YOLOE: Real-Time Seeing Anything

Ao Wang[1*] Lihao Liu[1*] Hui Chen[2] Zijia Lin[1] Jungong Han[3] Guiguang Ding[1,†]
[1]School of Software, Tsinghua University [2]BNRist, Tsinghua University
[3]Department of Automation, Tsinghua University

A. More Implementation Details

Data. We employ Objects365[6], GoldG [3] (including GQA[2] and Flickr30k [4]) for training YOLOE. Tab. 1 present their details. We utilize SAM-2.1-Hiera-Large [5] to generate high-quality pseudo labeling of segmentation masks with ground truth bounding boxes as prompts. We filter out ones with too few areas. To enhance the smoothness of mask edges, we apply Gaussian kernel to masks, using 3×3 and 7×7 kernels for small and large ones, respectively. Besides, we refine the masks following [1], which iteratively simplifies the mask contours. This reduces noise pixels while preserving overall structure.

Table 1. Data details for YOLOE training.

Dataset	Type	Box	Mask	Images	Anno.
Objects365 [6]	Detection	✓	✓	609k	8,530k
GQA [2]	Grounding	✓	✓	621k	3,662k
Flickr [4]	Grounding	✓	✓	149k	638k

Training. For all models, we adopt AdamW optimizer with an initial learning rate of 0.002. The batch size and weight decay are set to 128 and 0.025, respectively. The data augmentation includes color jittering, random affine transformations, random horizontal flipping, and mosaic augmentation. During transferring to COCO, for both *Linear probing* and *Full tuning*, we utilize the AdamW optimizer with an initial learning rate of 0.001, setting the batch size and weight decay to 128 and 0.025, respectively.

B. More Analyses for LRPC

To qualitatively show the efficacy of LRPC strategy, we visualize the number of anchor points retained for category retrieval after filtering. We present their average count under varying filtering threshold δ on the LVIS `minival` set in Fig. 1. It reveals that as δ increases, the number of retained anchor points decreases substantially across different models. This reduction significantly lowers computational overhead compared with the baseline scenario, which employs a total of 8400 anchor points. For example, for YOLOE-v8-S, with $\delta = 0.001$, the number of valid anchor points is reduced by 80%, enjoying 1.7× inference speedup with the same performance (see Tab. 7 in the paper). The results further confirm the notably redundancy of anchor points for category retrieval and verify the efficacy of LRPC.

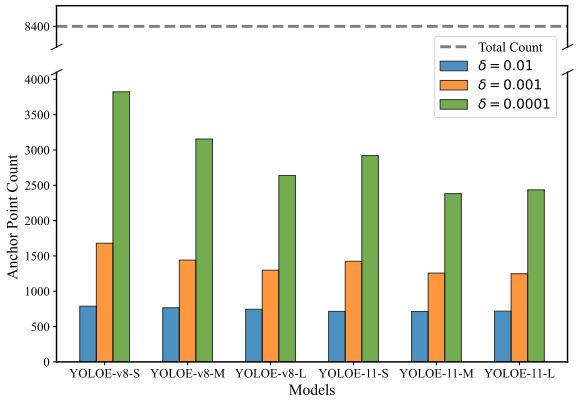

Figure 1. The count of retained anchor points under different filtering thresholds in LRPC. The dashed line means the total number.

C. More Visualization Results

To qualitatively show the efficacy of YOLOE, we present more visualization results for it in various scenarios.

Zero-shot inference on LVIS. In Fig. 2, we present the zero-shot inference capabilities of YOLOE on the LVIS. By leveraging the 1203 category names as text prompts, the model demonstrates its ability to detect and segment diverse objects across various images.

Prompt with customized texts. Fig. 3 presents the results with customized text prompts. We can see that YOLOE can interpret both generic and specific textual inputs, enabling precise object detection and fine-grained segmentation. Such capability allows users to tailor the model's behavior to meet specific requirements by defining

*Equal contributions. † Corresponding author.

input prompts at varying levels of granularity.

Prompt with visual inputs. In Fig. 4, we present the results of YOLOE with visual inputs as prompt. The visual inputs can take various forms, such as bounding box, point, or handcrafted shape. It can also be provided across the images. We can see that with visual prompt indicating the target object, YOLOE can accurately find other instances of the same category. Beside, it performs well across different objects and images, exhibiting robust capability.

Prompt-free inference. Fig. 5 shows the results of YOLOE with the prompt-free paradigm. We can see that in such setting, YOLOE achieves effective identification for diverse objects. This highlights its practicality in scenarios where predefined inputs are unavailable or impractical.

References

[1] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10 (2):112–122, 1973. 1

[2] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[3] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 1

[4] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[6] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

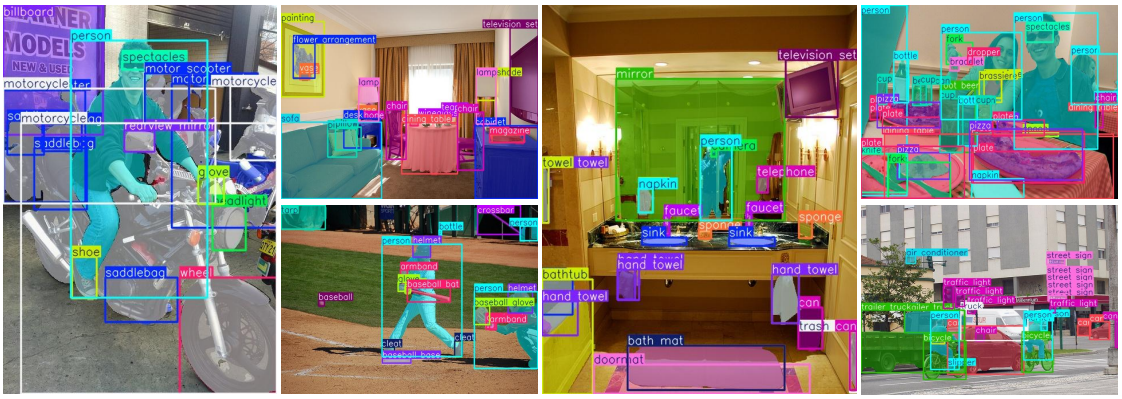

Figure 2. Zero-Shot inference on LVIS. The categories of LVIS are provided as text prompts.

{food, drink} {red car, black car, yellow car} {rock, tree, river, animal}

Figure 3. Prompt with customized texts. YOLOE adapts to both generic and specific text prompts for flexible usage.

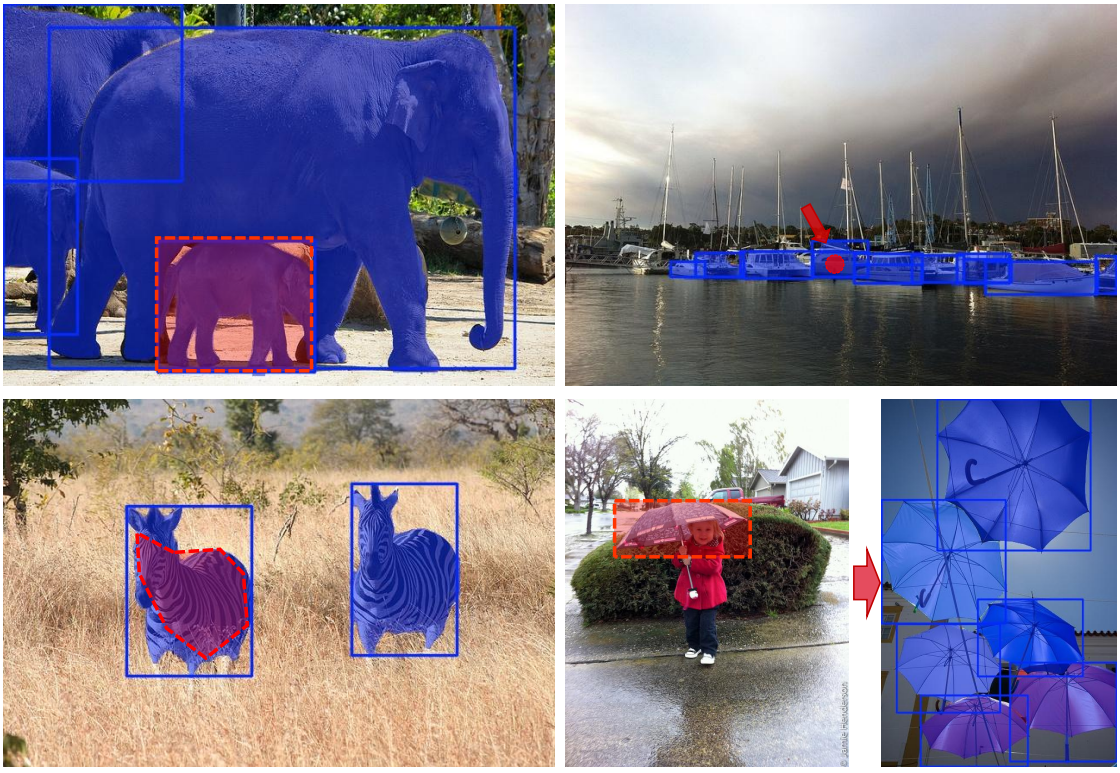

Figure 4. Prompt with visual inputs. YOLOE demonstrates the ability to identify objects guided by various visual prompts, like bounding box (top left), point (top right), handcrafted shape (bottom left). The visual prompt can also be applied across images (bottom right).

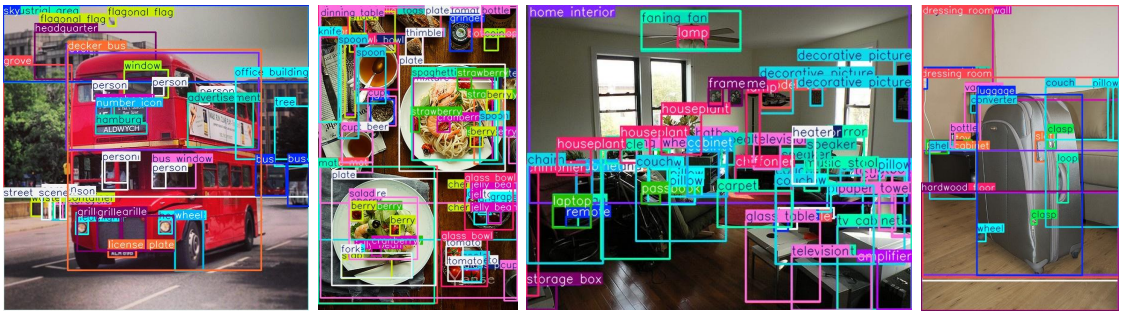

Figure 5. Prompt-free inference (omitting segmentation masks for clearer visualization). YOLOE can find diverse objects without prompt.