Discovering Objects that can Move

Zhipeng Bao, Pavel Tokmakov, Allan Jabri
Yu-Xiong Wang, Adrien Gaidon, Martial Hebert

Carnegie Mellon University · Toyota Research Institute · Berkeley University of California · University of Illinois Urbana-Champaign

CVPR June 19-24 2022 New Orleans Louisiana

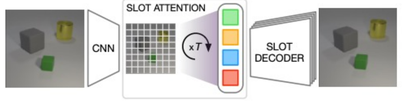
Code available

Object discovery

- Object discovery: separate objects from background without manual labels

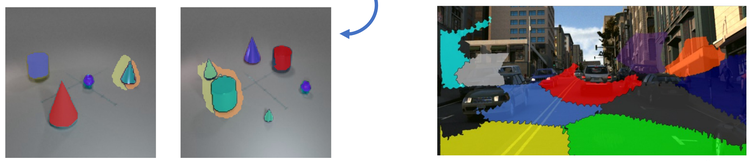
Baseline for object discovery

Slot attention (NeurIPS 20)



- Slots: 1D abstract object representations
- Learned with iterative spatial attention

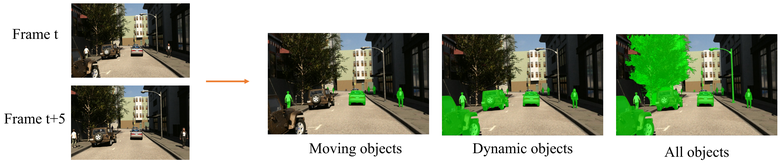
- Works well on a simple toy dataset



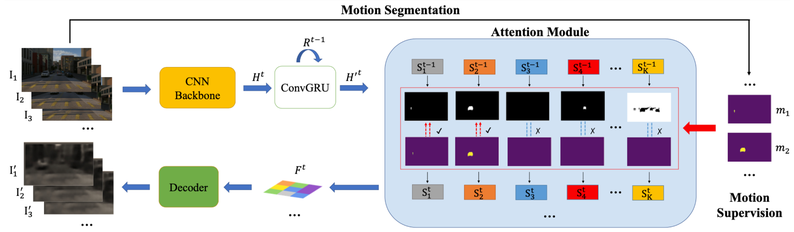
- Cannot resolve the object/background ambiguity in realistic scenes

Resolving object/background ambiguity

- Ambiguity of object definition is not resolvable for *static* images
- Videos provide a strong grouping cue -- independent object motion
- Focus on dynamic objects -- entities that *can* move independently

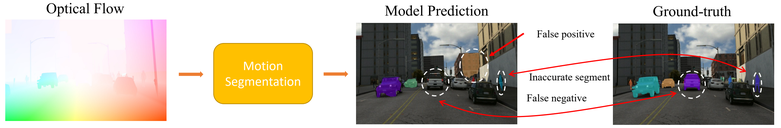


Our approach



- Conv-GRU based spatial-temporal feature extraction network
- Slot representation with learnable initial states
- Efficient one-shot slot decoding to save memory
- Motion cues to guide the attention masks

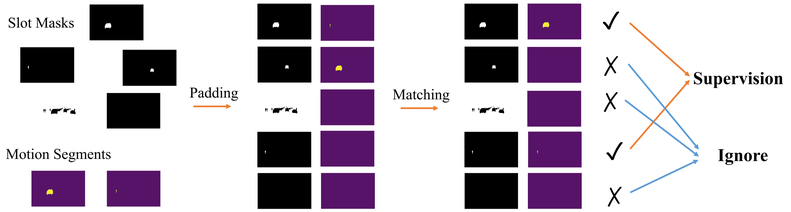
Motion segmentation



- Aim to segment objects that move independently from the camera using optical flow (e.g. Dave et al., ICCV 19)
- Noisy sparse motion predictions

Incorporating independent motion priors

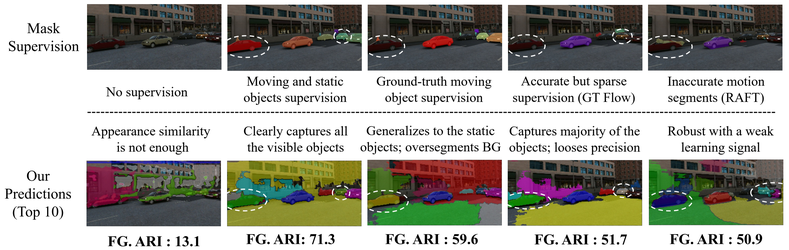
- Bipartite matching between the predicted masks and motion segments
- Only backpropagate positive signals for the matched masks



Benchmark

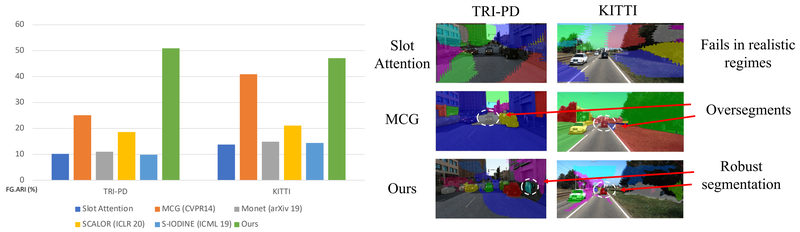
- High-quality realistic Parallel Domain (PD) dataset, released with code
- Evaluation metric: foreground ARI (FG. ARI) score

Ablation study



- Object discovery with different levels of supervisions
- Learn to discover the objects even with sparse and noisy motion segments
- The model generalizes to non-moving objects

Comparison to the state-of-the-art



- Appearance similarity is not sufficient in realistic regimes
- Surprisingly MCG outperforms most recent learning-based methods
- Motion cues are critical for resolving the object/background ambiguity