Zero-Shot Transfer Experiments

Source 논문명: Segment Anything in High Quality (https://arxiv.org/pdf/2005.05535.pdf)

Overview

- Qualitative Results (질적 비교)
- Open-world Dataset Results
- Famous Baseline Datasets Results
- Results depending on number of input points
- Results on Video Instance Segmentation

Qualitative Results:

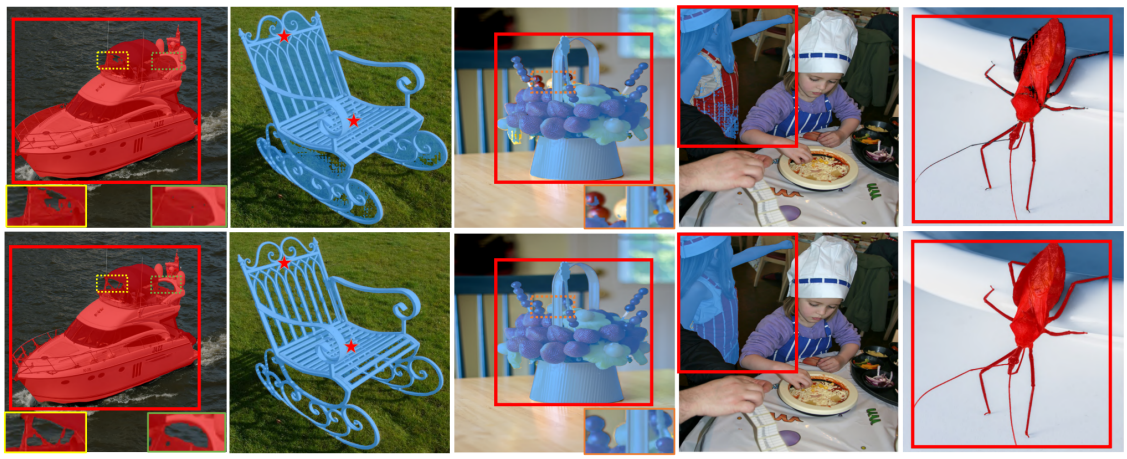


Figure 4: Visual results comparison between SAM (top row) **vs.** HQ-SAM (bottom row) in a *zero-shot transfer setting*, given the same red box or point prompt. HQ-SAM produces significantly more detailed-preserving results and also addresses the mask errors with broken holes.

💡 SAM보다 더 미세한 부분 (ex. 복잡한 디자인의 배경과 섞인 가는 영역)을 더 잘 segment 한다.

In Open-World Setting

[Dataset]

Segmentation에 있어 challenging하다고 알려진 UVO데이터셋 활용

Figure 3: **Examples of UVO**. UVO videos are exhaustively annotated with masks regardless of object categories. UVO features a wide-range of videos (*e.g.*, third-person/egocentric, professional/amateur, crowded/sparse objects) making it a challenging benchmark. Best viewed in color.

UVO dataset

[Result]

Table 5: Zero-shot open-world instance segmentation results comparison on UVO [41]. We use FocalNet-DINO [51] trained on the COCO dataset as our box prompt generator. $*^{strict}$ denotes the boundary region with a tighter threshold.

Model	AP_B^{strict}	AP_{B75}^{strict}	AP_{B50}^{strict}	AP_B	AP_{B75}	AP_{B50}	AP
SAM	8.6	3.7	25.6	17.3	14.4	37.7	29.7
HQ-SAM	**9.9**	**5.0**	**28.2**	**18.5**	**16.3**	**38.6**	**30.1**

: 더 좋은 성능을 보였다!

Big, high-resolution Dataset

[Dataset]

매우 큰 크기(**2048×1600 ~ 5000×3600)**의 이미지로 구성된 데이터셋에 비교 실험 적용

[Result]

Table 6: Zero-shot segmentation result comparison on the test set of high-quality BIG [6] benchmark using various types of input prompts. We employ PSPNet [53] to generate the coarse mask prompt.

Model	GT Box Prompt		Mask Prompt	
	mIoU	mBIoU	mIoU	mBIoU
SAM	81.1	70.4	66.6	41.8
HQ-SAM	**86.0**	**75.3**	**86.9**	**75.1**

결과: 두 실험(Ground-Truth object box를 제공했을 때 & 다른 inference 모델로 대략의 mask를 제공했을 때)에서 모두 SAM보다 더 나은 결과를 보였으나 특히 coarse mask를 제공했을 때 SAM에 비교적 더 robust한 성능을 보였음

Conventional baseline datasets

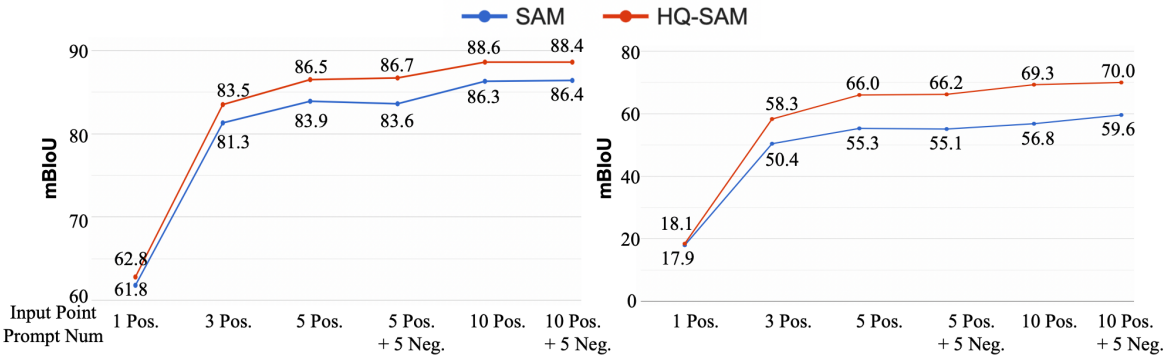
: 흔히 쓰이는 데이터셋인 COCO와 LVIS에도 활용해보았다고 밑의 [Result]에 보이다시피 SAM보다 더 좋은 성능을 기록했다.

[Result]

Table 7: Zero-shot instance segmentation results comparison on COCO [31] and LVISv1 [14]. For the COCO dataset, we use FocalNet-DINO [51] detector trained on COCO. For LVIS, we adopt ViTDet-H [28] trained on the LVIS dataset as our box prompt generator. For SAM, we use the ViT-L backbone and box prompt. We maintain the zero-shot segmentation capability of the original SAM while improving the mask quality on the boundary region.

Model	COCO		LVIS				
	AP_B	AP	AP_B^{strict}	AP_{B75}^{strict}	AP_B	AP_{B75}	AP
SAM	33.3	48.5	32.1	32.8	38.5	40.9	43.6
HQ-SAM	**34.4**	**49.5**	**32.5**	**33.5**	**38.8**	**41.2**	**43.9**

Input Points 수에 따른 성능 변화



(a) Performance comparison on **COIFT** (b) Performance comparison on **DIS**

Figure 5: Interactive segmentation results comparison using a varying number of input points on the COIFT [29] (zero-shot) and DIS [34] val set. HQ-SAM consistently outperforms SAM with various point numbers, and the relative improvement is more obvious with less prompt ambiguity.

결과: input prompt에서의 points수가 증가하는 것과 비례하게 성능이 좋아지는 것은 SAM과 HQ-SAM 모두 동일했다. 또한, 두 데이터셋(COIFT, DIS)에서 적용해본 모든 실험에서 SAM을 능가했고 input points에 대한 정보가 많을 수록(x축의 우측에 갈 수록) HQ-SAM의 성능이 SAM의 성능보다 더 뛰어난 것을 관찰했다.

Zero-shot High-quality Video Instance Segmentation

Table 8: Zero-shot Video Instance Segmentation comparison on the test set of the very accurately labeled HQ-YTVIS [20] benchmark. We utilize pre-trained Swin-L-based Mask2Fromer [4] on YTVIS [46] as our box prompt input while reusing its object association prediction.

Model	AP^B	AP^B_{75}	AP^B_{50}	AP^M	AP^M_{75}	AP^M_{50}
SAM	30.2	19.1	72.9	60.7	68.1	90.5
HQ-SAM	**34.0**	**24.3**	**79.5**	**63.6**	**70.5**	**91.1**

prompts and feed it into SAM and our HQ-SAM for mask prediction. In Table 8, HQ-SAM achieves remarkable gains of 3.8 points in Tube Boundary AP^B and 2.9 Tube Mask AP^M.

결과: 다른 실험과 마찬가지로 SAM을 능가했다.

⚠️ 한계점: SAM과 마찬가지로 VIT기반의 인코더를 활용하기 때문에 real-time으로 활용하기엔 느리다는 한계가 있다.