Segment Anything Model <SAM>

ComputerVision, Segmentation

- meta AI <2023.04.05>

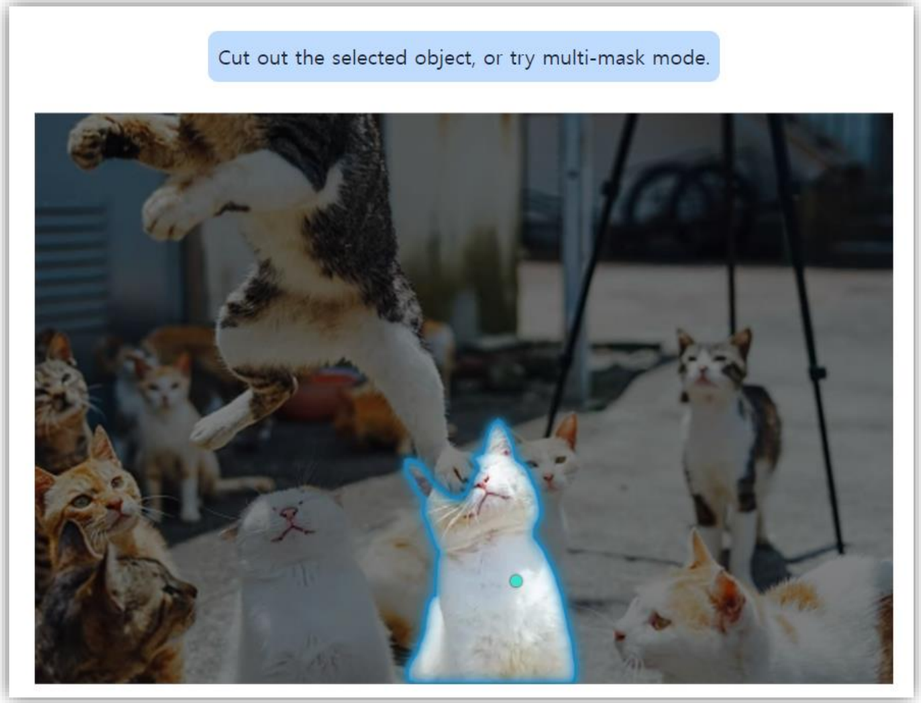
백 대 환

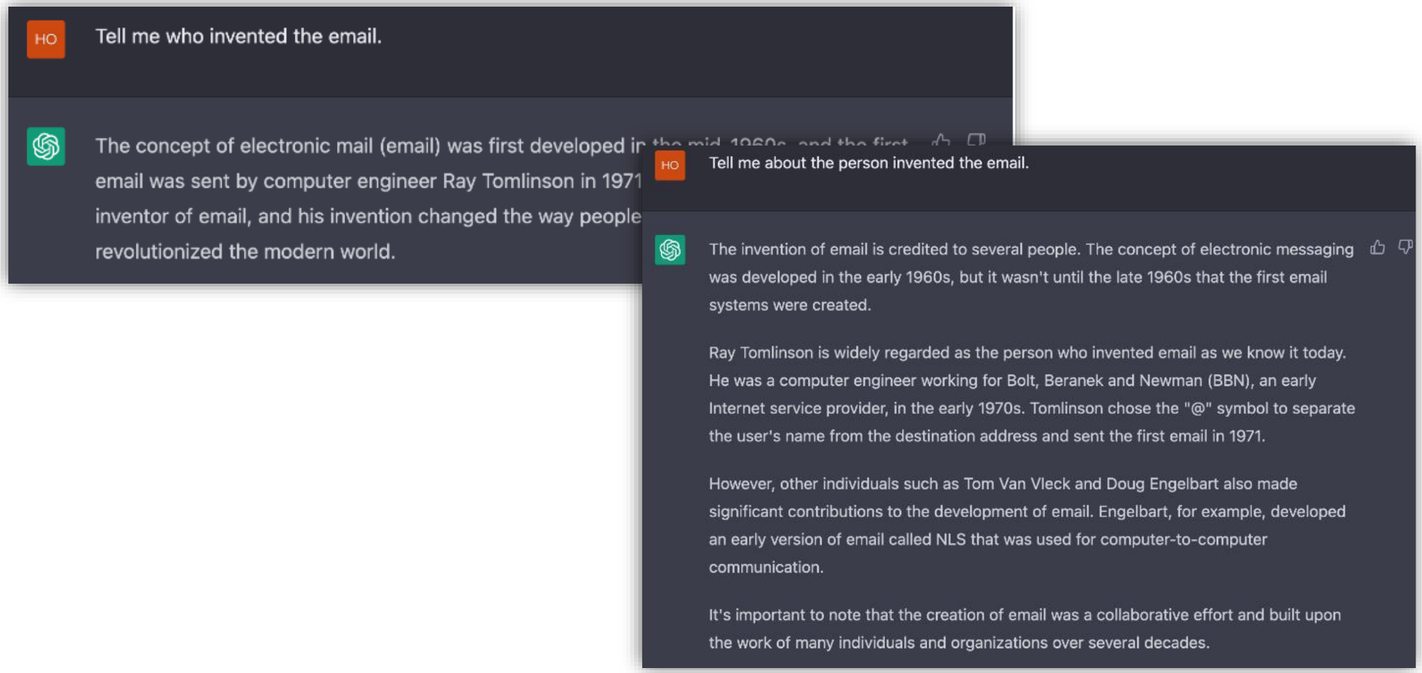
PaperReview

Contents

Introduction

- Segment Anything Model은 모든 분야에서 광범위하게 사용할 수 있는 image segmentation model에 대해서 설명.
- **Foundation**이 되는 모델이 될 수 있도록 하는 것이 목표.
- 광범위한 데이터셋 SA-1B 를 새롭게 만들고, 이것으로 학습시켜, **Powerful generalization**을 시킴
- **Prompt-engineering**을 통해 새로운 데이터 분포에서 다양한 task를 수행할 수 있도록 함.
- https://segment-anything.com/





Glossary

- **Task**: SAM 모델에서의 prompt는 다양한 것이 될 수 있다. 위 그림의 (a)에서 볼 수 있듯이 foreground와 background를 구분하는 점이 될 수도 있고, 러프한 bbox나 mask가 될 수도 있다. 그리고 타겟에 대해서 설명한 글이어도 상관없다. 이처럼 mask를 생성하고자 하는 대상에 대한 정보를 담은 어떤 것이든 prompt가 될 수 있다. 이를 통해서 다양한 downstream에 대해서 zero-shot transfer가 가능했다고 한다.
- **Pre-training**: prompt와 image를 함께 받아와서 대용량 데이터셋으로 valid한 mask를 생성해내는 pre-training을 한다. 이 때 'ambiguous'한 prompt에 대해서도 학습했다고 한다. 이렇게 ambiguous prompt와 함께 학습시켰을 때 어떠한 prompt가 오더라도 valid한 mask를 생성해 낼 수 있었다고 한다.
- **- 'ambiguous'한 prompt란**?: 예를 들어서 셔츠를 입고 있는 사람에서 셔츠에 prompt point가 찍혔다고 생각해보자. 그러면 이것이 사람 전체를 뜻하는 것인지, 아니면 셔츠만 타겟하는 것인지 알 수가 없다. 이런 경우 'ambiguous'하다고 한다.

Glossary

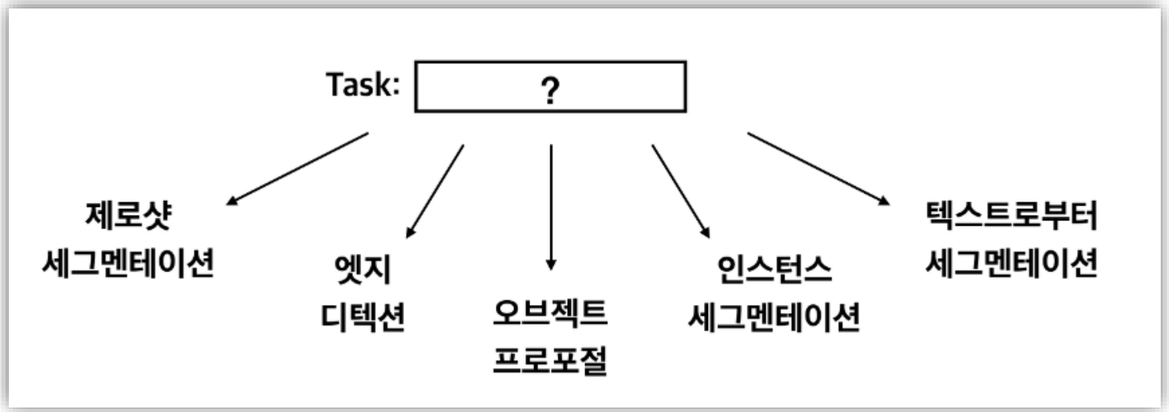
- **zero-shot transfer**: pre-training을 통해서 어떠한 prompt가 오더라도 valid한 마스크를 생성할 수 있는 모델이 있기 때문에, zero-shot transfer가 가능하다고 한다. 예를 들어 고양이에 bounding box가 쳐져 있다면, 이를 통해서 별다른 학습 없이 고양이를 segmentation하는 것이 가능하다.
- **related tasks**: segmentation에는 다양한 종류가 있다. 예를 들어, edge detection, super pixelization, interactive segmentation, object proposal generation, foreground segmentation, panoptic segmentation 등등이 있다. SAM은 prompt engineering을 통해 이렇게 다양한 task에 적용할 수 있다고 한다. 존재하는 다양한 object detector와 결합하여 사용할 수도 있다는 점에서 확장 가능성이 크다고 한다.
- **Data Engine**: prompting은 powerful한 방법이기하나 이미지의 prompt에 대해서 mask GT를 만드는 것도 human cost가 드는 일이다. 따라서 fully human annotated dataset을 사용하기 보다는 "data engine"을 통해서 GT를 생성했다고 한다. 처음에는 사람이 생성한 GT로 학습을 하고 (assisted-manual), 그 후에는 사람이 생성한 것과 generate된 마스크를 함께 학습하고 (semi-automatic), 마지막에는 foreground에 대해서 point가 주어진 prompt로 생성된 mask를 통해서만 학습했다고 한다. 이를 통해서 SA-1B라는 거대한 데이터셋을 만들 수 있었다고 함.

Motivation

- 최근 GPT4와 같은 Large Language Model(LLM)이 높은 **Zero-shot / Few Shot Generalization** 성능을 보임.
- LLM과 같이 대량의 데이터 셋을 **pre-train** 하고, down-stream task에 대해 높은 **zero-shot generalization** 성능을 보이는 모델을 **Foundation Model** 이라고 부름.
- 컴퓨터 비전 분야에서도 **CLIP, ALIGN** 같이 Vision-Language Dataset으로 Foundation Model을 만들려는 시도.

Segment Anything Task(Task, Data, Model)

- 1. What **task** will enable zero-shot generalization?
- 2. What is the corresponding **model** architecture?
- 3. What **data** can power this task and model?

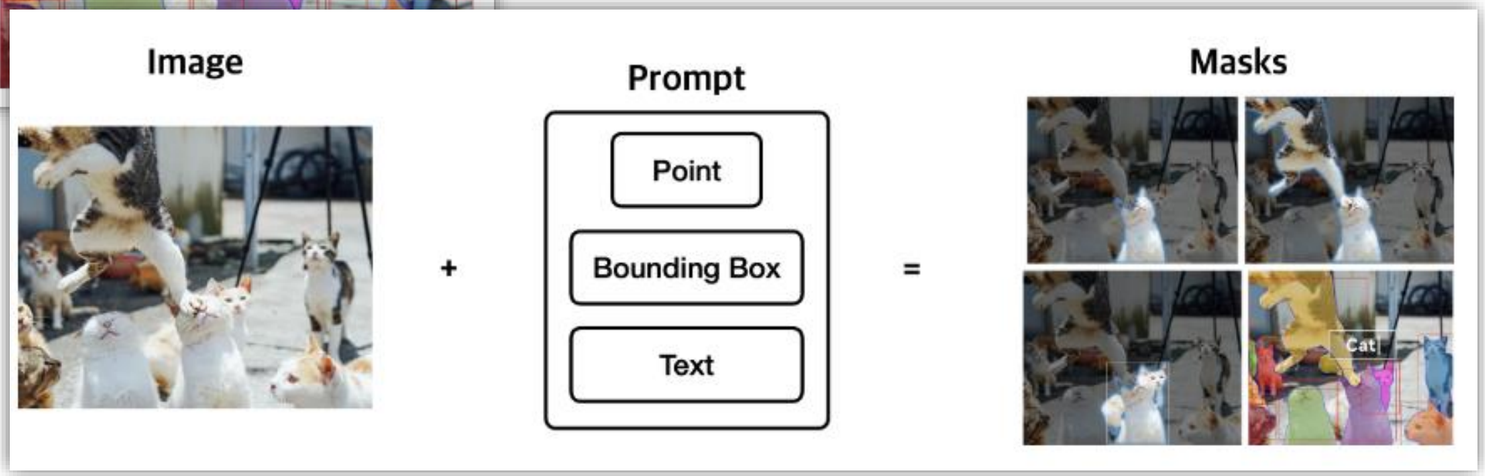


Segment Anything Task(Task, Data, Model)

➢ 1. What **task** will enable zero-shot generalization?

➢ **Promptable Segmentation**: prompt로는 점, 박스, 텍스트

➢ 어떤 prompt가 주어졌을 때, 유효한 Mask를 반환



Segment Anything Task(Task, Data, Model)

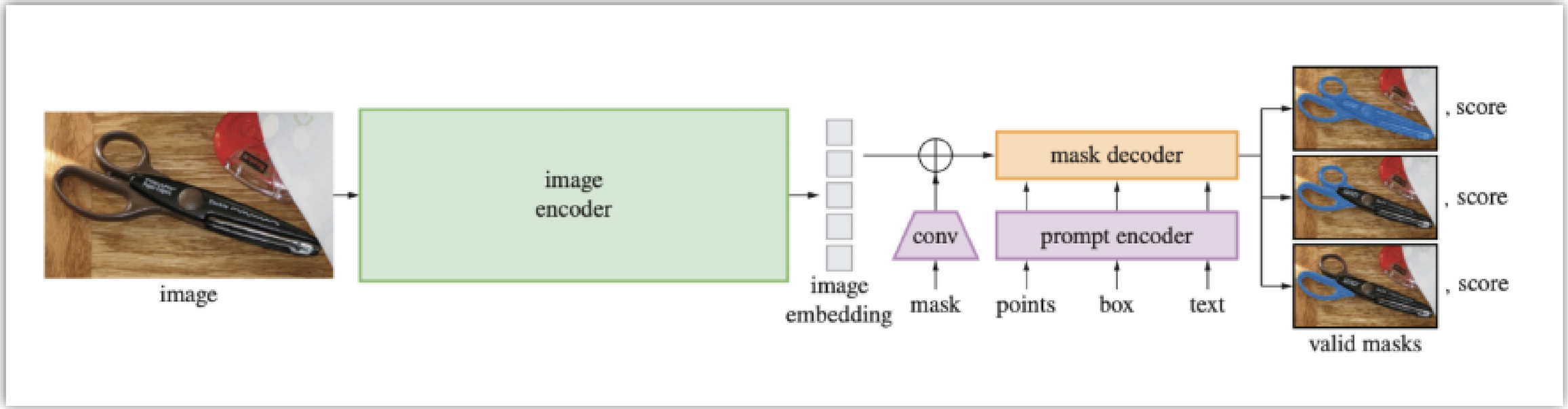
➢ 1. What **task** will enable zero-shot generalization?

➢ **Ambiguous(모호한) prompt** 가 주어졌을 때도 합리적인 mask 출력이 필요

➢ 해결법: **Mask를 여러 개 출력**하여 합리적인 mask를 출력할 확률을 높임.

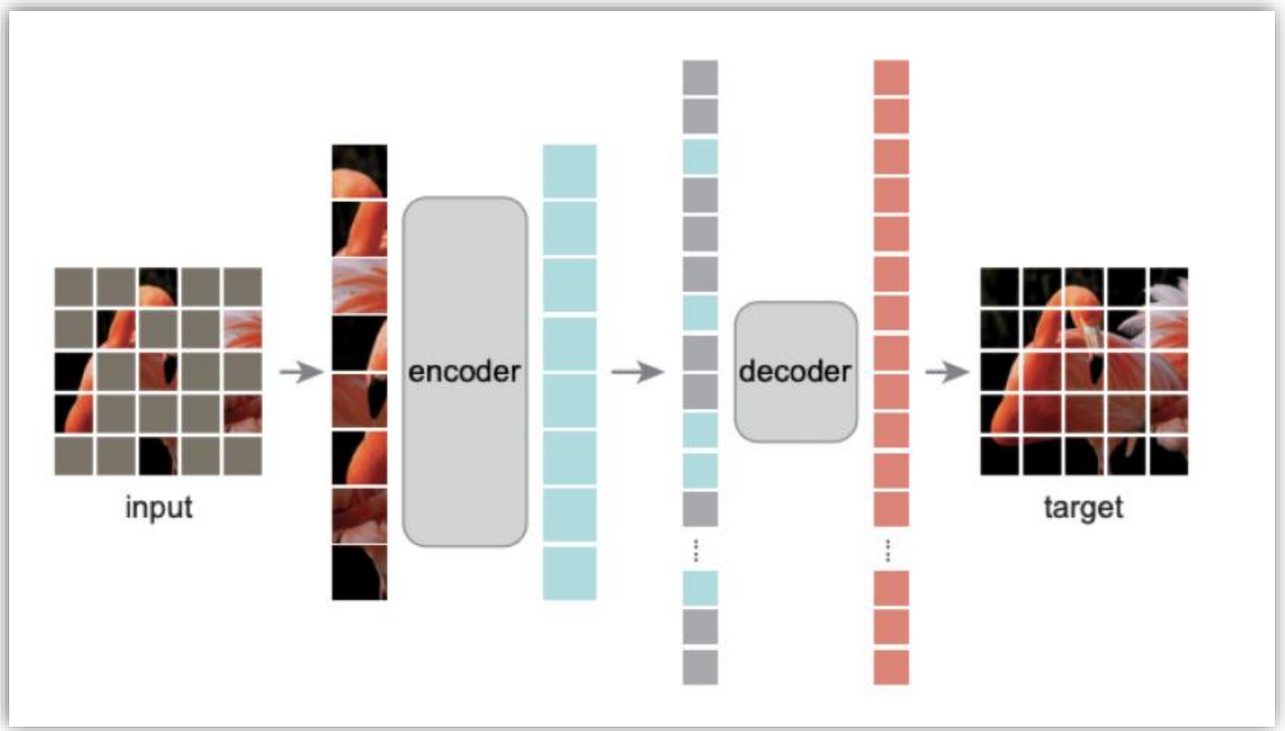
Segment Anything Task(Task, Model, Data)

- 2. What is the corresponding model architecture?
- Image encoder, Prompt encoder, Mask Decoder



Segment Anything Task(Task, Model, Data)

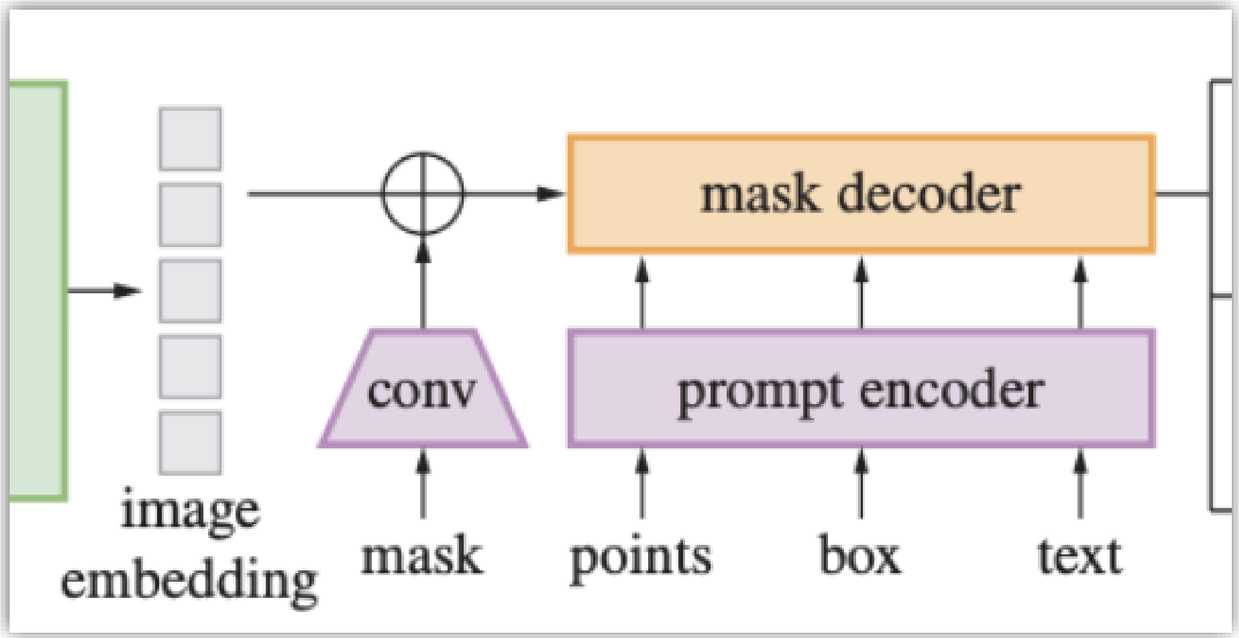
- 2. What is the corresponding model architecture?
- **Image encoder**, Prompt encoder, Mask Decoder
- Masked auto-encoder 방식으로 학습시킨 Vision transformer



Segment Anything Task(Task, Model, Data)

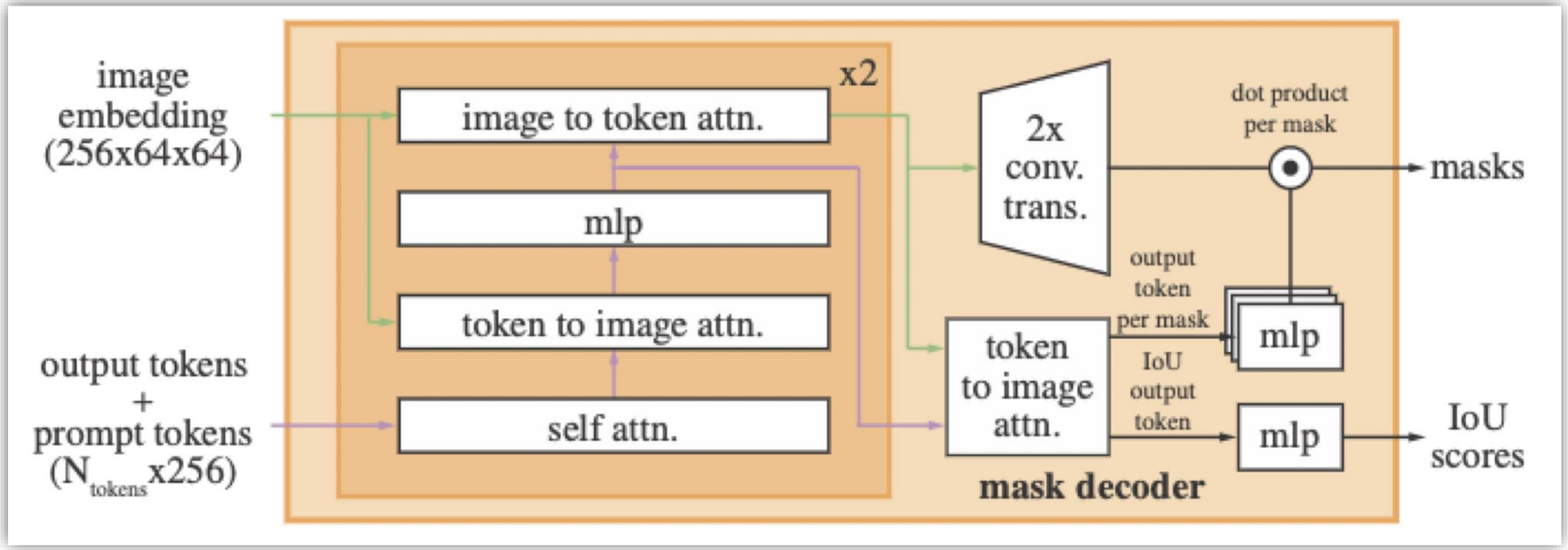
➢ 2. What is the corresponding model architecture?

- Image encoder, **Prompt encoder**, Mask Decoder
- 점이나 bounding 박스는 positional encoding을 사용, text는 CLIP Multimodal embedding
- Prompt encoder + mask decoder는 50ms 이내에 mask를 예측
- **Ambiguous(모호한) prompt**에 대비하기 위해 여러 개의 mask 예측 설계



Segment Anything Task(Task, Model, Data)

- 2. What is the corresponding model architecture?
- Image encoder, Prompt encoder, **Mask Decoder**
- 이미지 embedding과 프롬프트 embedding 간의 cross attention 메커니즘

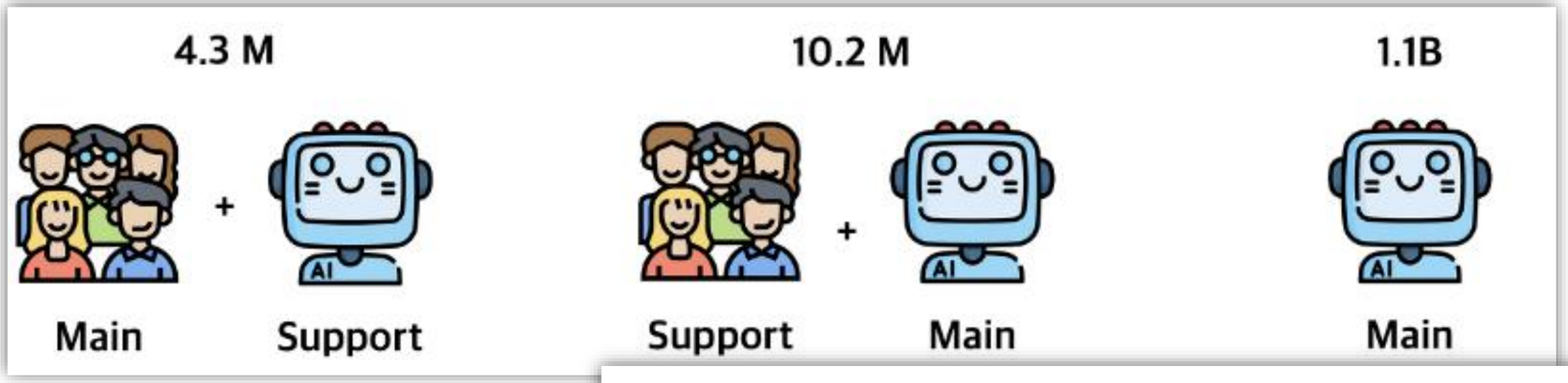


Segment Anything Task(Task, Model, Data)

➢ 3. What data can power this task and model?

➢ **Assisted Manual** -> Semi automatic -> Fully automatic

➢ 새로운 데이터에 대해서 AI가 먼저 segmentation을 해놓으면 사람이 이를 수정



Segment Anything Task(Task, Model, Data)

- 3. What data can power this task and model?
- Assisted Manual -> **Semi automatic** -> Fully automatic
- 기존의 segmentation 데이터 셋은 배제하고 1단계에서 모은 데이터 셋 만으로 SAM 모델 학습
- 마스크 590만개 추가 라벨링을 했습니다. (도합 1020만개)

Segment Anything Task(Task, Model, Data)

- 3. What data can power this task and model?
- Assisted Manual -> Semi automatic -> **Fully automatic**
- 1, 2 단계에서 모은 마스크 1020만개를 가지로 SAM 모델을 학습
- 이미지 1100만장에 대해 11억개의 마스크 라벨을 생성한게 **SA-1B** 데이터 셋

Segment Anything RAI Analysis

- RAI란 Responsible AI로 SA-1B와 **SAM이 얼마나 fair하느냐를 분석한 결과**
- 성별, 나이, 피부톤 등 유의미한 **성능차이가 없음**을 보여줌
- 사람들의 이미지를 분할하는 데 **공정하고 편향이 없는 모델**임을 시사

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	54.4 ±1.7	90.4 ±0.6
masculine	55.7 ±1.7	90.1 ±0.6
perceived age group		
older	62.9 ±6.7	92.6 ±1.3
middle	54.5 ±1.3	90.2 ±0.5
young	54.2 ±2.2	91.2 ±0.7

	mIoU at 1 point	mIoU at 3 points
perceived skin tone		
1	52.9 ±2.2	91.0 ±0.9
2	51.5 ±1.4	91.1 ±0.5
3	52.2 ±1.9	91.4 ±0.7
4	51.5 ±2.7	91.7 ±1.0
5	52.4 ±4.2	92.5 ±1.4
6	56.7 ±6.3	91.2 ±2.4

Table 2: SAM's performance segmenting people across perceived gender presentation, age group, and skin tone. 95% confidence intervals are shown. Within each grouping, all confidence intervals overlap except older *vs*. middle.

Experimental Results: Zero-shot Transfer

- 5개의 하위 task, SA-1B와 다른 distribution을 가진 새로운 데이터셋에 대해서 평가
- 1. edge detection
- 2. segment everything
- 3. object proposal generation
- 4. segment detected objects
- 5. segment objects from free-form text

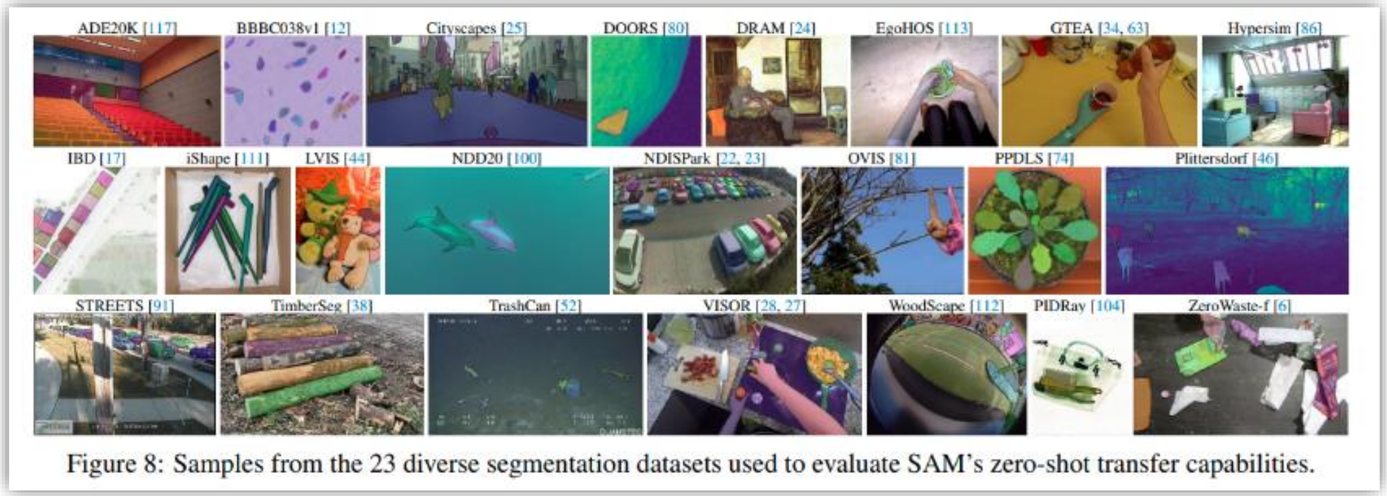
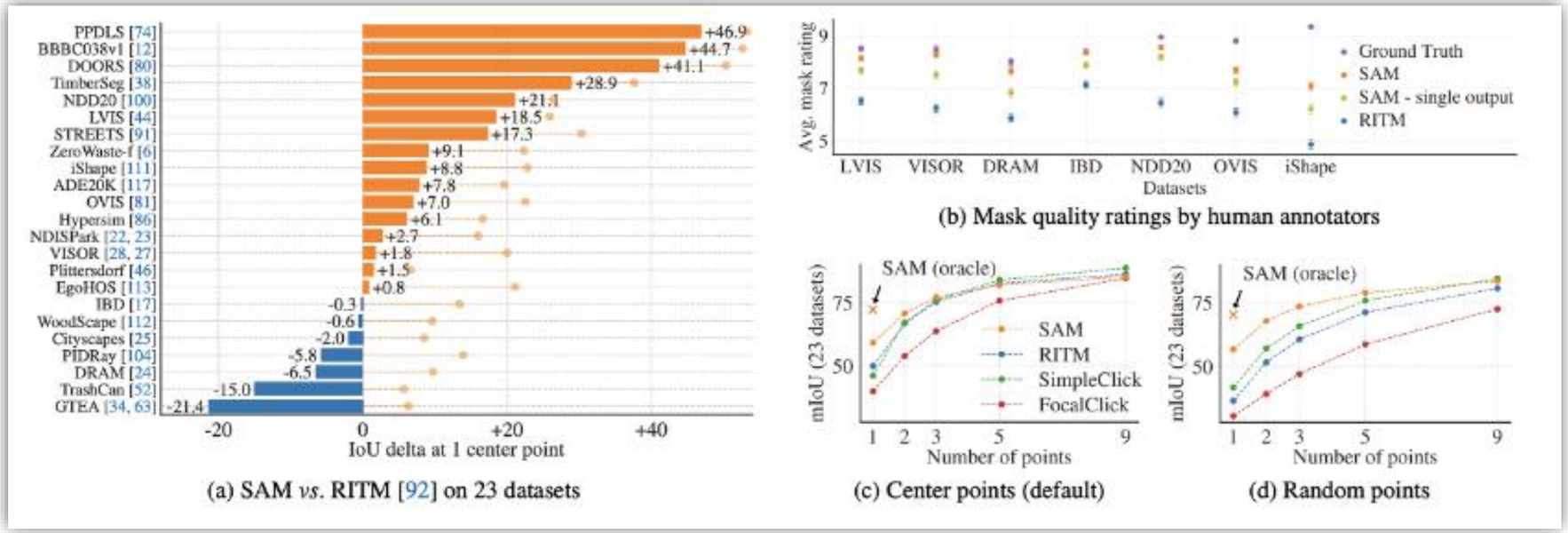


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

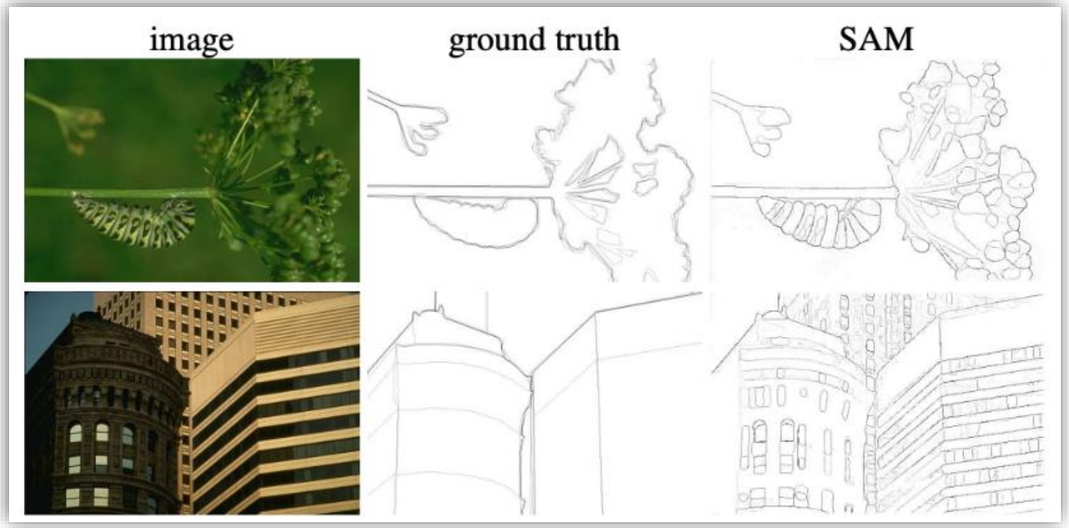
Experimental Results: Zero-shot Transfer

- Single Point Valid Mask Evaluation
- 23개의 데이터셋 중 16개에서 SOTA 모델인 **RITM보다 높은 성능**을 보임
- 연한 점으로 표시된 지표는 **Ambiguity를 고려**했을 때 SAM 모델이 얼만큼 더 뛰어난 성능을 보여주었는지
- (c), (d)에서는 prompt로 주는 점의 개수를 달리할 때의 결과. 개수가 많아질수록 그 차이가 적어지지만 점의 개수가 적을 때는 SAM이 월등한 성적을 보임.



Experimental Results: Zero-shot Transfer

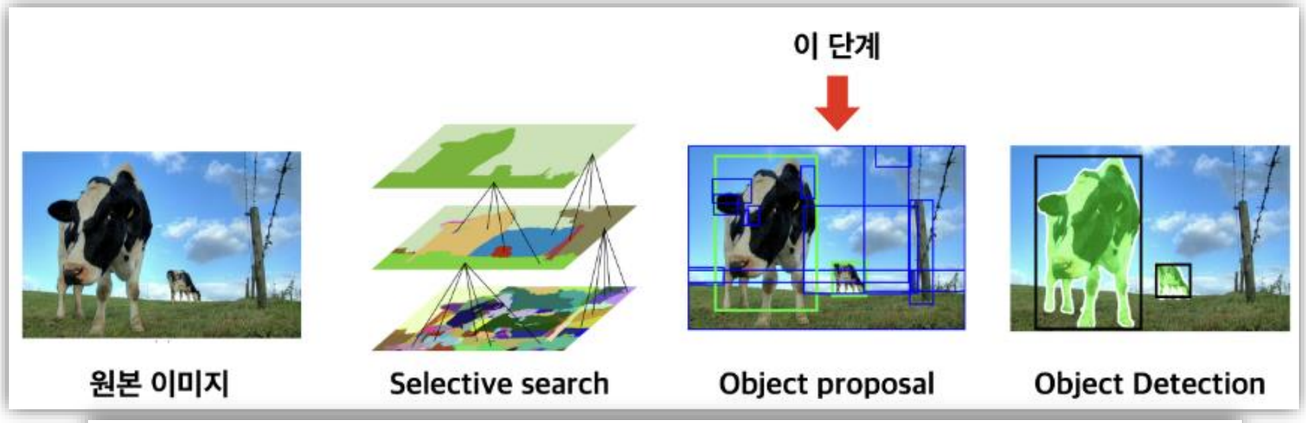
- Zero-Shot Edge Detection
- Edge Detection에 대해서는 전혀 학습시키지 않았음에도 불구하고 **준수한 성능을** 보임. SAM보다 좋은 성능을 낸 모델들은 테스트 데이터셋인 BSDS500의 트레이닝 데이터셋으로 학습된 것임을 감안하면, 아주 좋은 성능이라고 말함.

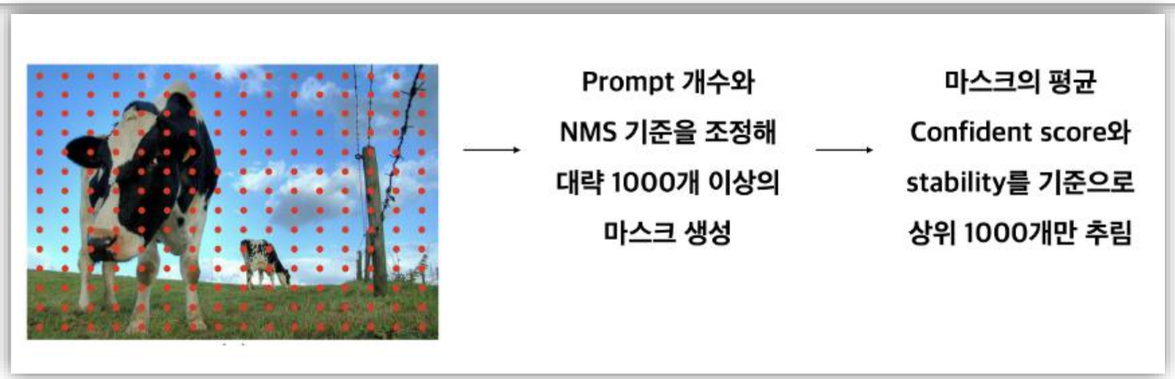


method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Experimental Results: Zero-shot Transfer

- **Object Proposal**
- object detection 분야에서 많이 연구된 분야로 물체가 있을만한 후보 영역을 찾는 태스크
- 작은 물체를 제외하고는 SOTA 시스템과 비슷한 성능
- rare한 물체는 더 잘 찾는 것으로 보아 더 general한 특징

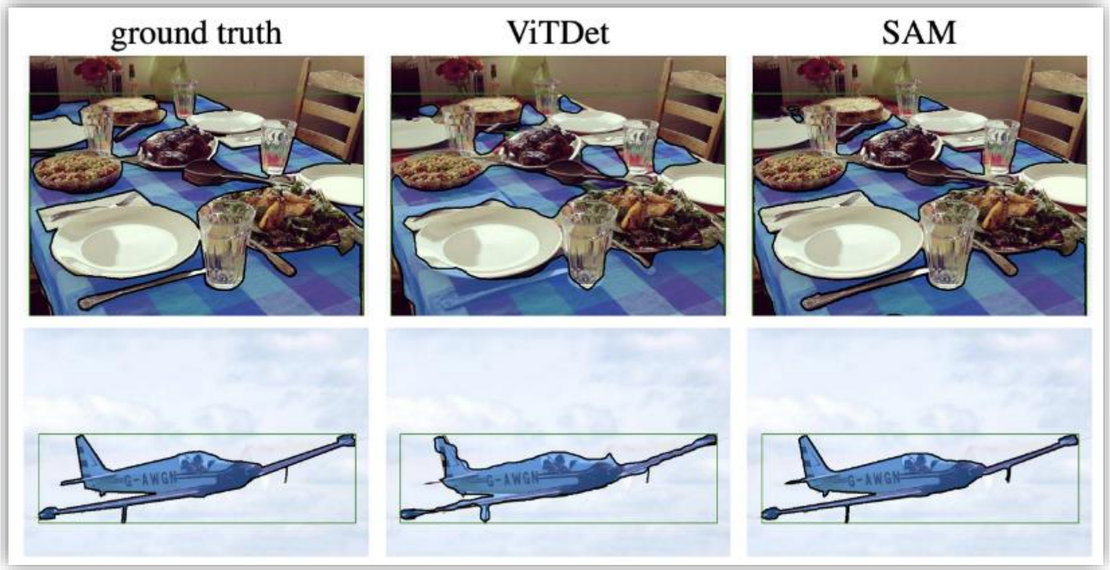




	mask AR@1000						
method	all	small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
zero-shot transfer methods:							
SAM – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Experimental Results: Zero-shot Transfer

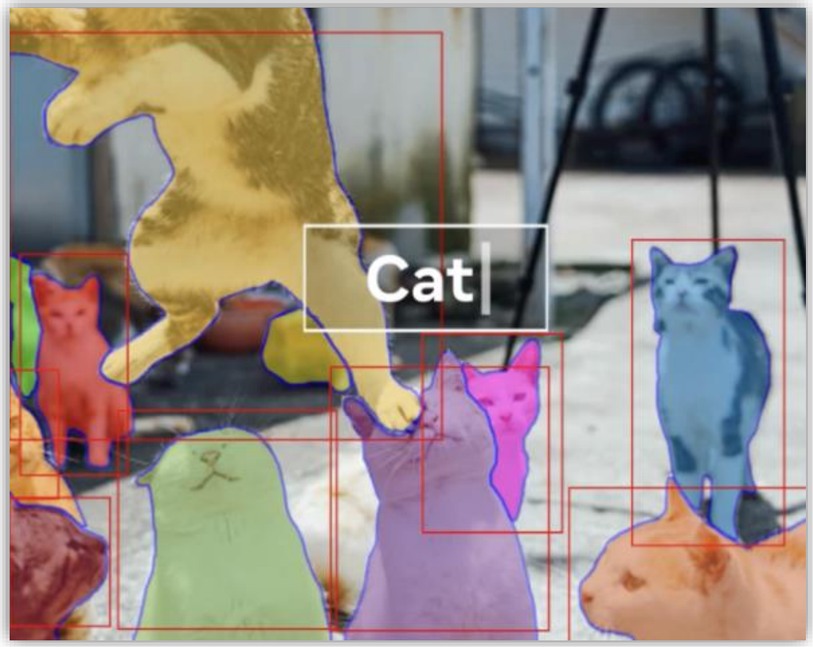
- **Instance Segmentation**
- Object Detection 결과로 출력된 물체에 대해서 Segmentation하는 태스크
- (c), (d)에서는 prompt로 주는 점의 개수를 달리할 때의 결과. 개수가 많아질수록 그 차이가 적어지지만 점의 개수가 적을 때는 SAM이 월등한 성적을 보임.

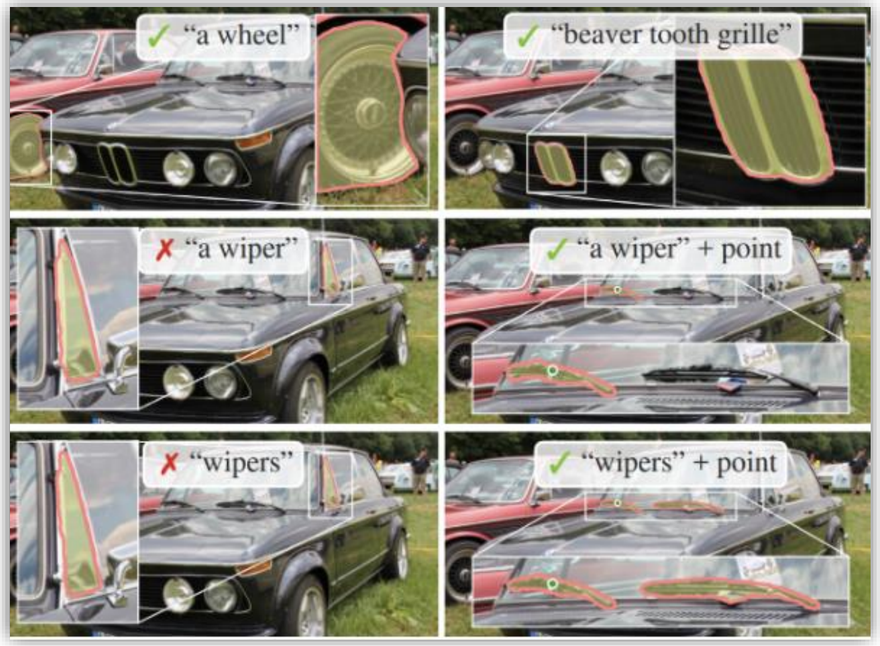


	COCO [66]				LVIS v1 [44]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

Experimental Results: Zero-shot Transfer

- Text to mask
- Text로 Segmentation하는 태스크
- 정량적인 평가가 논문에서 빠졌고, 데모에서도 제외된 것으로 보아 성능이 뛰어나진 않았을 것으로 추측





Conclusion

- Computer Vision계의 Foundation Model을 만들자는 문제 의식에서 출발해서 Task, Data, Model을 직접 고안
- 한번 Image embedding을 얻은 후에는 prompt를 통해 아주 빠르고 자유롭게 마스크를 만들 수 있으니 활용성까지 잡음

- 다음에 할 논문 : SAM + Tracking을 이용한 Track Anything (TAM) 논문
- https://arxiv.org/pdf/2304.11968.pdf