4D Unsupervised Object Discovery

Yuqi Wang[1,2] Yuntao Chen[3] Zhaoxiang Zhang[1,2,3]
[1] CASIA [2] UCAS [3] HKISI_CAS
{wangyuqi2020,zhaoxiang.zhang}@ia.ac.cn

Abstract

Object discovery is a core task in computer vision. While fast progresses have been made in supervised object detection, its unsupervised counterpart remains largely unexplored. With the growth of data volume, the expensive cost of annotations is the major limitation hindering further study. Therefore, discovering objects without annotations has great significance. However, this task seems impractical on still-image or point cloud alone due to the lack of discriminative information. Previous studies underlook the crucial temporal *information and constraints naturally behind multi-modal inputs. In this paper, we propose* 4D unsupervised object discovery*, jointly discovering objects from 4D data – 3D point clouds and 2D RGB images with temporal information. We present the first practical approach for this task by proposing a* ClusterNet *on 3D point clouds, which is* jointly iteratively optimized *with a 2D localization network. Extensive experiments on the large-scale Waymo Open Dataset suggest that the localization network and ClusterNet achieve competitive performance on both class-agnostic 2D object detection and 3D instance segmentation, bridging the gap between unsupervised methods and full supervised ones. Codes and models will be made available at* https://github.com/Robertwyq/LSMOL.

Method Overview

As shown in Figure 1, the input is a set of video clips recorded in both 2D video frames I^t and 3D point clouds P^t at frame t during training. Since the point cloud and image data provide complementary information about location and appearance, they can serve as the natural cues guiding the training process mutually. During inference, the trained localization network L_{θ_1} is applied to still-image for 2D object detection, and the trained ClusterNet N_{θ_2} is applied to the point cloud for 3D instance segmentation.

$$\theta_1^*, \theta_2^* = \arg\min_{\theta_1,\theta_2} f(L_{\theta_1}(I^t), N_{\theta_2}(P^t), t) \qquad (1)$$

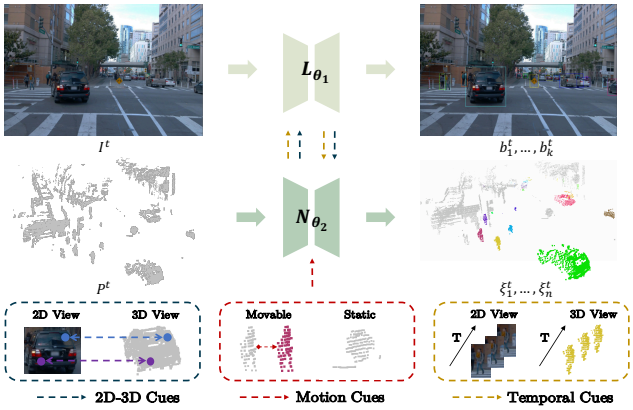

Figure 1. The pipeline of 4D unsupervised object discovery. The input is the corresponding 2D frames and 3D point clouds. The task needs to discover objects on both images and point clouds without manual annotations. The overall process can be divided into two steps: (1) 3D instance initialization and (2) joint iterative optimization. (1) 3D instance initialization: motion cues serve as the initial cues for training the ClusterNet. (2) Joint iterative optimization: the localization network and ClusterNet are optimized jointly by 2D-3D cues and temporal cues.

The algorithm can be formulated into a joint optimization function f in Eq.1. θ_1 and θ_2 are the parameters of the network need to optimize. Temporal information t serve as the natural constraint in function f. The localization network L_{θ_1} utilized Faster R-CNN as default. We propose a ClusterNet N_{θ_2} for 3D instance segmentation. The major challenge is the optimization for function f without annotations. To overcome the challenge, we seek for *motion cues*, *2D-3D cues* and *temporal cues* to serve as the supervision. All these cues are extracted naturally in the informative 4D data. (1) *motion cues*, represented as 3D scene flow, can distinguish movable segments from the background. It uses to train the ClusterNet N_{θ_2} initially. (2) *2D-3D cues*, reflecting the mapping between LiDAR points and RGB pixels, can be used as a bridge to optimize the L_{θ_1} and N_{θ_2} iteratively. It indicates the output of either network can be further used to optimize another network. (3) *temporal cues*, encouraging the temporal-consistent discovery in 2D and 3D view, can serve as the constraint to optimize the function together.