Unsupervised Open-Vocabulary Object Localization in Videos

Ke Fan[1,*], Zechen Bai[2,*], Tianjun Xiao[2], Dominik Zietlow[2], Max Horn[2], Zixu Zhao[2],
Carl-Johann Simon-Gabriel[2], Mike Zheng Shou[3], Francesco Locatello[2],
Bernt Schiele[2], Thomas Brox[2], Zheng Zhang[2,†], Yanwei Fu[1,†], Tong He[2]
[1]Fudan University [2]Amazon Web Services [3]National University of Singapore
kfan21@m.fudan.edu.cn, {baizeche, tianjux, zhaozixu, zhaz, htong}@amazon.com
{zietld, hornmax, cjsg, locatelf, bschiel, brox}@amazon.de
mikeshou@nus.edu.sg, yanweifu@fudan.edu.cn

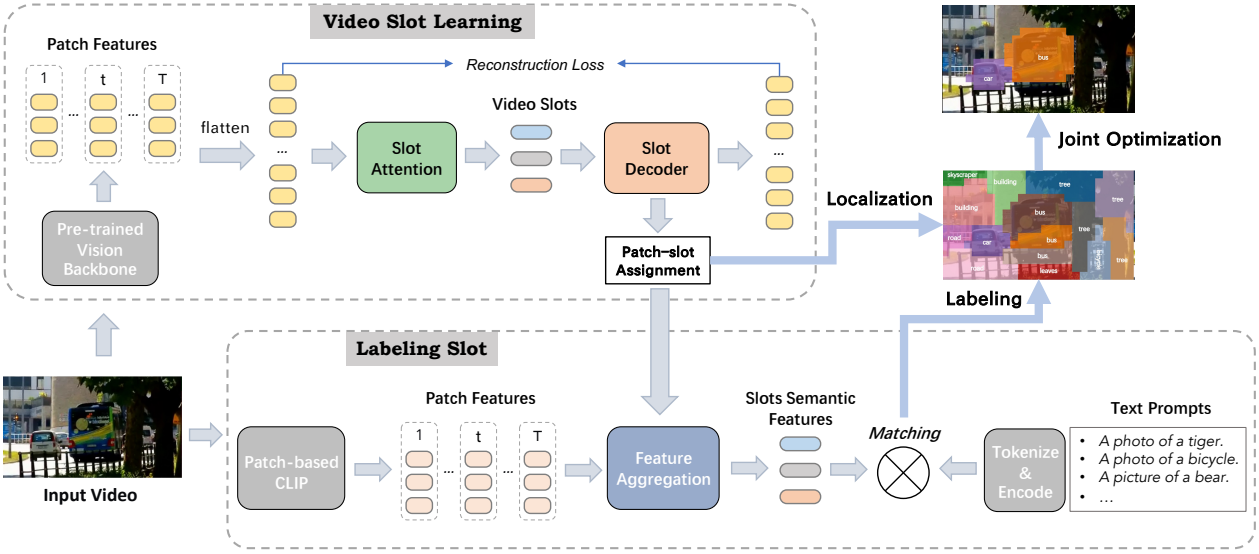

Figure 1. Given an input video, we first localize objects by slot attention with a video encoder pretrained with self-supervision. Next, we extract semantic features for each slot by a patch-based CLIP finetuned from its vanilla version. Then, slots are named by matching slot semantic features to text features from a curated list of text prompts. Finally, the named slots are optimized to alleviate over-segmentation caused by part-whole hierarchies.

In this paper, we show that recent advances in video representation learning and pre-trained vision-language models allow for substantial improvements in self-supervised video object localization. We propose a method that first localizes objects in videos via an object-centric approach and then assigns text to the obtained objects.

To localize the object with frame consistency, we extend slot attention to video-level grouping directly without using transition function to model temporal dynamics. Similar to DINOSAUR, we reconstruct extracted features from a pre-trained self-supervised vision backbone instead of pixels.

To properly name each object, we rely on the CLIP model to match slots with text features. However, the ordinary CLIP is based on contrastive learning between class tokens, making it hard to locate the semantic information for each patch of the image correctly. To deal with this problem, we propose a method to adapt the pre-trained CLIP visual encoder to a *patch-based CLIP* visual encoder using only *unlabeled images*. For further improvement, we utilized the semantic information to filter out the background and focus on the foreground objects of our interests, and merge the adjacent foregrounds with similar semantics.

The resulting video object localization is entirely unsupervised apart from the implicit annotation contained in CLIP, and it yields good results on regular video benchmarks.

∗ Equal contribution; † Corresponding authors; This paper has been presented on ICCV 2023.