Segment Anything

Alexander Kirillov et al.

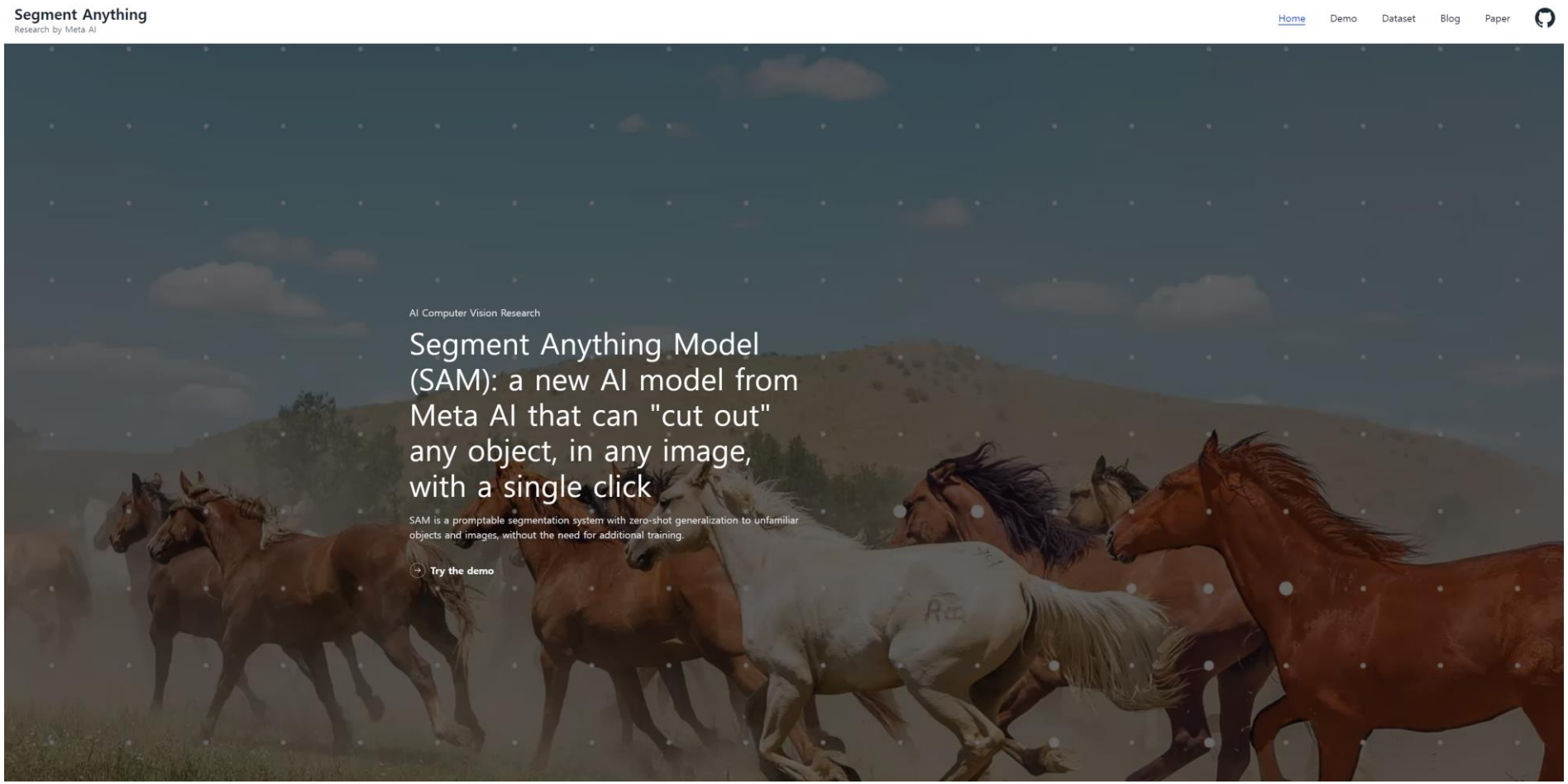
Meta AI Research, FAIR

arXiv

Presented by Minho Park

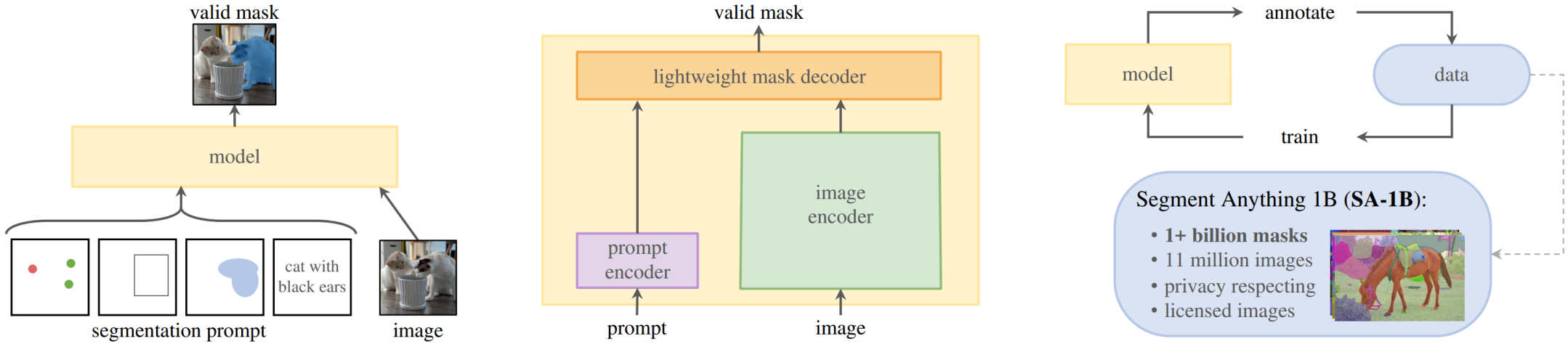
Segment Anything: Demo

- https://segment-anything.com/



Segment Anything Project

- **Segment Anything (SA) project: Segmentation에서 foundation model을 만들어보자.**
- NLP처럼 대규모 데이터로 하나의 pre-training task를 학습한 후 prompt-tuning을 통해 zero-shot transfer를 해결해보겠다.

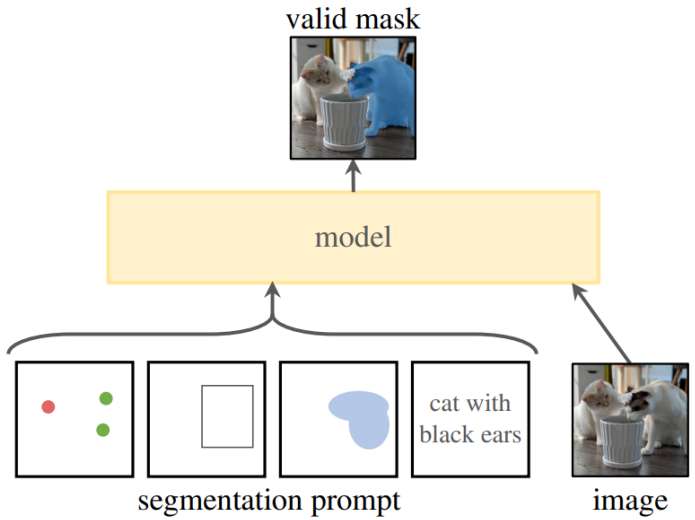


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 1: We aim to build a foundation model for segmentation by introducing three interconnected components: a prompt-able segmentation *task*, a segmentation *model* (SAM) that powers data annotation and enables zero-shot transfer to a range of tasks via prompt engineering, and a *data* engine for collecting SA-1B, our dataset of over 1 billion masks.

Task: Promptable Segmentation Task

- **Promptable segmentation: “prompt”가 들어올 때, “valid한 mask”를 제공함.**
- Input (Prompt): segment를 가리킬 수 있는 정보
  - pos/neg points, a rough box or mask, free-form text 등
- Output (Valid mask): prompt를 통해 뽑힐 수 있는 segment 중 하나
  - 만약 가방을 가리키면 “가방” or “가방을 맨 사람” 둘 다 정답임.



(a) **Task**: promptable segmentation

셋 모두가 정답임
Pre-training에서 셋 중 하나를
말하면 정답으로 쳐줘야 함.

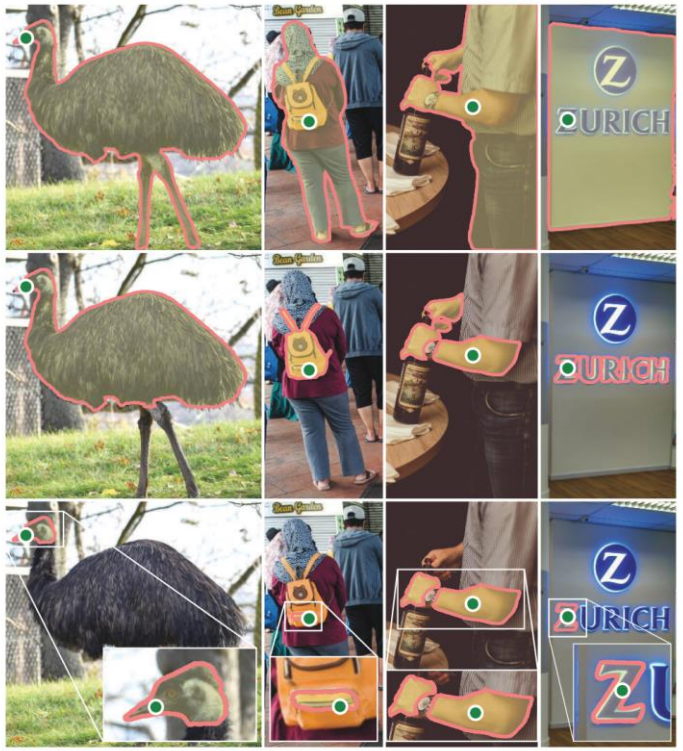
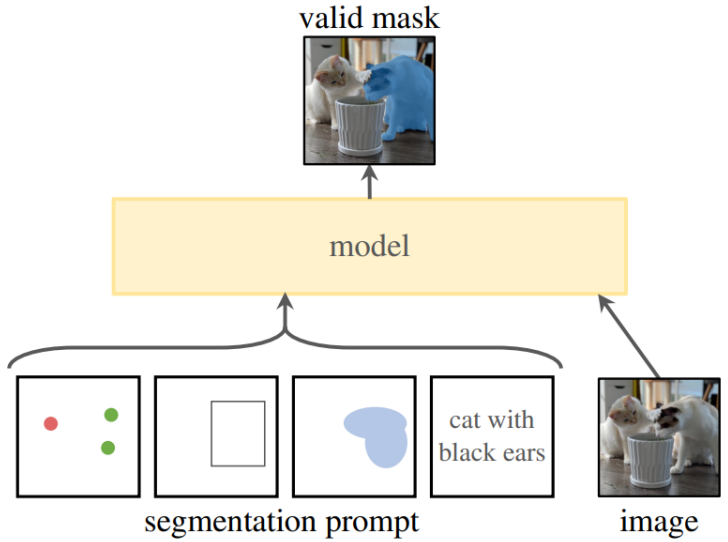


Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

Task: Promptable Segmentation Task

- Multi-task segmentation systems: Fixed set of tasks만 수행 가능함.
  - I.e., training, test tasks가 동일.
  - E.g., joint semantic, instance, and panoptic segmentation
- Segment anything: Inference time에 새로운 다른 tasks도 수행 가능함.
  - Object detector로 bounding box를 prompt로 주면 instance segmentation이 가능함.

크게 네 가지 prompt를 먼저 정의하였음.



(a) **Task**: promptable segmentation

Model: Segment Anything Model

- SAM이 되기 위한 조건: 1) support flexible prompts, 2) amortized real-time, 3) ambiguity-aware.
- **SAM: Powerful image encoder + Prompt encoder + Lightweight mask decoder**

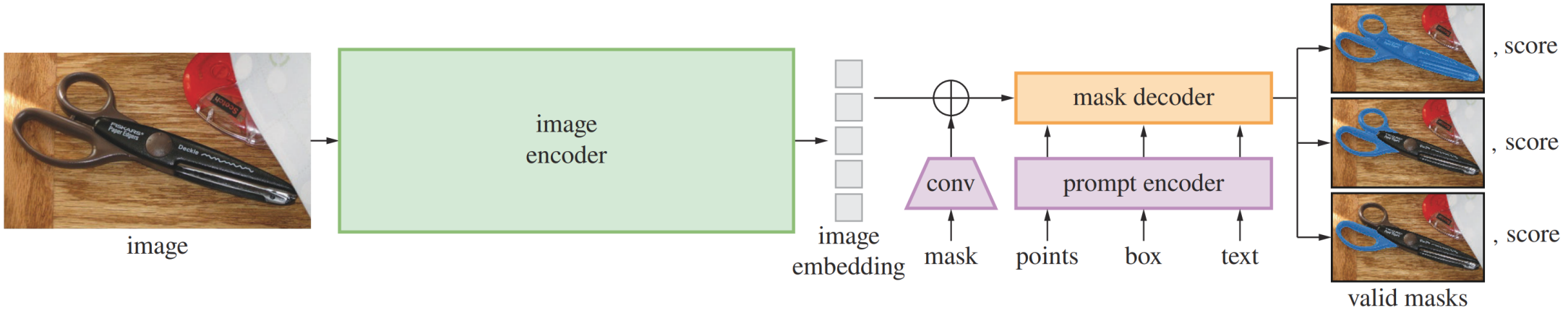


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Model: Segment Anything Model

- **Image encoder: MAE pre-trained Vision Transformer (ViT)**
- Image encoder는 이미지에 대해서 한 번만 통과시켜서 amortized real-time을 얻어냄.

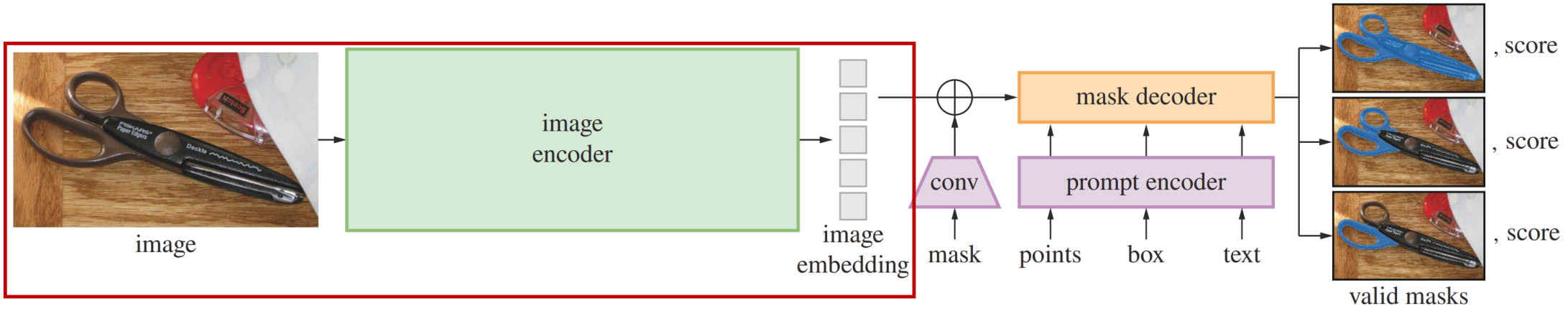


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Model: Segment Anything Model

- **Prompt encoder: Sparse (points, boxes, text) encoder and dense (masks) encoder.**
- Sparse encoder: positional encoding + learned embeddings for each prompt type
  - Text의 경우 pre-trained CLIP text encoder.
- Dense encoder: convolutions

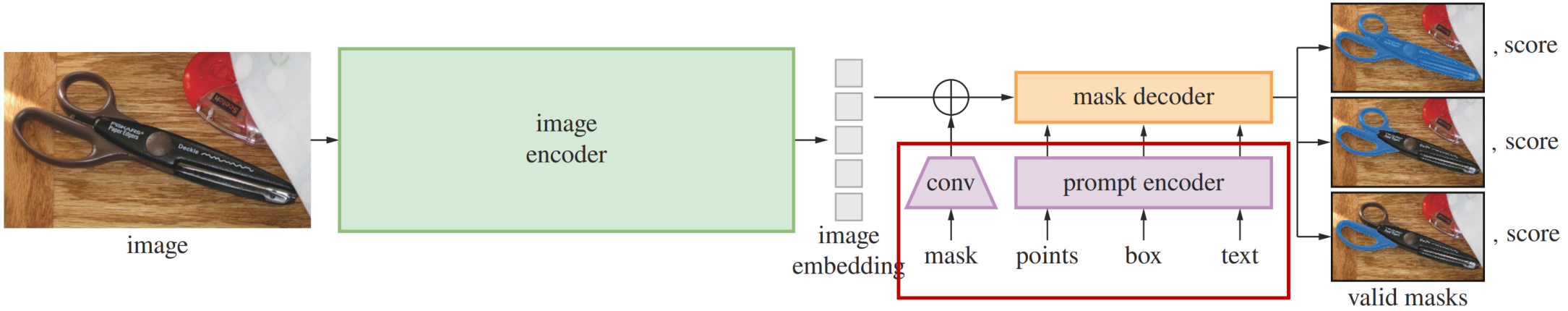


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Model: Segment Anything Model

- **Mask decoder: Modification of a Transformer decoder block** + dynamic mask prediction head
- Dense prompts는 image embedding에 더해지고 sparse prompts는 **"TwoWayTransformer"**를 통과
  - TwoWayAttentionBlock: SA (sparse prompt) + CA (prompt-to-image) + CA (image-to-prompt)
- 최종적으로 condition을 받은 image embedding을 사용함.

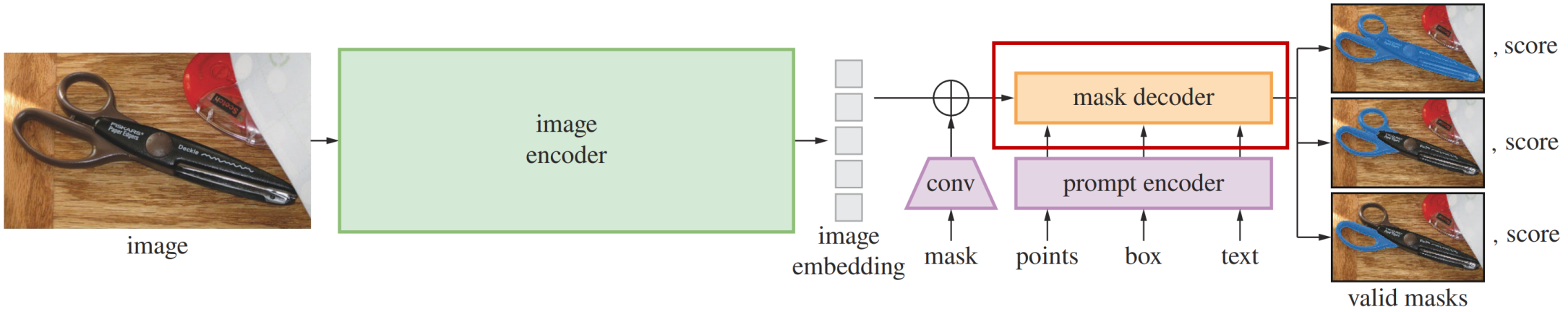


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

https://github.com/facebookresearch/segment-anything/blob/main/segment_anything/modeling/mask_decoder.py#L16
https://github.com/facebookresearch/segment-anything/blob/main/segment_anything/modeling/transformer.py#L16

Model: Segment Anything Model

- **Mask decoder:** Modification of a Transformer decoder block + **dynamic mask prediction head**
- 1) 생성된 image embedding을 upsampling
- 2) Ambiguity 해결을 위해 총 세 가지 출력을 내기 위한 dynamic mask prediction head.
  - Single prediction head는 이도 저도 아닌 애매한 segment를 출력함.

```
self.output_upscaling = nn.Sequential(
    nn.ConvTranspose2d(transformer_dim, transformer_dim // 4, kernel_size=2, stride=2),
    LayerNorm2d(transformer_dim // 4),
    activation(),
    nn.ConvTranspose2d(transformer_dim // 4, transformer_dim // 8, kernel_size=2, stride=2),
    activation(),
)
```

Upsampling layer (Transposed Conv + LN + Activation)

init

```
self.output_hypernetworks_mlps = nn.ModuleList(
    [
        MLP(transformer_dim, transformer_dim, transformer_dim // 8, 3)
        for i in range(self.num_mask_tokens)
    ]
)
```

forward

```
hyper_in_list: List[torch.Tensor] = []
for i in range(self.num_mask_tokens):
    hyper_in_list.append(self.output_hypernetworks_mlps[i](mask_tokens_out[:, i, :]))
hyper_in = torch.stack(hyper_in_list, dim=1)
b, c, h, w = upscaled_embedding.shape
masks = (hyper_in @ upscaled_embedding.view(b, c, h * w)).view(b, -1, h, w)
```

Hypernetwork MLP (self.num_mask_tokens = 3)

https://github.com/facebookresearch/segment-anything/blob/main/segment_anything/modeling/mask_decoder.py#L16
https://github.com/facebookresearch/segment-anything/blob/main/segment_anything/modeling/transformer.py#L16

Model: Segment Anything Model

- **Resolving ambiguity: Predict multiple output masks for a single prompt.**
- 대부분의 경우에 “whole”, “part”, “subpart” 세 가지 중 헷갈리는 경우가 많아 3개의 head를 달았음.
  - 학습할 때에는 셋 중 가장 작은 loss만 backward.
- Test-time에 셋 중 가장 confident한 sample을 출력하기 위해 score를 학습함.
  - GT는 실제 IoU.

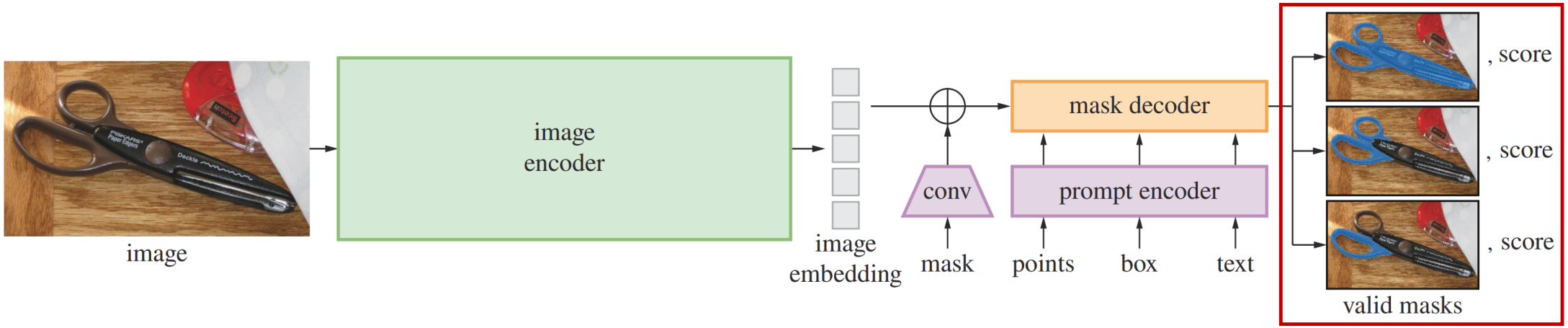
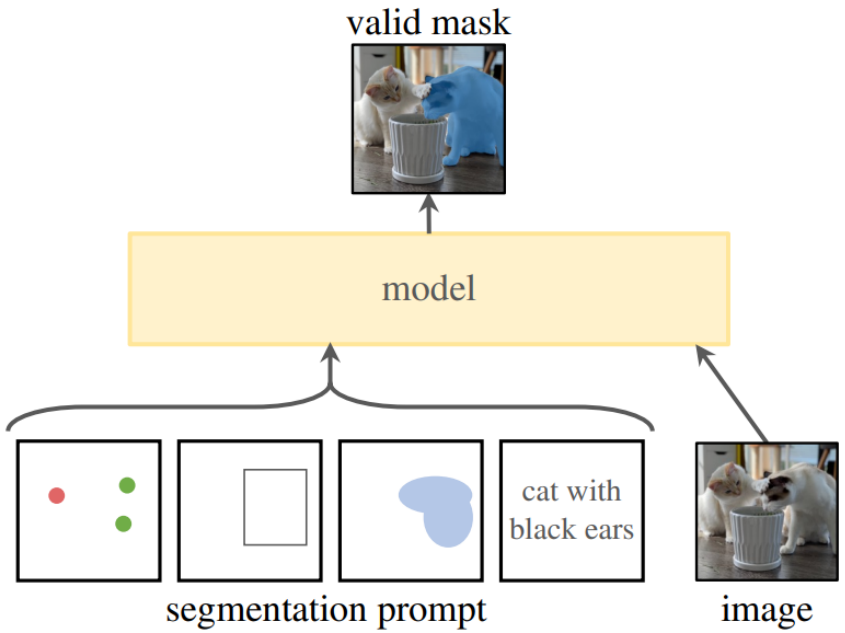


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

Model: Text-guided Segmentation

- Text-image pairs dataset을 segmentation mask 별로 구하기 어렵기 때문에, CLIP을 활용함.
- Text와 Image간의 embedding space가 거의 일치한다고 알려진 CLIP이기 때문에, **CLIP Image encoder를 이용하여 학습 후 inference에서 text를 CLIP text encoder에 통과시켜서 추론.**
  - **CLIP의 한계를 그대로 이어 받았을 것이기 때문에, 이 부분은 보완이 필요함.**



(a) **Task**: promptable segmentation

Segment Anything Data Engine

- Foundation model을 학습하기 위해서는 대규모 데이터셋이 필요함.
  - Segmentation은 crawling을 통해 얻을 수 없음.
- 대규모 데이터셋을 얻기 위해, data engine을 구축하였음.
  - 1) Assisted-manual stage
  - 2) Semi-automatic stage
  - 3) Fully automatic stage

Data Engine: Assisted-manual stage

- **일반적인 annotation 방법론처럼 foreground / background object points를 annotators가 직접 주어 labeling을 진행함.**
  - SAM은 일단 common public segmentation datasets으로 학습.
  - SAM은 미리 image embedding을 뽑아두어서 browser-based interactive segmentation tool로 사용함.
- “Stuff”, “things” 가리지 않고 자유롭게 labeling.
  - Mask당 30초 걸리도록 권장하였음 (quality control).

Point-interactive segmentation model을 이용한 labeling 과정 예시 (RITM)

Sofiiuk et al. "Reviving iterative training with mask guidance for interactive segmentation." *ICIP.* 2022.

Data Engine: Assisted-manual stage

- **Retraining strategy: 데이터가 모임에 따라서 새로운 annotated masks만으로 SAM을 다시 학습.**
  - Masks가 쌓이면서 MAE도 ViT-B에서 ViT-H로 바꿈.
  - **총 6번 재학습하였음.**
  - 재학습할수록 labeling 속도가 빨라져서 최종적으로 **34초에서 14초까지 줄었음.**
  - COCO mask labeling보다 6.5배 빠르며, bounding box labeling보다 2배만 느림.
- **We collected 4.3M masks from 120k images in this stage.**

Data Engine: Semi-automatic stage

- Assisted-manual stage보다 **masks의 diversity를 높이기 위해 제안**됨.
- SAM에 pre-trained object detector를 통해 모든 검출된 object를 segmentation하고 **annotator에게 더 추가할 objects가 없는지 묻는 식으로 진행.**
- Retraining strategy 사용 (5회)
- **We collected an additional 5.9M masks in 180k images (10.2M total).**
  - 이미지당 44장에서 72장으로 늘었음.
- Labeling하기에 더 challenging하기 때문에 다시 mask 당 34초 걸림.

Data Engine: Fully automatic stage

- Fully automatic하게 진행할 것임.
- 32 x 32 regular grid로 점을 주어서 이미지 내에 있는 모든 mask를 잡아낼 예정.
- 이제 ambiguity-aware model를 학습함.
- **Confident and stable masks를 수집할 것임.**
  - Confident: 높은 score의 mask만 수집
  - Stable: 0.5 – d, 0.5 + d로 각각 thresholding 했을 때 비슷한 결과를 내는 mask만 수집.
  - NMS를 통해 duplicate 제거
- 작은 mask를 잡아내기 위해서 multiple zoomed-in image crops를 segmentation.
- 최종적으로 11M 이미지에서 1.1B high-quality masks를 구축하였음.

Segment Anything Dataset

- Images: 3300 x 4950 pixels on average
  - 용량 문제로 짧은 변이 1500 pixels인 이미지로 downsampling 함.
- **SA-1B only includes automatically generated masks.**
  - 퀄리티를 측정 해봤을 때 생성된 데이터셋이 믿을만하기 때문에 1.1B 중 semi-automatic masks를 다 빼고 fully automatic stage에서 얻은 1B을 제공함.
- Mask quality:
  - We computed IoU between each pair and found that 94% of pairs have greater than 90% IoU.
  - For comparison, prior work estimates inter-annotator consistency at 85-91% IoU.
- 11x more images and 400x more masks than the second largest, Open Images.

Segment Anything Dataset

- Statistics

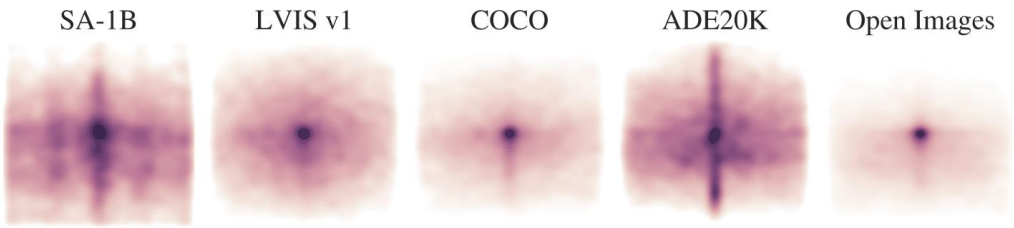


Figure 5: Image-size normalized mask center distributions.

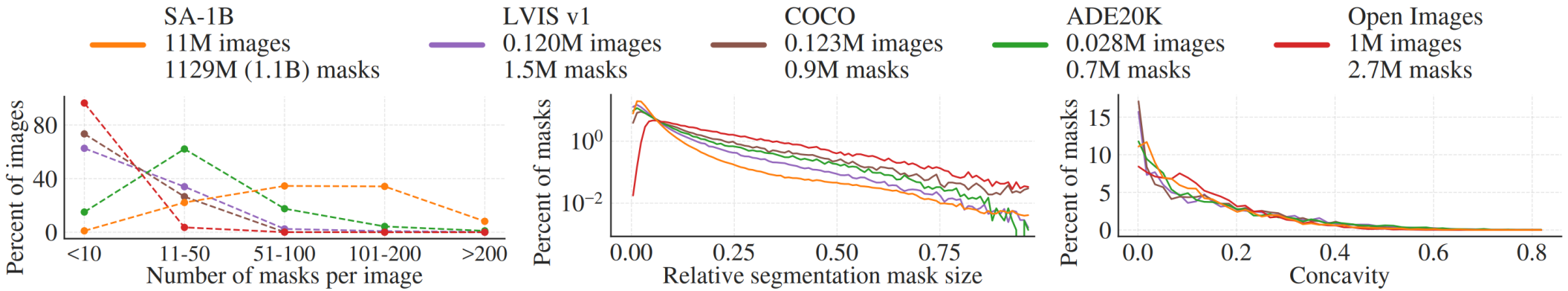
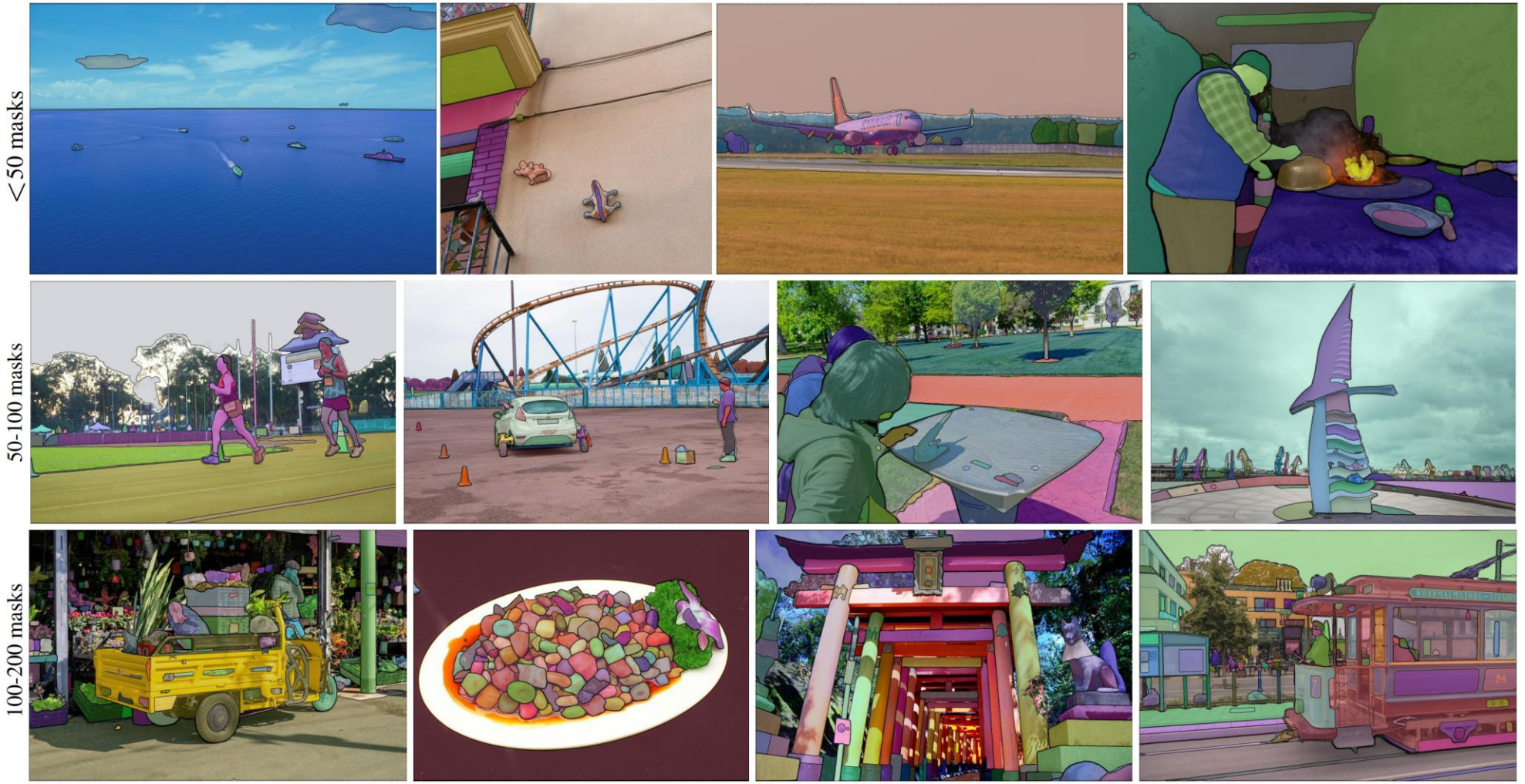


Figure 6: Dataset mask properties. The legend references the number of images and masks in each dataset. Note, that SA-1B has 11× more images and 400× more masks than the largest existing segmentation dataset Open Images [60].

Segment Anything Dataset



Segment Anything Dataset

Responsible AI

	# countries	SA-1B #imgs	SA-1B #masks	% images SA-1B	% images COCO	% images O.I.
Africa	54	300k	28M	2.8%	3.0%	1.7%
Asia & Oceania	70	3.9M	423M	36.2%	11.4%	14.3%
Europe	47	5.4M	540M	49.8%	34.2%	36.2%
Latin America & Carib.	42	380k	36M	3.5%	3.1%	5.0%
North America	4	830k	80M	7.7%	48.3%	42.8%
high income countries	81	5.8M	598M	54.0%	89.1%	87.5%
middle income countries	108	4.9M	499M	45.0%	10.5%	12.0%
low income countries	28	100k	9.4M	0.9%	0.4%	0.5%

Table 1: Comparison of geographic and income representation. SA-1B has higher representation in Europe and Asia & Oceania as well as middle income countries. Images from Africa, Latin America & Caribbean, as well as low income countries, are underrepresented in all datasets.

	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.

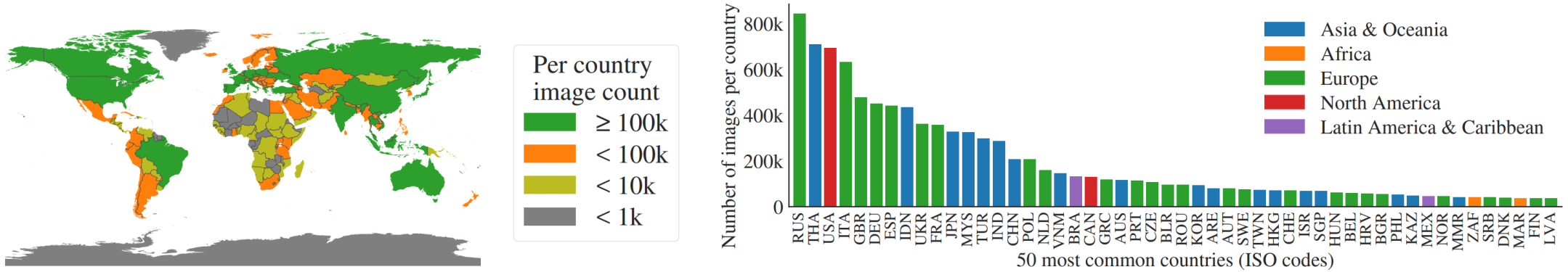


Figure 7: Estimated geographic distribution of SA-1B images. Most of the world's countries have more than 1000 images in SA-1B, and the three countries with the most images are from different parts of the world.

Experiments: Zero-shot Transfer

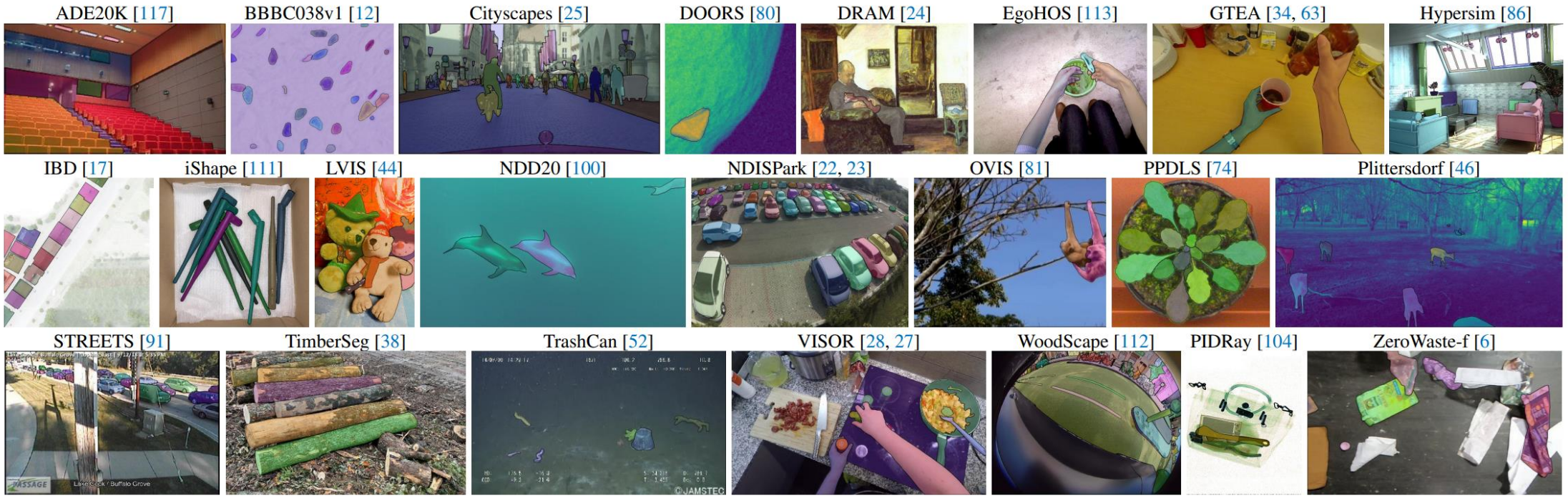
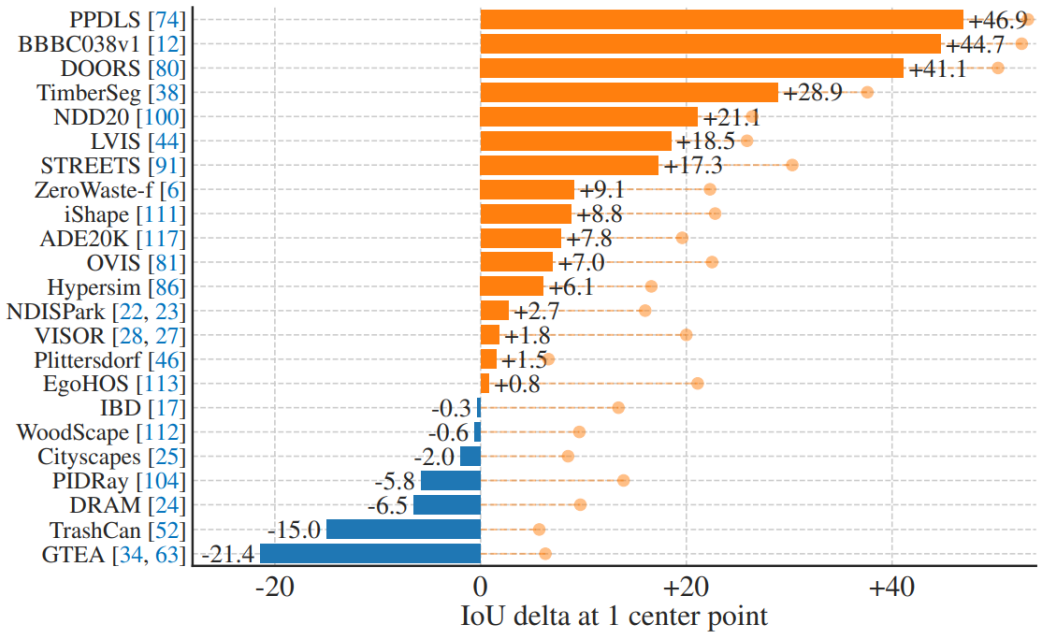
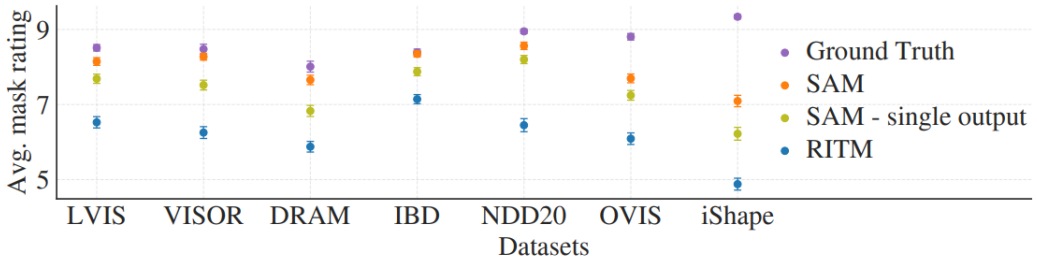


Figure 8: Samples from the 23 diverse segmentation datasets used to evaluate SAM's zero-shot transfer capabilities.

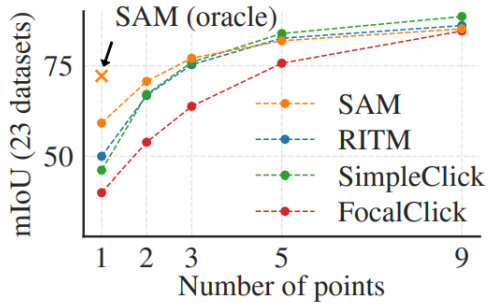
Experiments: Zero-shot Transfer



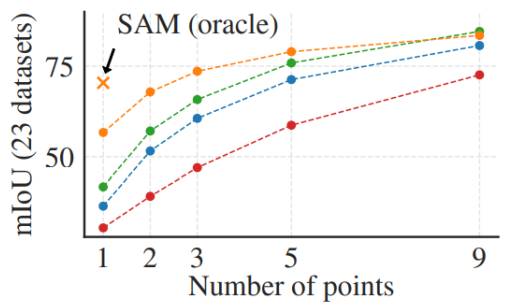
(a) SAM *vs*. RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



(c) Center points (default)



(d) Random points

Figure 9: Point to mask evaluation on 23 datasets. (a) Mean IoU of SAM and the strongest single point segmenter, RITM [92]. Due to ambiguity, a single mask may not match ground truth; circles show "oracle" results of the most relevant of SAM's 3 predictions. (b) Per-dataset comparison of mask quality ratings by annotators from 1 (worst) to 10 (best). All methods use the ground truth mask center as the prompt. (c, d) mIoU with varying number of points. SAM significantly outperforms prior interactive segmenters with 1 point and is on par with more points. Low absolute mIoU at 1 point is the result of ambiguity.