Towards Robust Video Object Segmentation with Adaptive Object Calibration

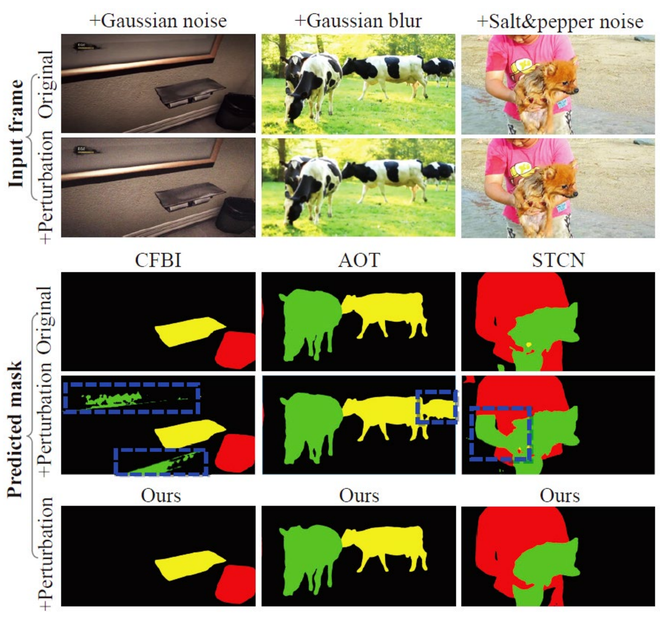
Xiaohao Xu, [1,2] Jinglu Wang, [2] Xiang Ming, [2] Yan Lu [2]
[1] Huazhong University of Science & Technology
[2] Microsoft Research Asia

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

paper repo

Introduction

- **Task**: **Video Object Segmentation (VOS)** aims at segmenting target objects in a video clip given the ground truth mask at the reference frame.
- **Motivation**: Our finding indicates that existing **VOS models are fragile to natural perturbations**.

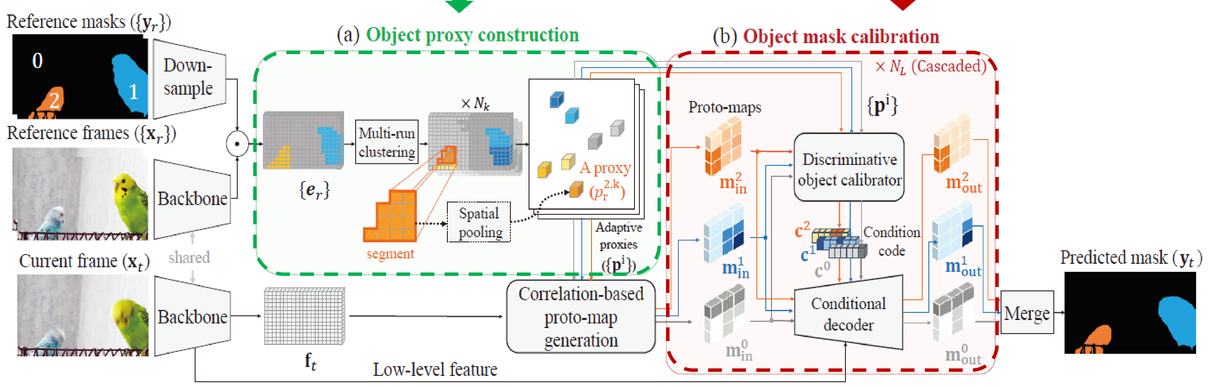


Our model with adaptive object calibration shows superior robustness against perturbations.

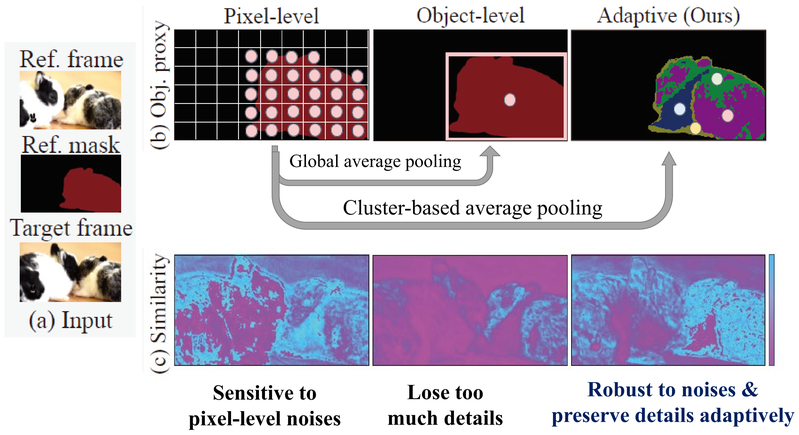
- **Solution**: The key insight is to:
  - introduce an **adaptive object proxy representation** for referenced objects robustly, which **reduces errors incurred by unstable pixel-level matching**.
  - **calibrate the object masks** by updating object representation and masks in an interleaving manner progressively, achieving **discrimination among co-existing objects**.

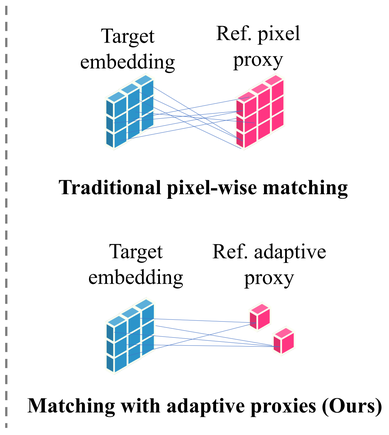
Method

Robust VOS Framework = Robust object representation + Robust mask decoding



Adaptive Object Proxy (AOP) for Robust Object Representation





Discriminative Object Calibration (DOC) for Robust Mask Decoding

incorporate **discrepancy** between different objects for inter-object **discrimination**

Results

Method	YouTube-VOS18	DAVIS17 Valid	DAVIS17 Test-dev
STM	79.4	81.8	72.3
CFBI	81.4	81.9	74.8
AOT-B	83.2	82.1	75.5
Ours-Base	**83.6**	**83.1**	**76.5**
STCN	83.0	**85.4**	76.1
AOT-L	83.7	83.8	78.3
Ours-MF	**84.0**	83.8	**79.3**

Comparisons on standard VOS benchmarks w.r.t overall performance (J&F).

Method	After-perturbation accuracy (↑)	Perturbation robustness (↓)
CFBI	79.4	1.6
AOT-B	81.6	1.7
Ours-Base	**82.3**	**1.4**
STCN	79.7	3.0
AOT-L	81.7	1.9
Ours-MF	**82.7**	**1.4**

Pilot study of perturbation robustness for VOS models on the perturbed dataset YouTube-VOS-P.

Qualitative results of our model.