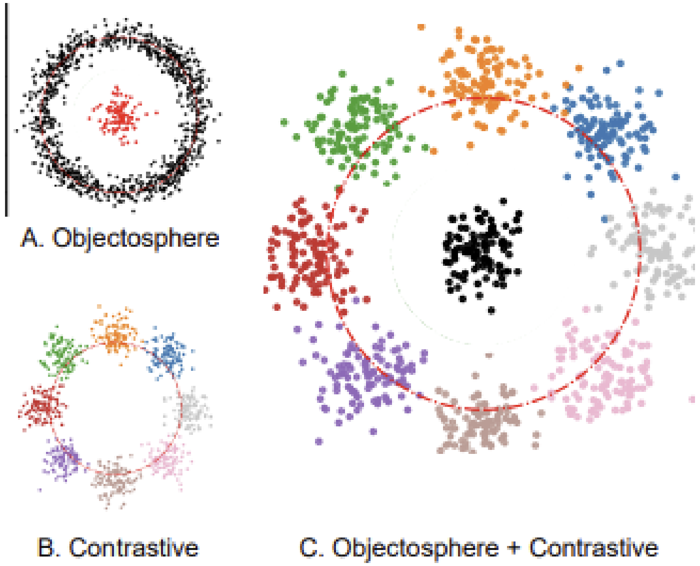


Figure 3. 2D visualization of the expected output of the contrastive decoder. The behavior of the objectosphere loss is shown in A, where all points coming from known classes (black) lie around the red (outer) circle of radius ξ, see Eq. (9), and the points from unknown classes lie around the origin. The contrastive loss is shown in B, where features lie on the unit circle. Together, they lead to a behavior similar to the one depicted in C.

GSPS

Post Processing

$$s_{\mathrm{unk},\,p} = \frac{1}{2}\left(s^{\mathrm{sem}}_{\mathrm{unk},\,p} + s^{\mathrm{cont}}_{\mathrm{unk},\,p}\right)$$

$$s^{\mathrm{sem}}_{\mathrm{unk},\,p} = 1 - s(p)$$

$$s^{\mathrm{cont}}_{\mathrm{unk},\,p} = \max\left(0, \left(1 - \frac{\|\boldsymbol{f}_p\|^2}{\xi}\right)\right)$$

$$s(p) = \max_k s_k(\boldsymbol{f}_p)$$

$$s_k(\boldsymbol{f}_p) = \exp\left(-\frac{1}{2}(\boldsymbol{f}_p - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{f}_p - \boldsymbol{\mu}_k)\right)$$

ξ: : hypersphere의 반지름
w_q : 클래스 불균형 보정용 가중치
f_p : 픽셀 p에 대한 pre softmax feature
μ_k: 클래스 k의 분포의 평균값
Σ_k^{-1} 클래스 k의 공분산 행렬의 역행렬

Unknown Class Discovery

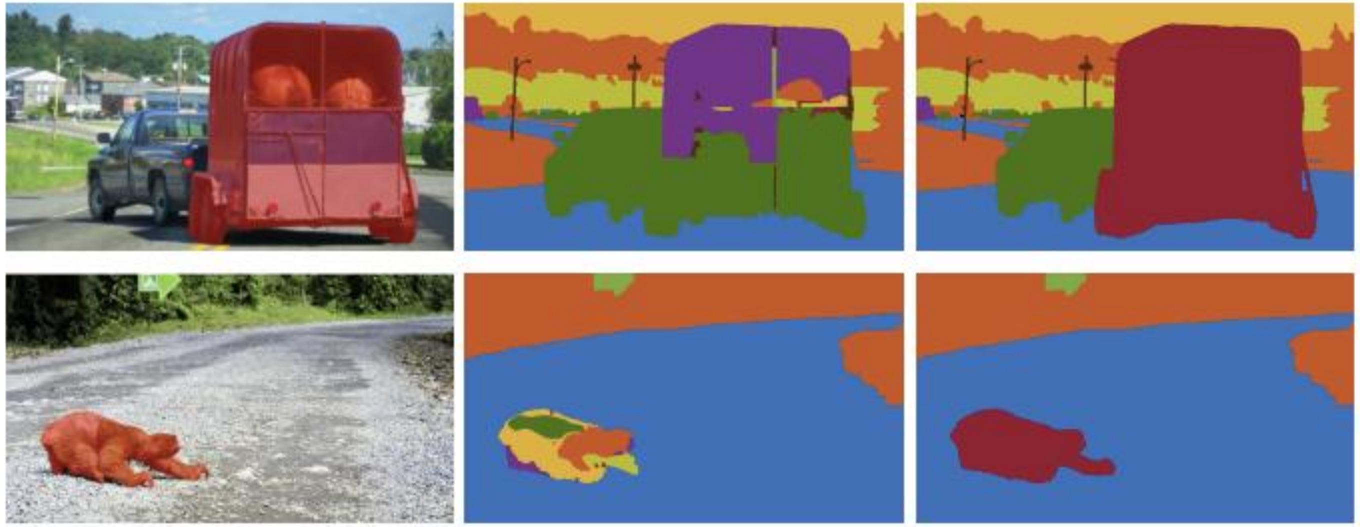
$$\mathcal{F} = \{\boldsymbol{f}_u^1, \ldots, \boldsymbol{f}_u^G\}$$

Unknown classes set

Class Similarity

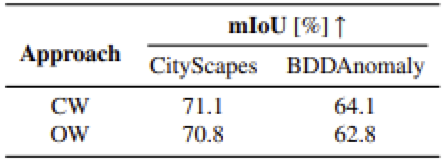
$$\tilde{k} = \operatorname{argmax}_k s_k(\boldsymbol{f}_p)$$

Unknown 픽셀의 feature fp에 대해 각 known class k에 대한 gaussian scaore 계산을 통해 비슷한 클래스 분류

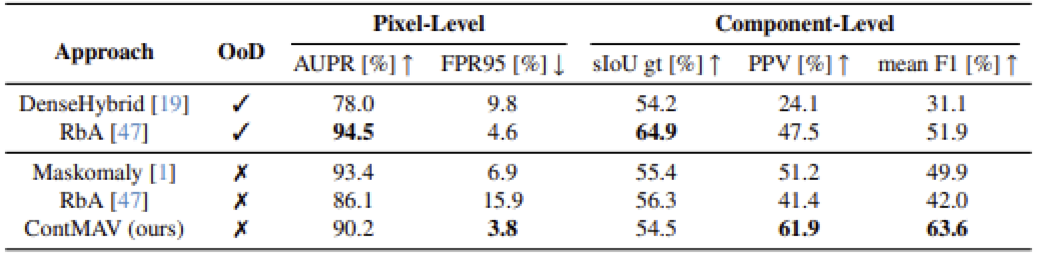


(a) Input RGB (b) Closed-world prediction (c) Open-world prediction

Figure 4. Results from the validation set of SegmentMeIfYouCan. We show the input RGB overlayed with the ground truth unknown mask (a), the prediction of our closed-world model (b), and the prediction of our approach for open-world segmentation (c). In the open-world prediction, the unknown class is shown in red.



Approach	mIoU [%] ↑	
	CityScapes	BDDAnomaly
CW	71.1	64.1
OW	70.8	62.8



Approach	OoD	Pixel-Level		Component-Level		
		AUPR [%] ↑	FPR95 [%] ↓	sIoU gt [%] ↑	PPV [%] ↑	mean F1 [%] ↑
DenseHybrid [19]	✓	78.0	9.8	54.2	24.1	31.1
RbA [47]	✓	**94.5**	4.6	**64.9**	47.5	51.9
Maskomaly [1]	✗	93.4	6.9	55.4	51.2	49.9
RbA [47]	✗	86.1	15.9	56.3	41.4	42.0
ContMAV (ours)	✗	90.2	**3.8**	54.5	**61.9**	**63.6**

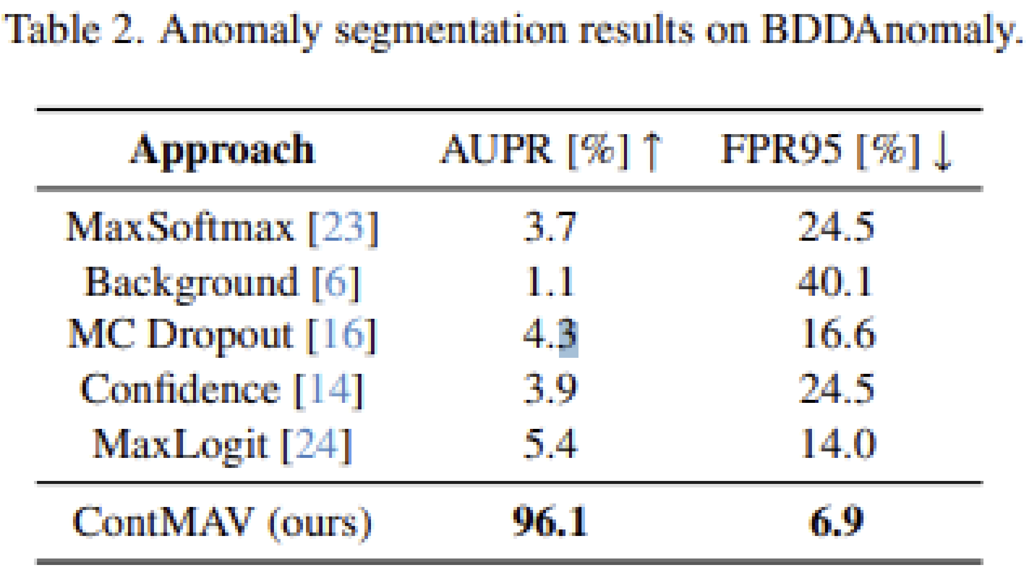


Table 2. Anomaly segmentation results on BDDAnomaly.

Approach	AUPR [%] ↑	FPR95 [%] ↓
MaxSoftmax [23]	3.7	24.5
Background [6]	1.1	40.1
MC Dropout [16]	4.3	16.6
Confidence [14]	3.9	24.5
MaxLogit [24]	5.4	14.0
ContMAV (ours)	**96.1**	**6.9**

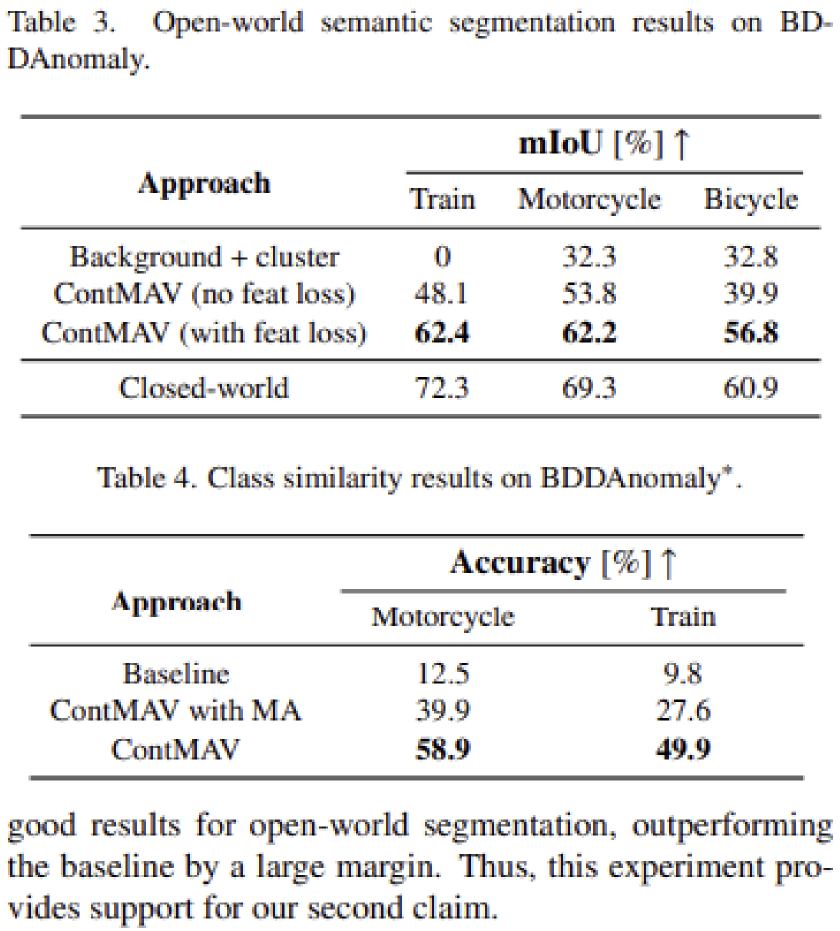


Table 3. Open-world semantic segmentation results on BDDAnomaly.

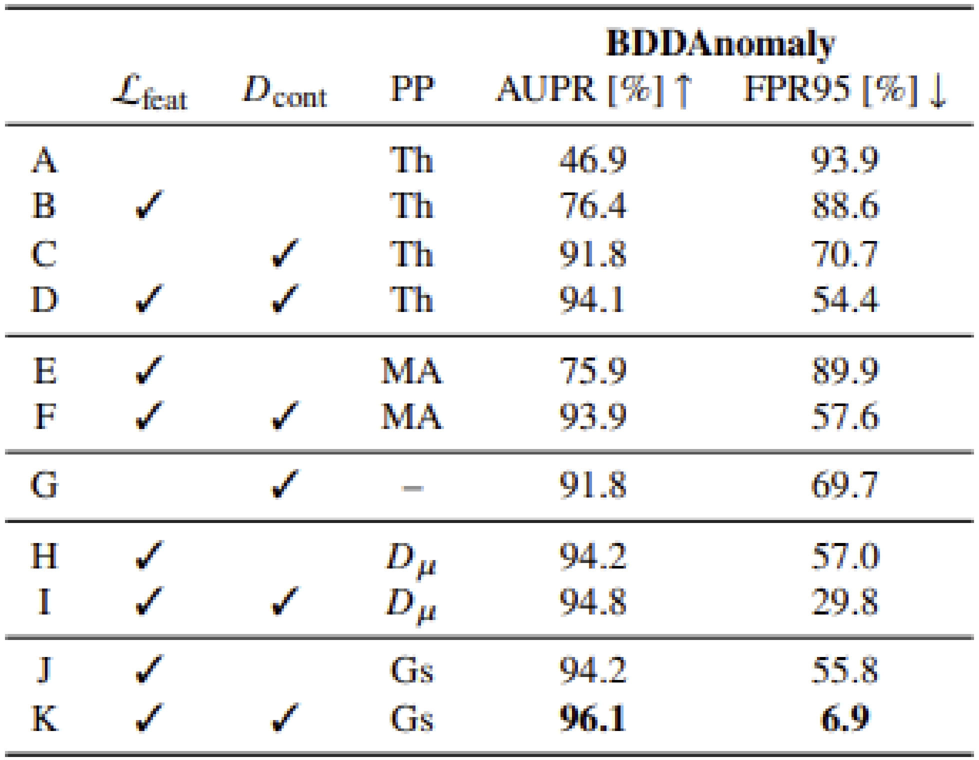
Approach	mIoU [%] ↑		
	Train	Motorcycle	Bicycle
Background + cluster	0	32.3	32.8
ContMAV (no feat loss)	48.1	53.8	39.9
ContMAV (with feat loss)	**62.4**	**62.2**	**56.8**
Closed-world	72.3	69.3	60.9

Table 4. Class similarity results on BDDAnomaly*.

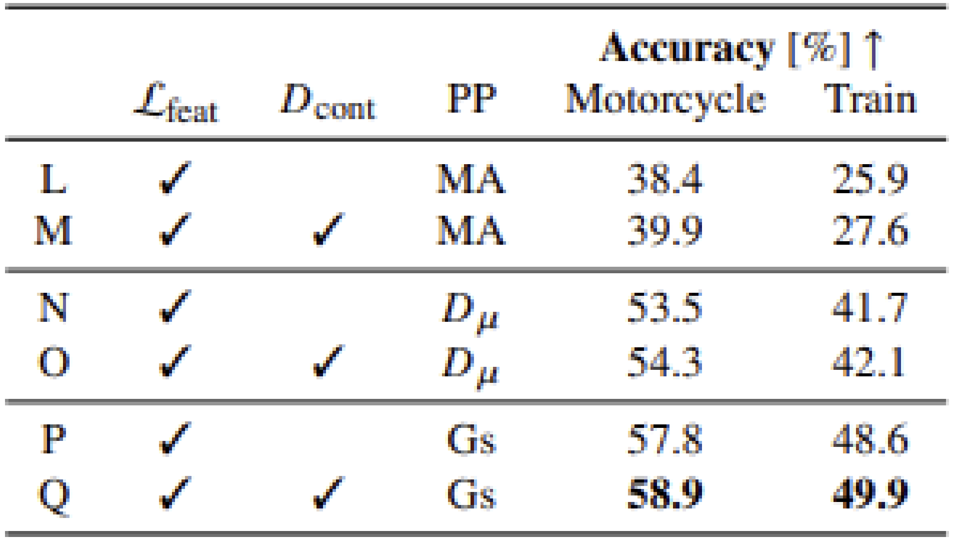
Approach	Accuracy [%] ↑	
	Motorcycle	Train
Baseline	12.5	9.8
ContMAV with MA	39.9	27.6
ContMAV	**58.9**	**49.9**

good results for open-world segmentation, outperforming the baseline by a large margin. Thus, this experiment provides support for our second claim.

Ablation study



	$\mathcal{L}_{\text{feat}}$	D_{cont}	PP	BDDAnomaly AUPR [%] ↑	 FPR95 [%] ↓
A			Th	46.9	93.9
B	✓		Th	76.4	88.6
C		✓	Th	91.8	70.7
D	✓	✓	Th	94.1	54.4
E	✓		MA	75.9	89.9
F	✓	✓	MA	93.9	57.6
G		✓	–	91.8	69.7
H	✓		D_μ	94.2	57.0
I	✓	✓	D_μ	94.8	29.8
J	✓		Gs	94.2	55.8
K	✓	✓	Gs	**96.1**	**6.9**



	$\mathcal{L}_{\text{feat}}$	D_{cont}	PP	Accuracy [%] ↑ Motorcycle	 Train
L	✓		MA	38.4	25.9
M	✓	✓	MA	39.9	27.6
N	✓		D_μ	53.5	41.7
O	✓	✓	D_μ	54.3	42.1
P	✓		Gs	57.8	48.6
Q	✓	✓	Gs	**58.9**	**49.9**

클래스별 feature distribution

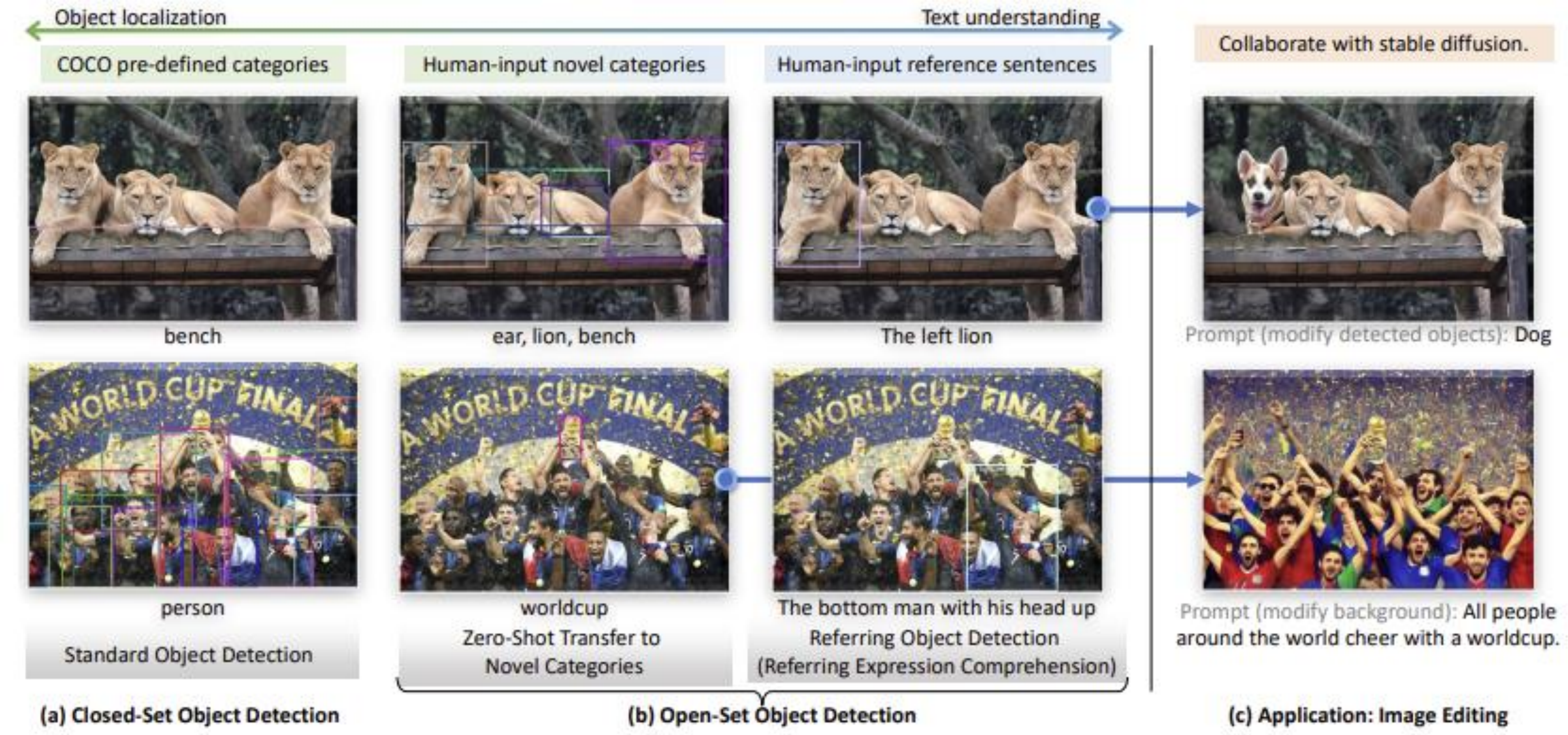
+

Contrastive, objectoshpere

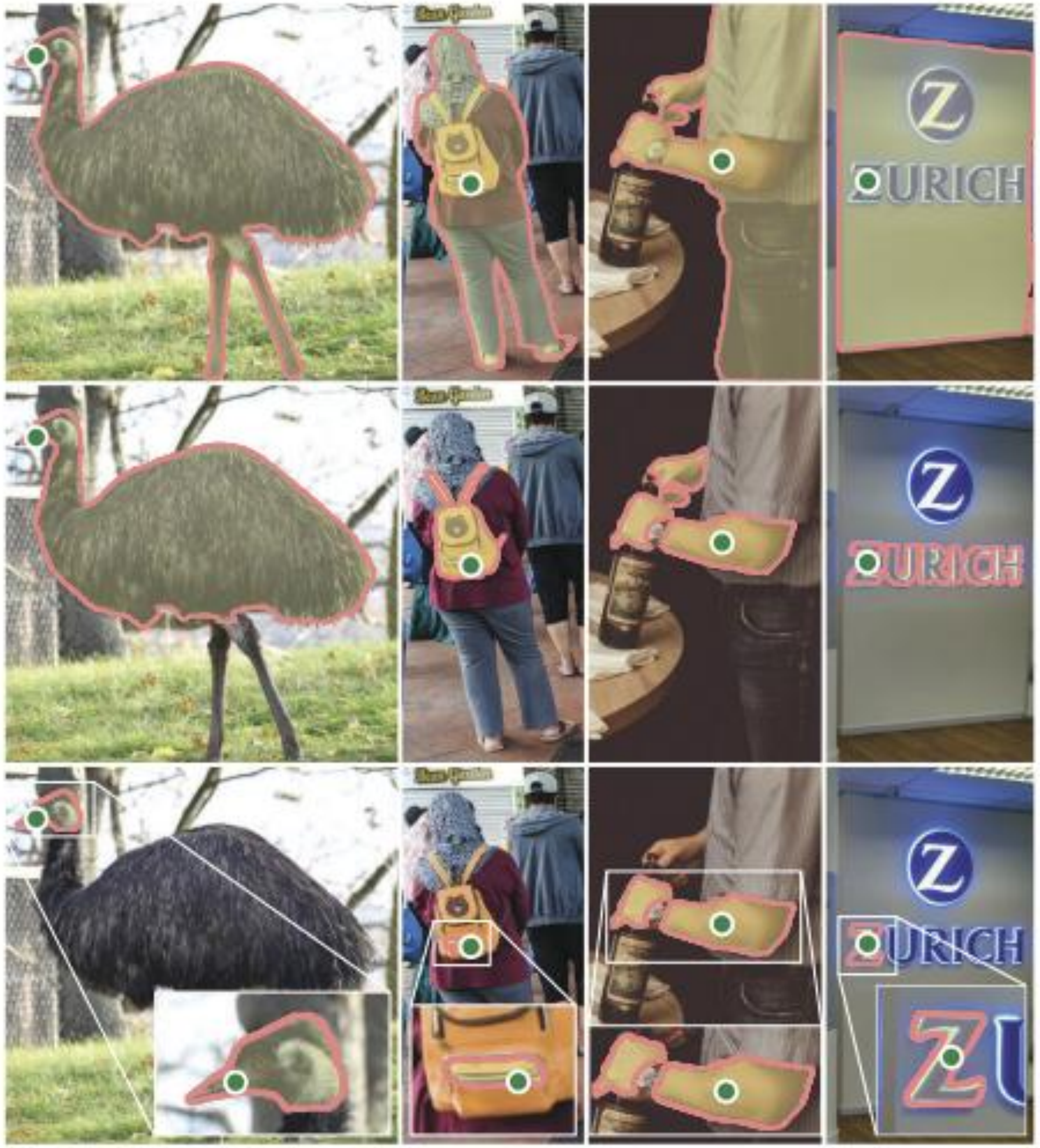
⬇

추가 데이터셋 없이 open world 문제 해결

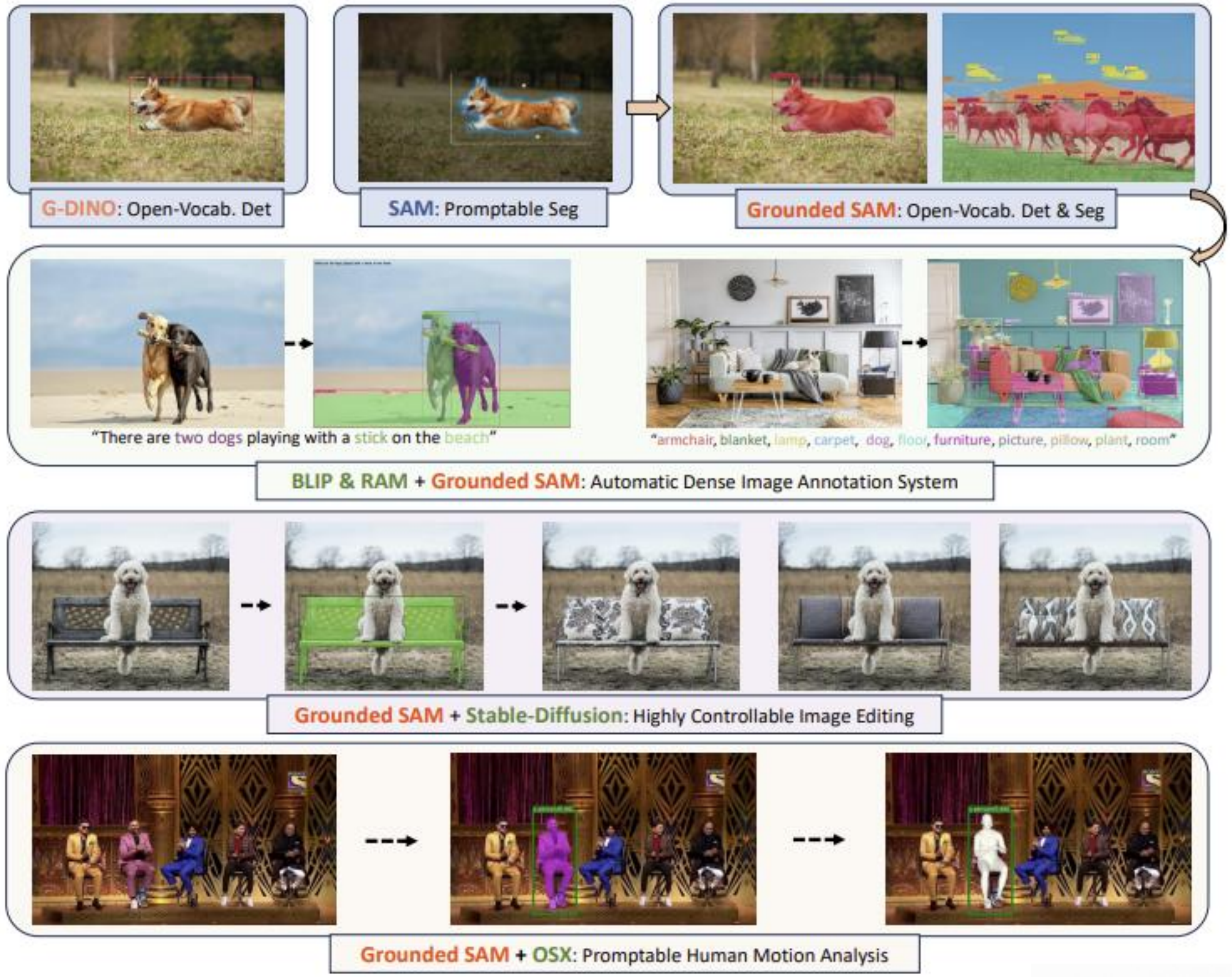
Grounding DINO


Object localization
Text understanding
COCO pre-defined categories
Human-input novel categories
Human-input reference sentences
Collaborate with stable diffusion.
bench
ear, lion, bench
The left lion
Prompt (modify detected objects): Dog
person
worldcup
The bottom man with his head up
Prompt (modify background): All people around the world cheer with a worldcup.
Standard Object Detection
Zero-Shot Transfer to Novel Categories
Referring Object Detection (Referring Expression Comprehension)
(a) Closed-Set Object Detection
(b) Open-Set Object Detection
(c) Application: Image Editing

Segment Anything



Grounded SAM



G-DINO: Open-Vocab. Det

SAM: Promptable Seg

Grounded SAM: Open-Vocab. Det & Seg

"There are two dogs playing with a stick on the beach"

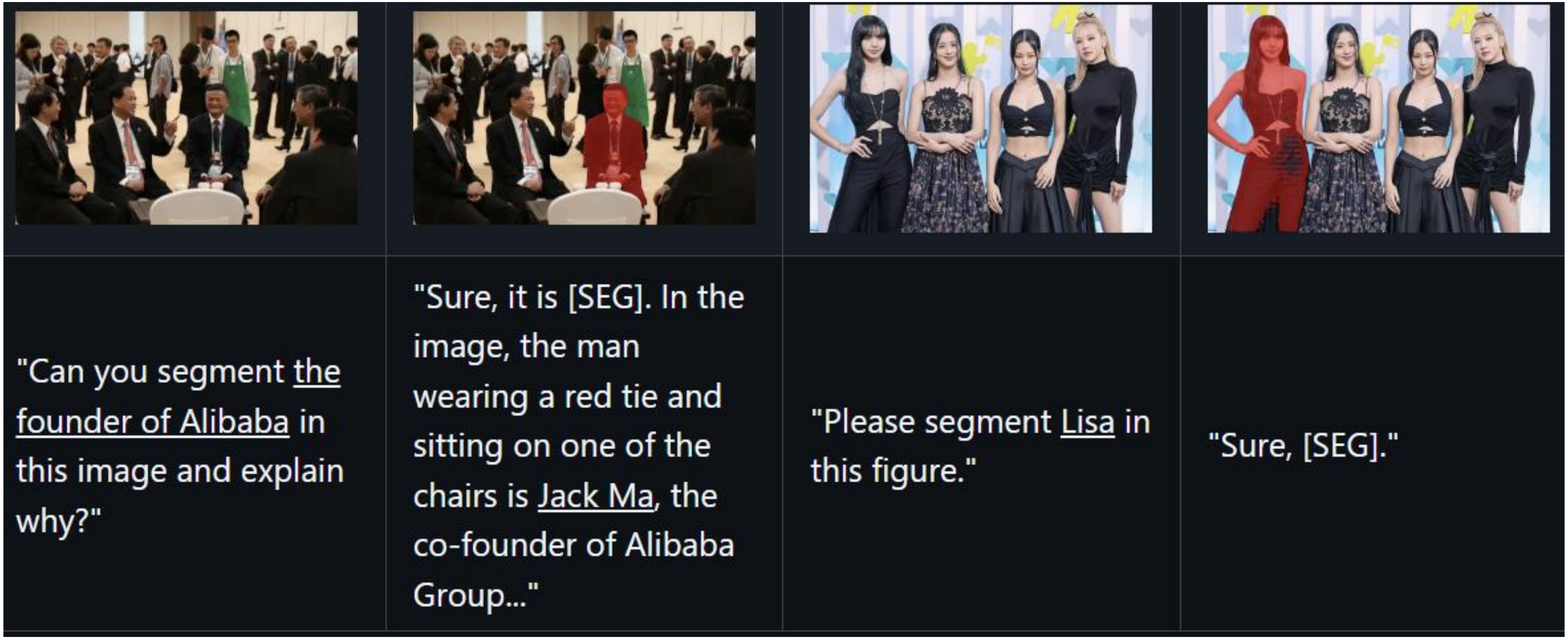
"armchair, blanket, lamp, carpet, dog, floor, furniture, picture, pillow, plant, room"

BLIP & RAM + Grounded SAM: Automatic Dense Image Annotation System

Grounded SAM + Stable-Diffusion: Highly Controllable Image Editing

Grounded SAM + OSX: Promptable Human Motion Analysis

LISA: Reasoning Segmentation via Large Language Model



"Can you segment the founder of Alibaba in this image and explain why?"	"Sure, it is [SEG]. In the image, the man wearing a red tie and sitting on one of the chairs is Jack Ma, the co-founder of Alibaba Group..."	"Please segment Lisa in this figure."	"Sure, [SEG]."