Object Discovery form Motion-Guided Tokens

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

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

Code available

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

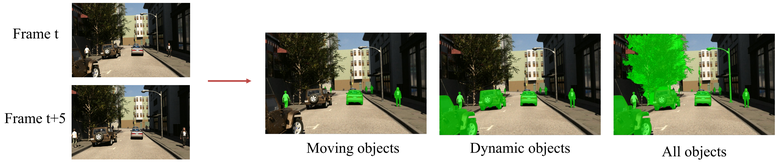
Object discovery

- Object discovery: separate objects from background without manual labels

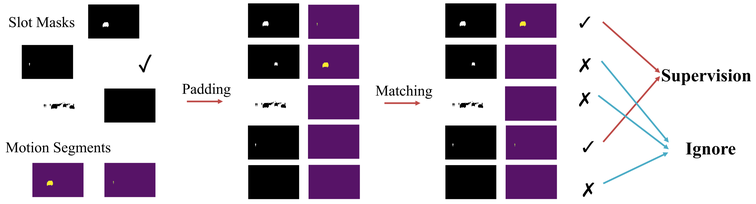
- Slot attention: a powerful object discovery pipeline

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

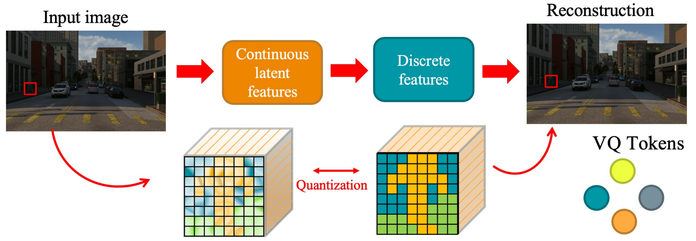


- Motion-guided object discovery [Bao et al., CVPR 22]

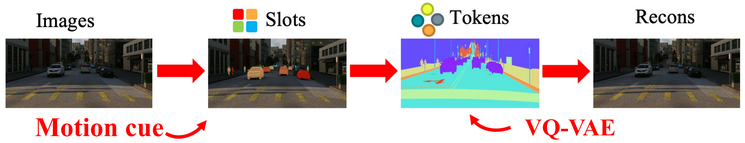


Background: tokenization via VQ-VAE

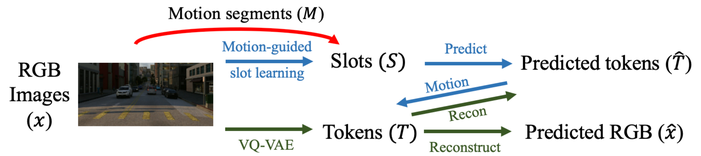
- Tokenization in the latent space [Oord et al., NeurIPS 17]



Our approach: motion-guided tokens



- Tokens serve as mid-level reconstruction space
- Bind instance-level slots to mid-level tokens

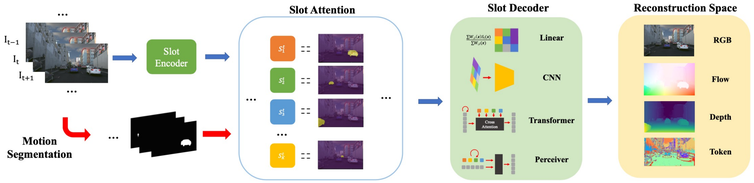


- Joint optimization: link the motion cue and tokens

$$\mathcal{L}_{joint} = \left|\left|\mathrm{sg}(T) - \hat{T}\right|\right|^2 + \left|\left|\mathrm{sg}(\hat{T}) - T\right|\right|^2$$

$\mathrm{sg}(\cdot)$: stop gradient

A unified object discovery system

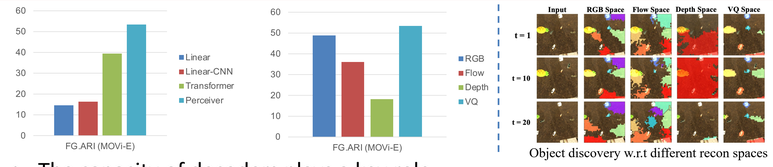


- Comprehensively evaluate the contributions of prior works
- Ablate key design choices in a unified framework

Benchmark

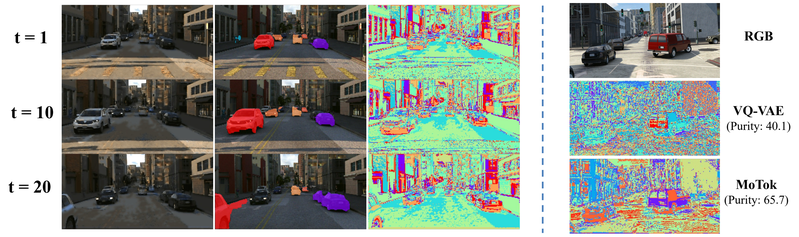
- Evaluation metric: video foreground adjusted random index (FG. ARI)

Effect of reconstruction spaces and decoders



- The capacity of decoders plays a key role
- VQ-space + perceiver decoder yields the best performance

Slots and tokens visualizations



- MoTok enables the emergence of interpretable object-specific mid-level features

Comparison to the state-of-the-art

- Motion cue is the key for real-world object discovery
- Vector quantization helps better utilize the motion guidance