ETH zürich VIS Visual Intelligence and Systems

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Mask-Free Video Instance Segmentation

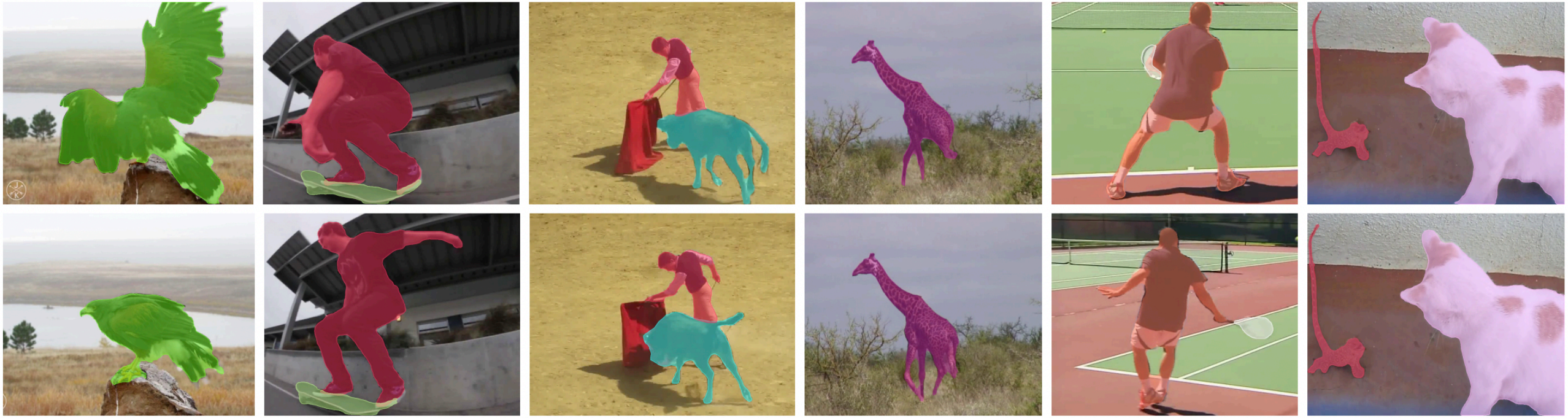
Lei Ke, Martin Danelljan, Henghui Ding, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu

Paper Tag: THU-PM-215

(github.com/SysCV/MaskFreeVIS)

Method

- We propose MaskFreeVIS, for high performance VIS without any mask annotations.
- We leverage the rich temporal consistency constraints by introducing the Temporal KNN-patch Loss.

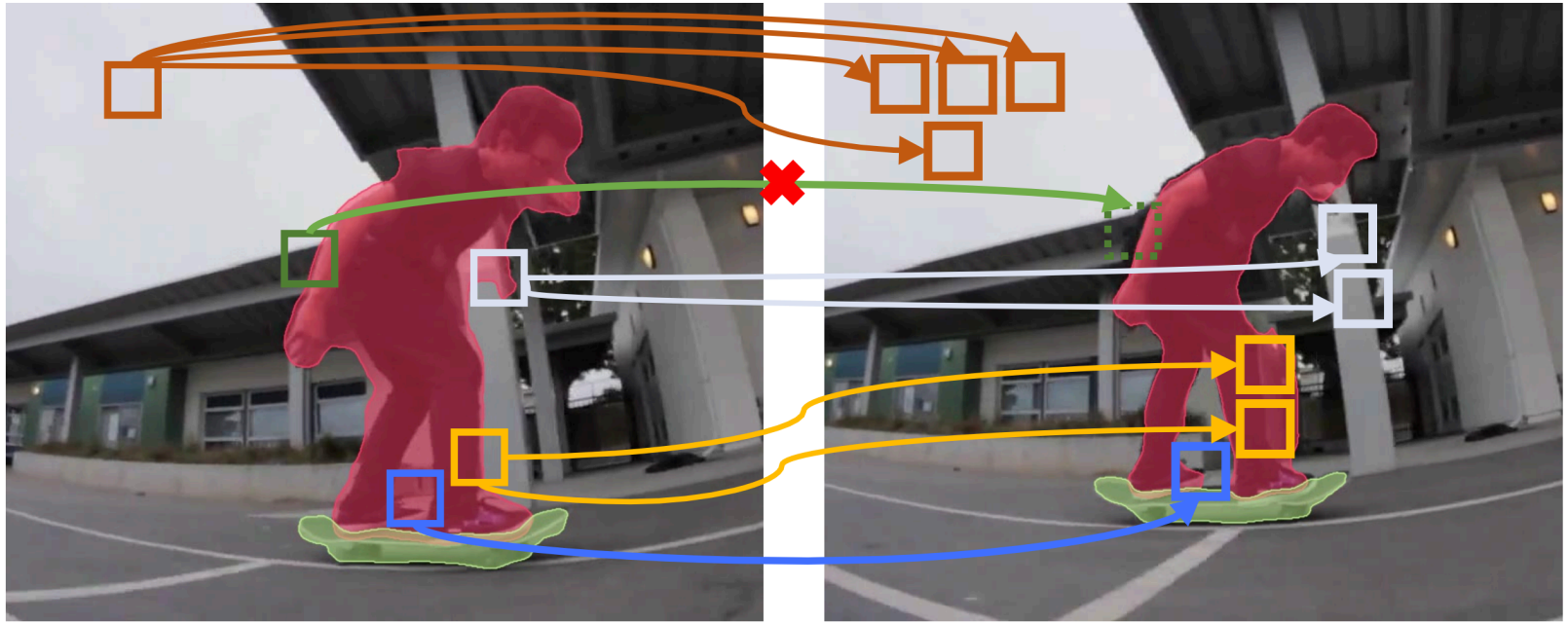


Mask prediction of our MaskFreeVIS trained **without** using any masks annotation.

Mask-Free VIS

Enforce mask consistency between matches.

Temporal One-to-**K** Patch Correspondence



VIS Mask Prediction at Frame T

VIS Mask Prediction at Frame T+1

Blue: Unique one-to-one match.

Orange: Multiple matches in homogeneous regions.

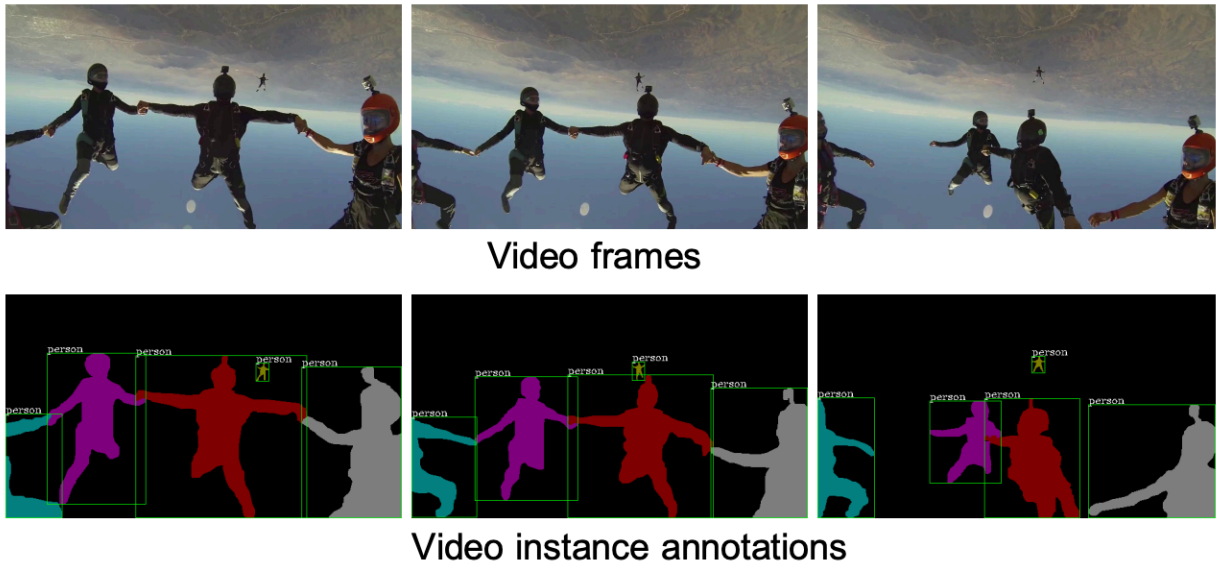
White and Yellow: multiple matches on image edges.

Green: No match due to occlusion.

Video Instance Segmentation

Annotation Requirement for VIS methods

- Bounding boxes, object ids and instance masks per frame for all objects.
- Building a large-scale VIS/MOTS benchmark is very expensive.



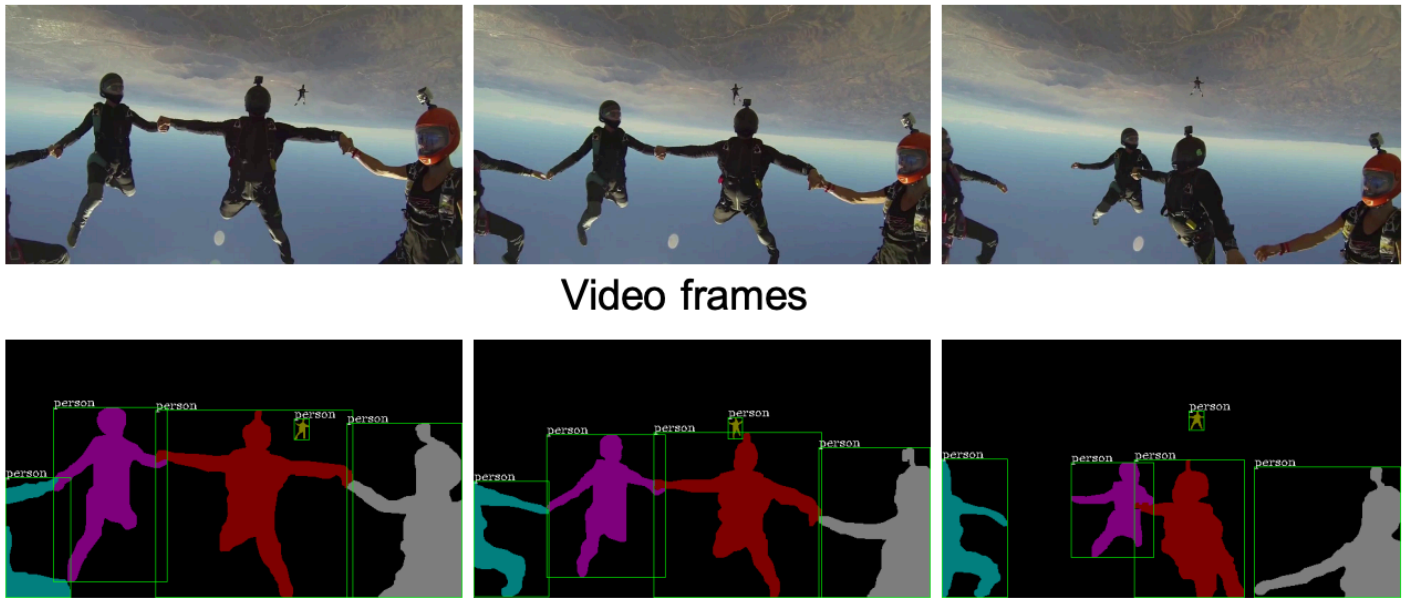
(a) Image source: Video Instance Segmentation. ICCV, 2019.

(b) Video source: OVIS

Video Instance Segmentation

Annotation Cost Analysis

- On COCO, it takes on average 79.2s per instance to create a coarse polygon-based object mask while box only takes 7s (**11 times faster**) [1]. Classification takes 0.9s.
- **Video boxes annotation is significantly faster/cheaper than video masks.**



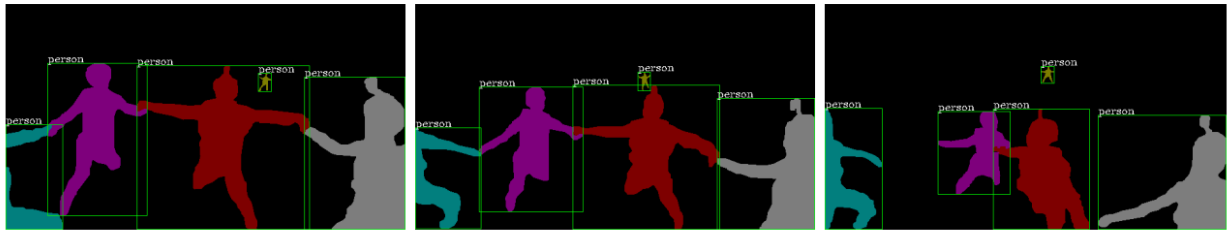
(a) Image source: Video Instance Segmentation. ICCV, 2019.

[1] Pointly-Supervised Instance Segmentation. CVPR, 2022.

Can we train VIS models **w/o** Mask Labeling?

- **Goal**: Training a VIS model **without** using any video or even image masks, but still achieving high-performing results.

Video frames



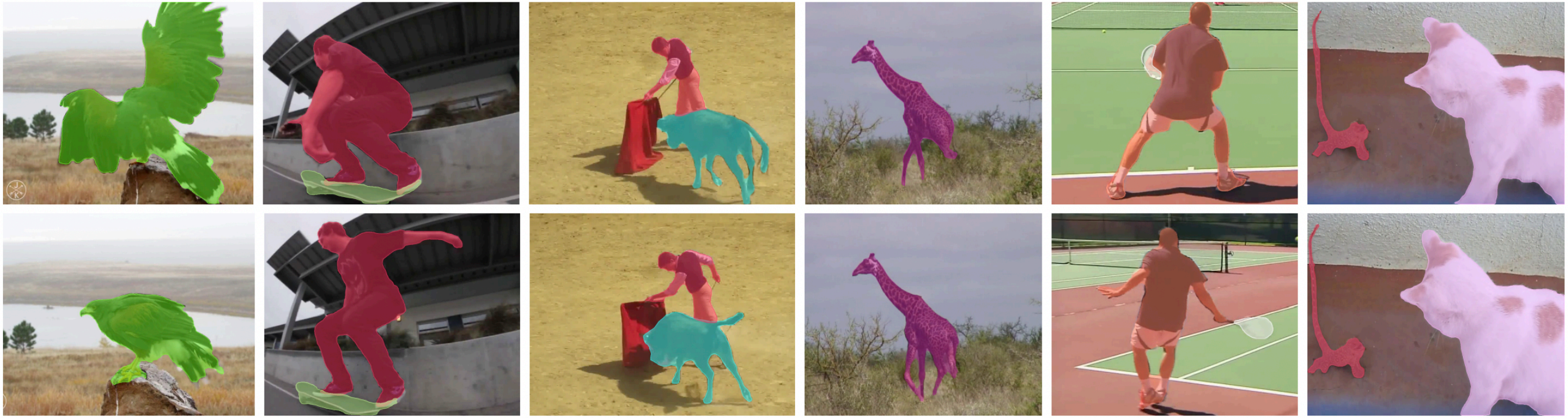
Video instance annotations

Only using the bounding boxes during training.

- **Insight**: Considering the object **bounding box** as the coarse object masks, and optimize them with temporal consistency.

Method

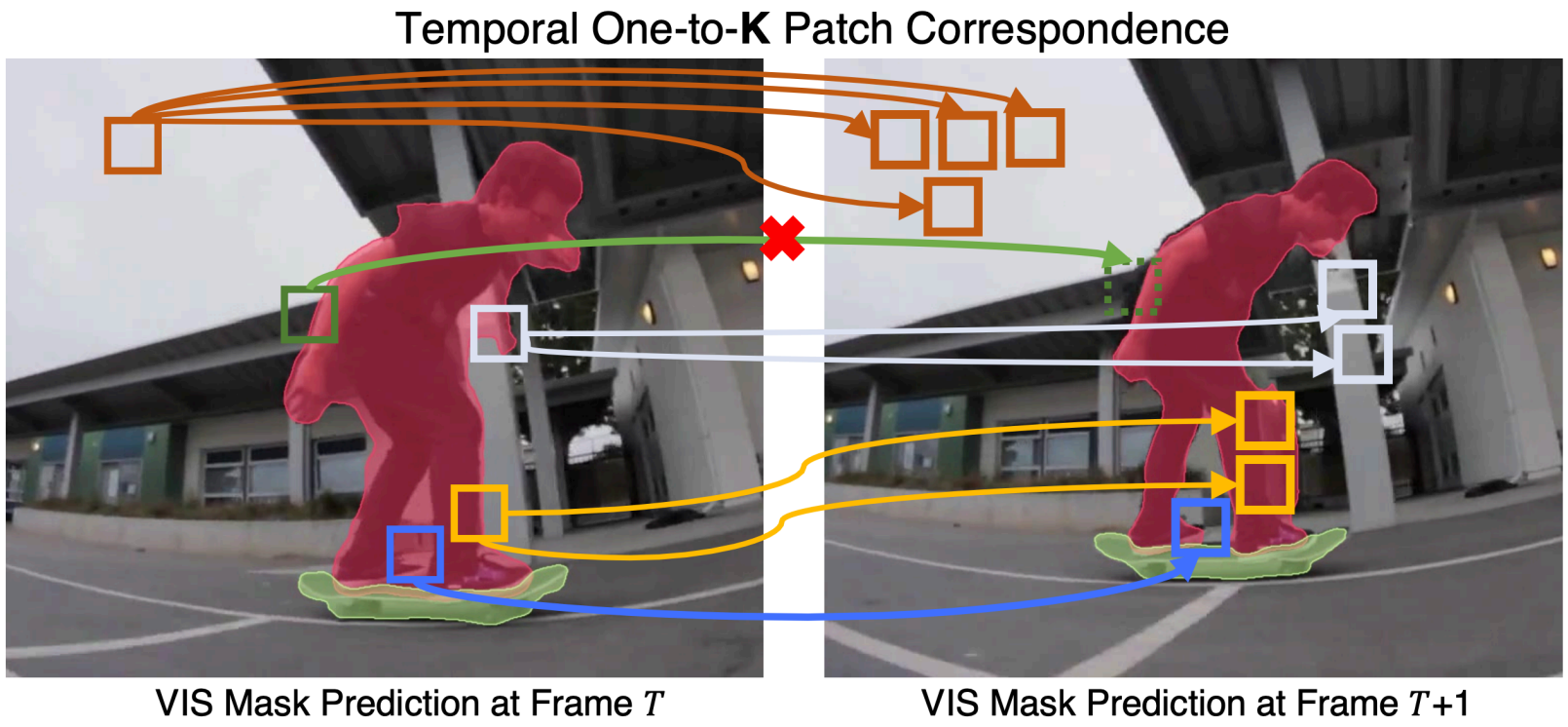
- We propose MaskFreeVIS, for high performance VIS without any mask annotations.
- We leverage the rich temporal consistency constraints by introducing the Temporal KNN-patch Loss.



Mask prediction of our MaskFreeVIS trained **without** using any masks annotation.

Mask-Free VIS

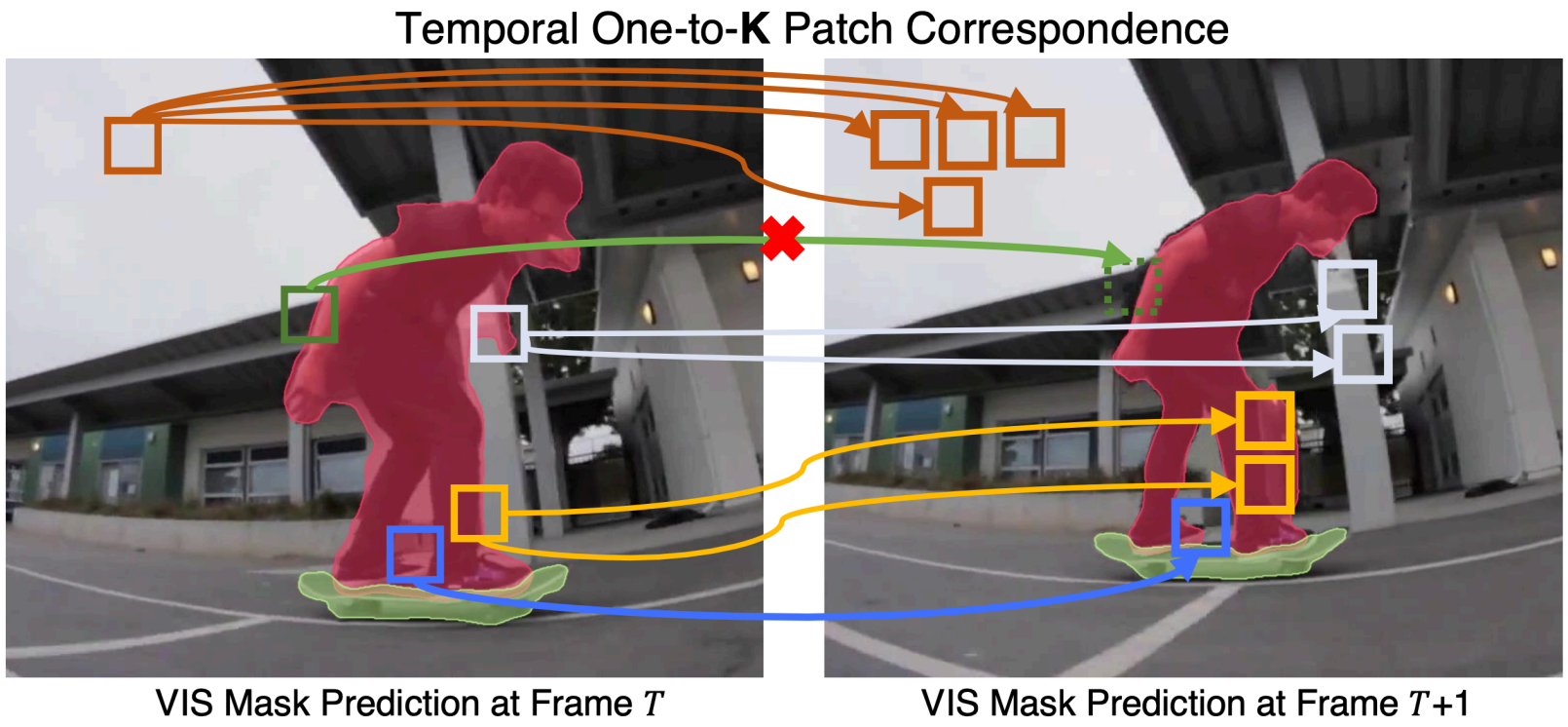
- The core component TK-Loss:



  - 1) Mask consistency between **one-to-K** patch correspondences.
  - 2) **No** trainable parameters.
  - 3) TK-Loss replaces the conventional video masks losses.

Mask-Free VIS

- The core component TK-Loss:



Blue: Unique one-to-one match.

Orange: Multiple matches in homogeneous regions.

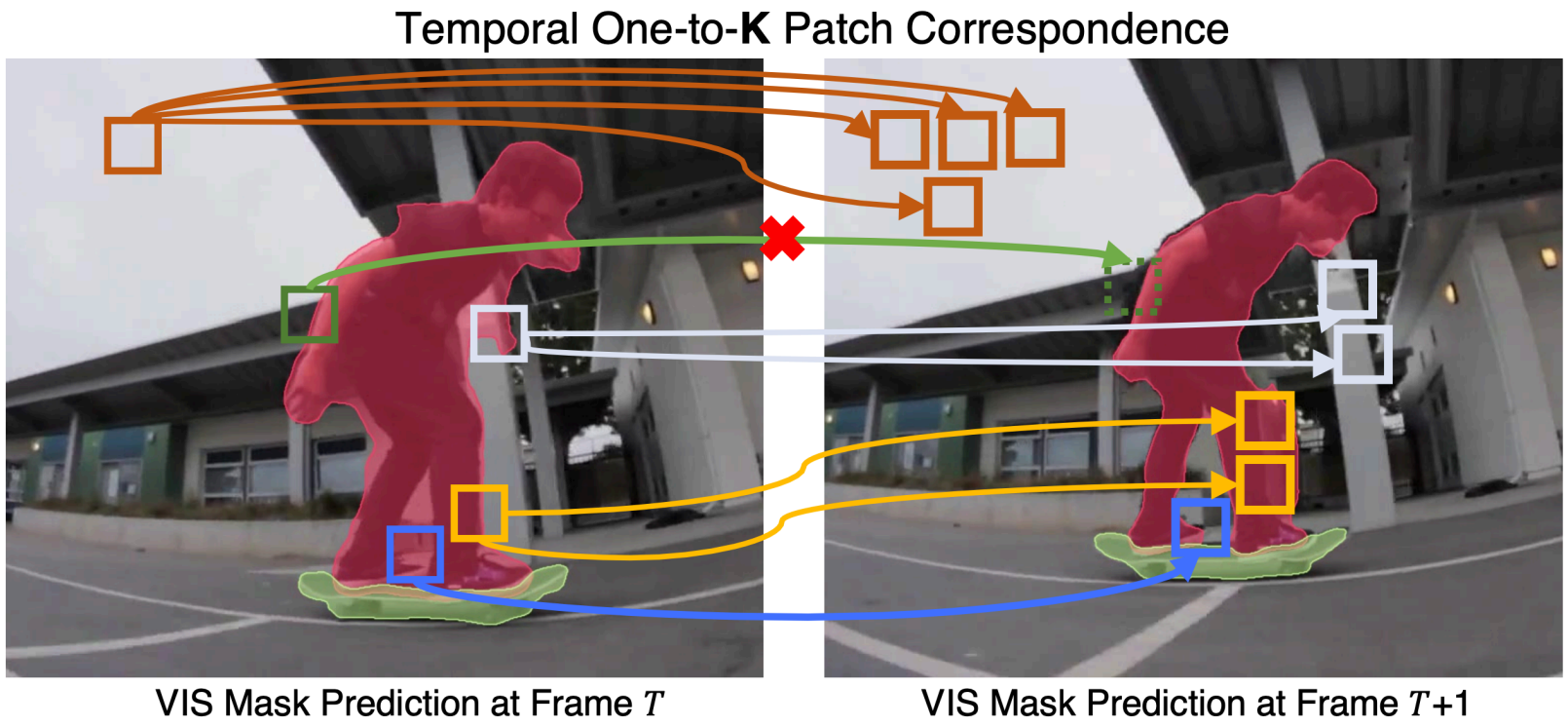
White and Yellow: multiple matches on image edges.

Green: No match due to occlusion.

- 1) Mask consistency between **one-to-K** patch correspondences.
- 2) **No** trainable parameters.
- 3) TK-Loss replaces the conventional video masks losses.

Mask-Free VIS

- The core component TK-Loss:



Blue: Unique one-to-one match.

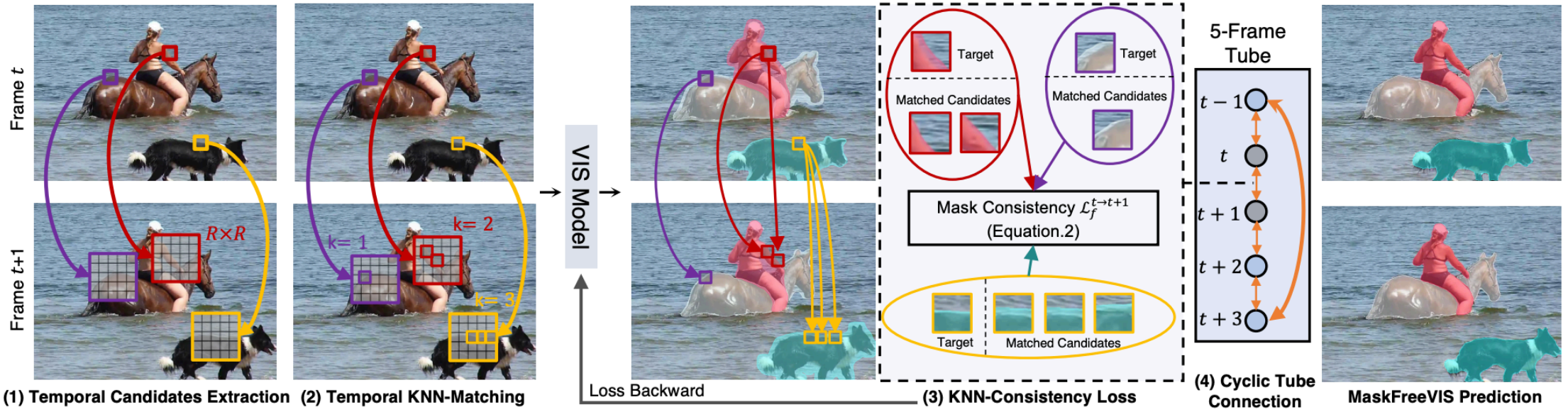
Orange: Multiple matches in homogeneous regions.

White and Yellow: multiple matches on image edges.

Green: No match due to occlusion.

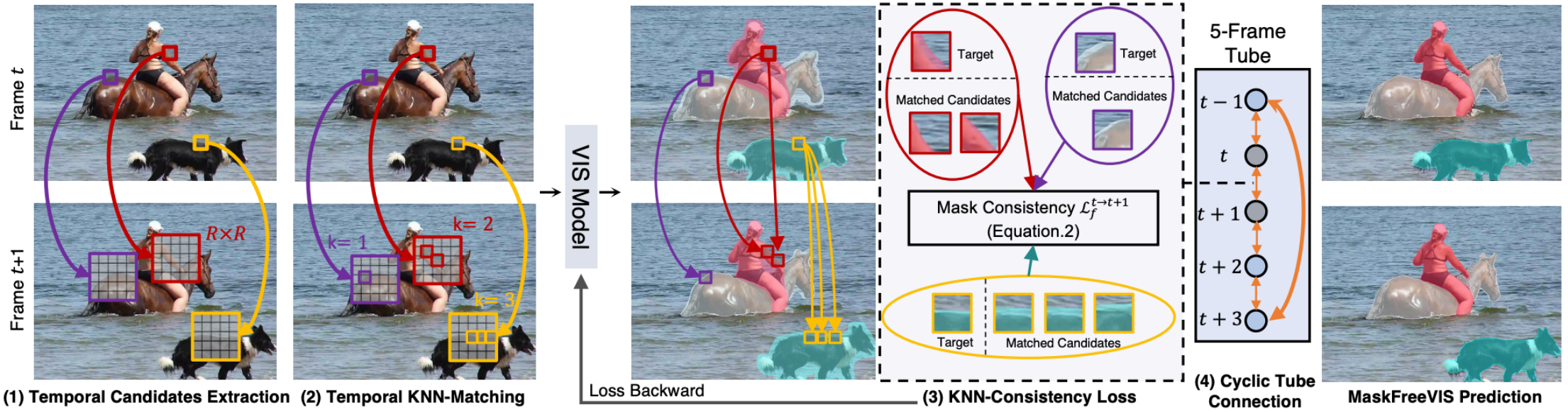
- 1) Mask consistency between **one-to-K** patch correspondences.
- 2) **No** trainable parameters.
- 3) TK-Loss replaces the conventional video masks losses.

Mask-Free VIS



- TK-Loss has four steps:
  1) **Patch Candidate Extraction**: Patch candidates searching across frames with radius **R**.
  2) **Temporal KNN-Matching**: Match **K** high-confidence candidates by patch affinities.
  3) **KNN-Consistency Loss**: Enforce mask consistency objective among the matches.
  4) **Cyclic Tube Connection**: Temporal loss aggregation in the 5-frame tube.

Mask-Free VIS



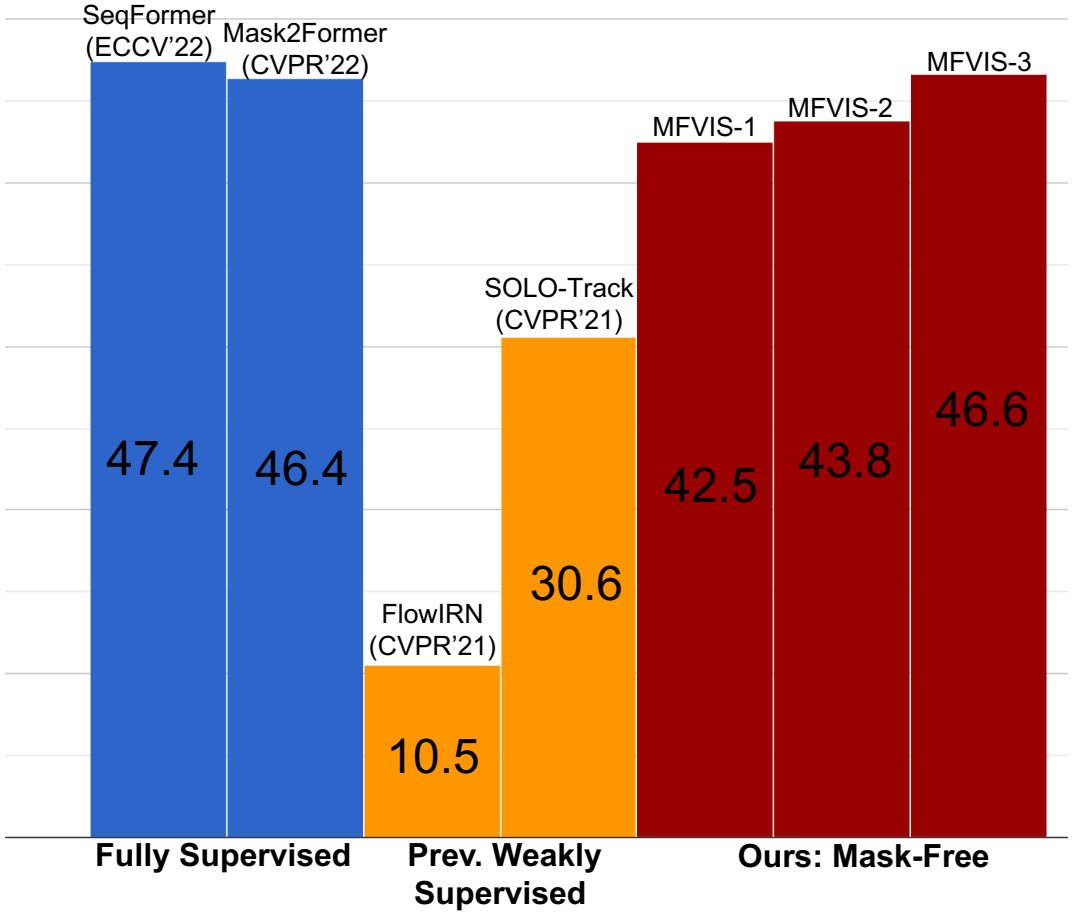
- TK-Loss has four steps:
  1) **Patch Candidate Extraction**: Patch candidates searching across frames with radius **R**.
  2) **Temporal KNN-Matching**: Match **K** high-confidence candidates by patch affinities.
  3) **KNN-Consistency Loss**: Enforce mask consistency objective among the matches.
  4) **Cyclic Tube Connection**: Temporal loss aggregation in the 5-frame tube.

Quantitative VIS Results

Achieving **91.5**% (42.5 vs. 46.4) of its fully-supervised baseline performance on YTVIS.

Method	Video Mask	Image Mask	Pseudo Video	AP
SeqFormer (ECCV'22)	✓	✓	✓	47.4
VMT (ECCV'22)	✓	✓	✓	47.9
Mask2Former (CVPR'22)	✓	✓	✓	47.8
MaskFreeVIS (ours)	✗	✓	✓	46.6
Mask2Former (CVPR'22)	✓	✓	✗	46.4
MaskFreeVIS (ours)	✗	✗	✗	42.5

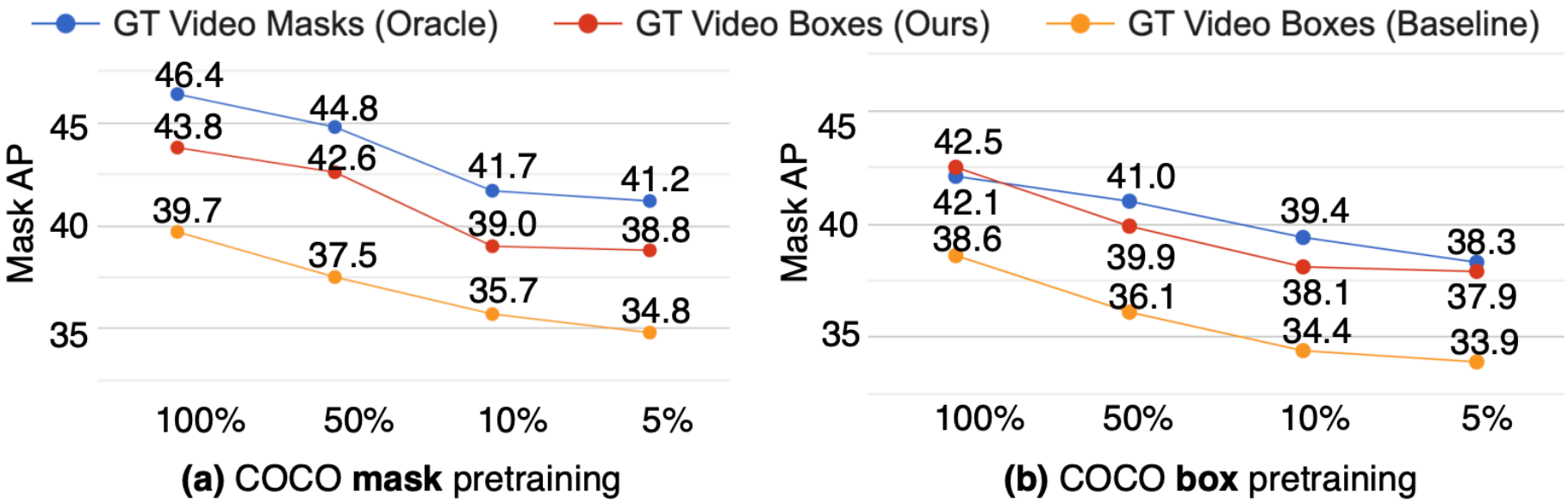
Comparison on YTVIS benchmark



MFVIS-1: **w/o** any mask usage.
MFVIS-2: using pre-trained COCO mask model as initialization.
MFVIS-3: using COCO masks as pseudo video masks labels.

Quantitative VIS Results

Results on with various percentages of the YTVIS training data.



Uniformly sample frames and their labels for each video.

Pseudocode

- TK-Loss has four steps:
  1) **Patch Candidate Extraction.**
  2) **Temporal KNN-Matching.**
  3) **KNN-Consistency Loss.**
  4) **Cyclic Tube Connection.**

Algorithm 1 Temporal KNN-patch Loss.

Input: Tube length T, mask predictions M, frame width W, height H, radius R, patch distance threshold D.

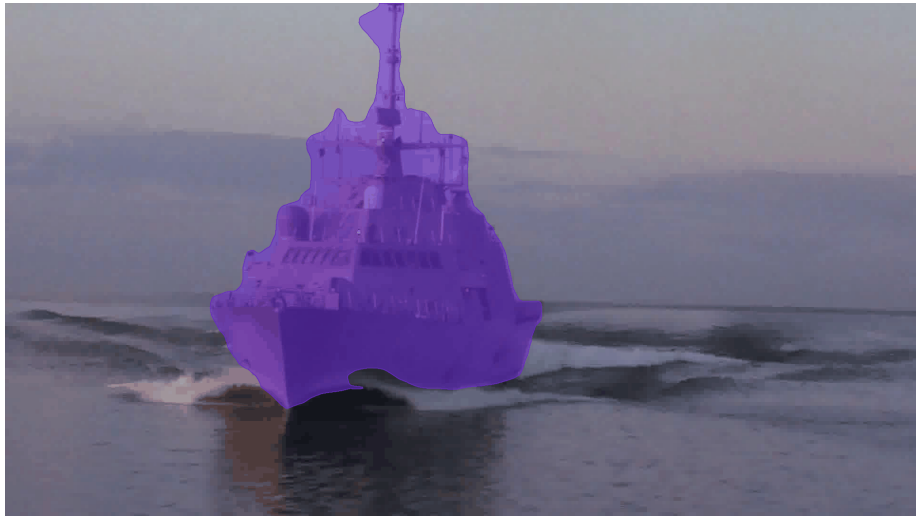
Output: TK-Loss $\mathcal{L}_{\text{temp}}$

$\texttt{top}K$ denotes selecting top K patch candidates with the maximum patch similarities computed using L_2 distance $\texttt{Dis}(\cdot,\cdot)$
$L_{\texttt{cons}}$ denotes mask consistency loss (Equation 3 of the paper)
$\mathcal{L}_{\text{temp}} \leftarrow 0.$
for $t = 1, \ldots, T$ **do**
  $\hat{t} \leftarrow (t+1)\%T$
  $\mathcal{L}_f^{t\to\hat{t}} \leftarrow 0.$
  for $j = 1, \ldots, H \times W$ **do**
    # 1) Patch Candidate Extraction:
    $\mathcal{S}_{p_j}^{t\to\hat{t}} \leftarrow \{\hat{p}_i\}_i$, where $\|p_j - \hat{p}_i\| \leq R$
    # 2) Temporal KNN-Matching:
    $\mathcal{S}_{p_j}^{t\to\hat{t}} \leftarrow \texttt{top}K(\mathcal{S}_{p_j}^{t\to\hat{t}})$, where $\texttt{Dis}(p_j, \hat{p}_i) \leq D$
    # 3) Consistency Loss
    $\mathcal{L}_f^{t\to\hat{t}} \leftarrow \mathcal{L}_f^{t\to\hat{t}} + \sum_{\hat{p}_i \in \mathcal{S}_{p_j}^{t\to\hat{t}}} L_{\texttt{cons}}(M_{p_j}^t, M_{\hat{p}_i}^{\hat{t}})$
  end for
  # 4) Cyclic Connection
  $\mathcal{L}_{\text{temp}} \leftarrow \mathcal{L}_{\text{temp}} + \mathcal{L}_f^{t\to\hat{t}}/(H \times W)$
end for

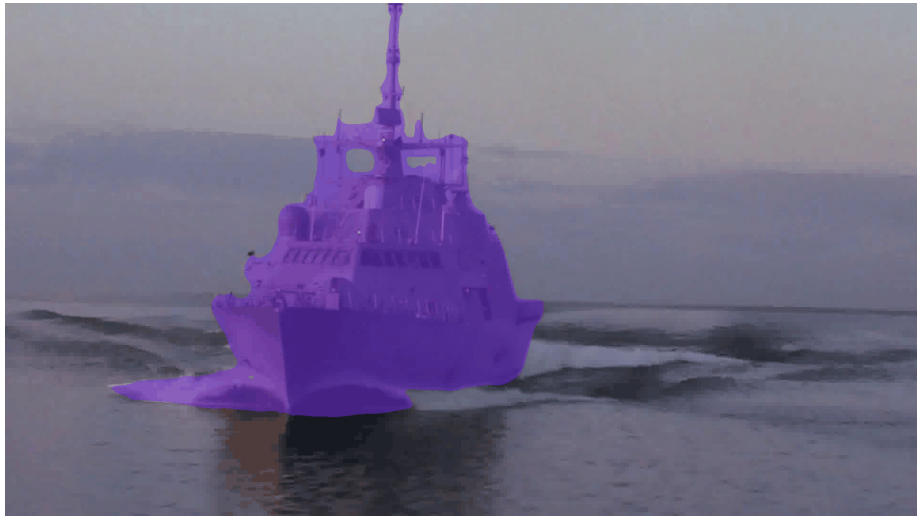
Qualitative VIS Results Comparison

This section includes qualitative comparisons between our MaskFreeVIS, Baseline and Oracle Mask2Former.

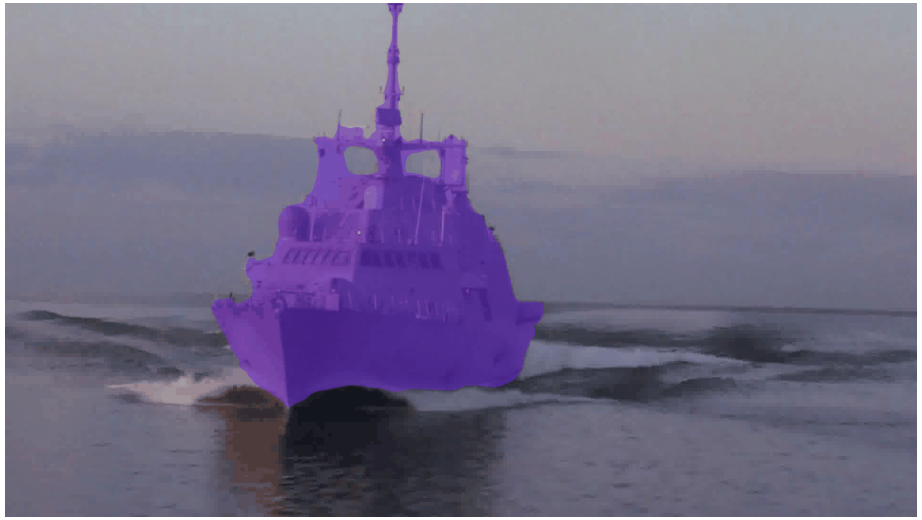
Results on YouTube-VIS **w/o** video masks usage



Baseline (Mask2Former + BoxInst)

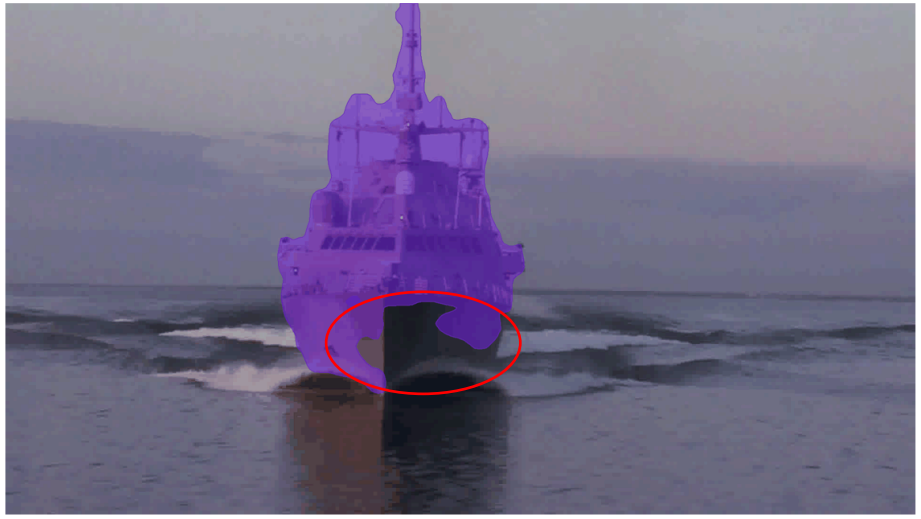


Oracle (Mask2Former + GT Video Masks Training)

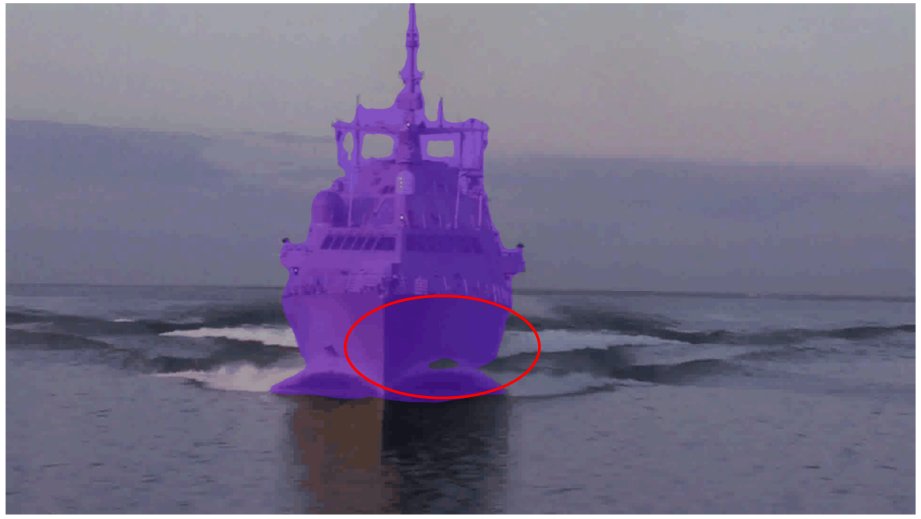


Ours (MaskFreeVIS)

Results on YouTube-VIS **w/o** video masks usage



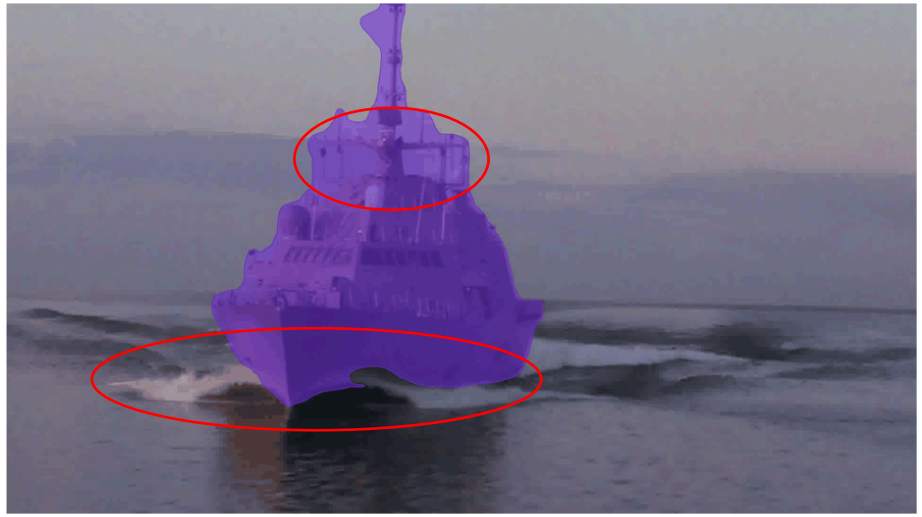
Baseline (Mask2Former + BoxInst)



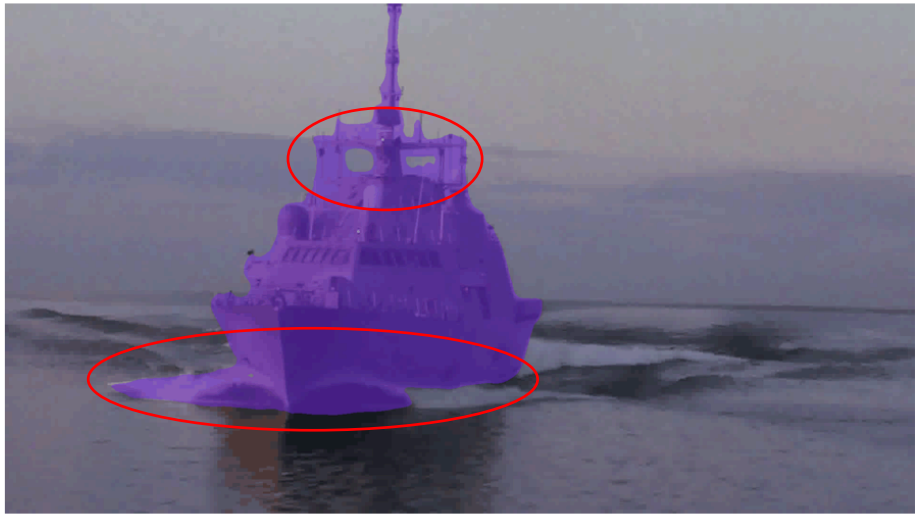
Oracle (Mask2Former + GT Video Masks Training)

Ours (MaskFreeVIS)

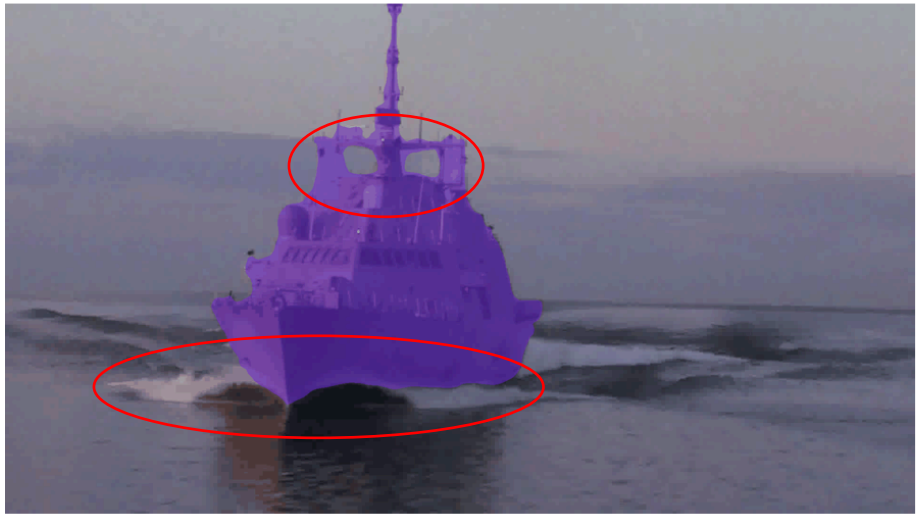
Results on YouTube-VIS **w/o** video masks usage



Baseline (Mask2Former + BoxInst)

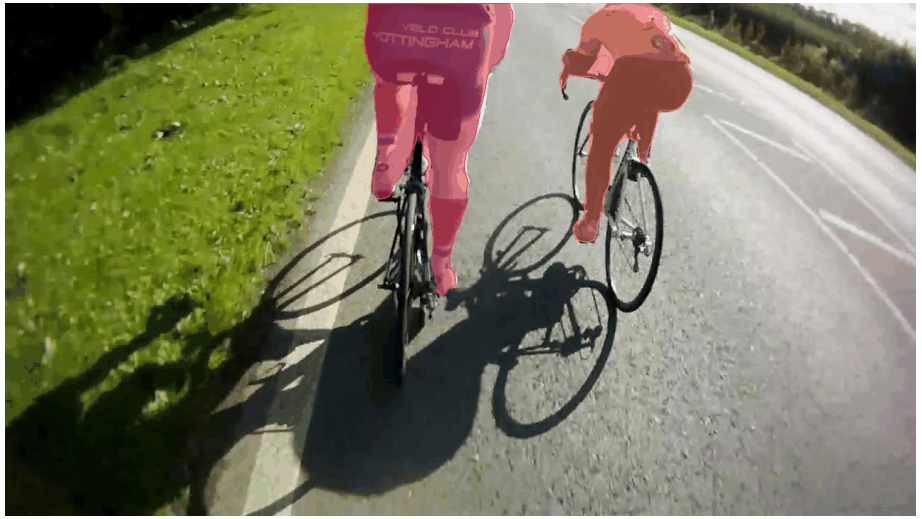


Oracle (Mask2Former + GT Video Masks Training)

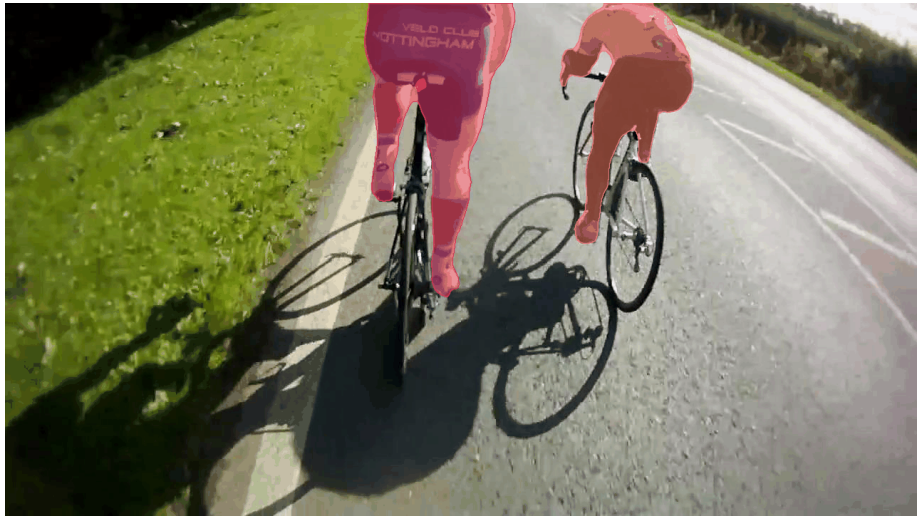


Ours (MaskFreeVIS)

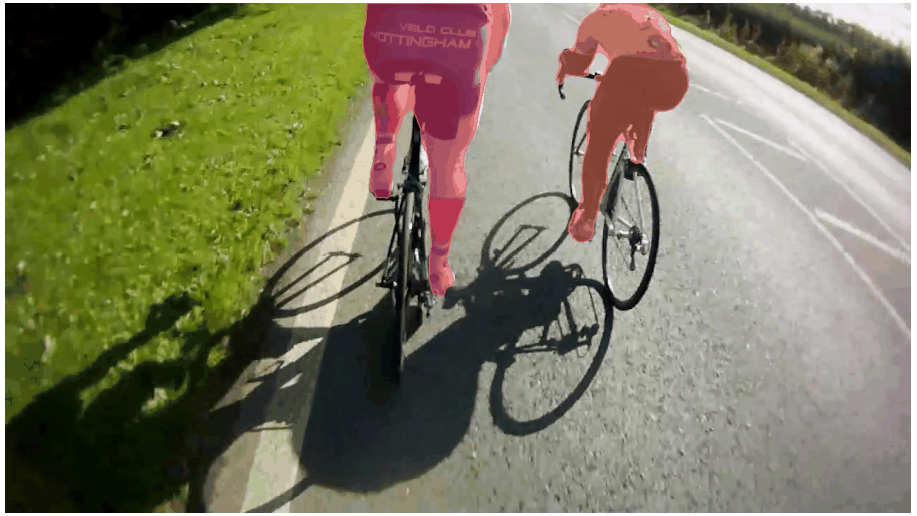
Results on YouTube-VIS **w/o** video masks usage



Baseline (Mask2Former + BoxInst)

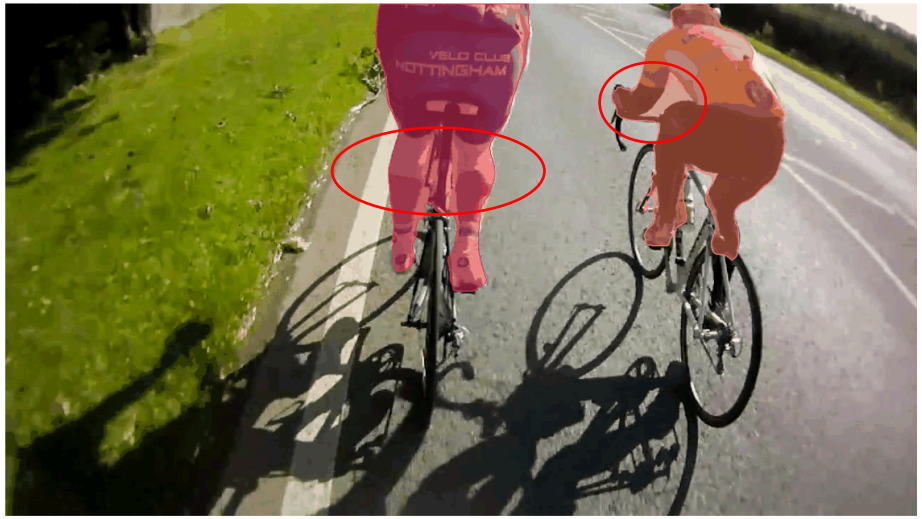


Oracle (Mask2Former + GT Video Masks Training)



Ours (MaskFreeVIS)

Results on YouTube-VIS **w/o** video masks usage

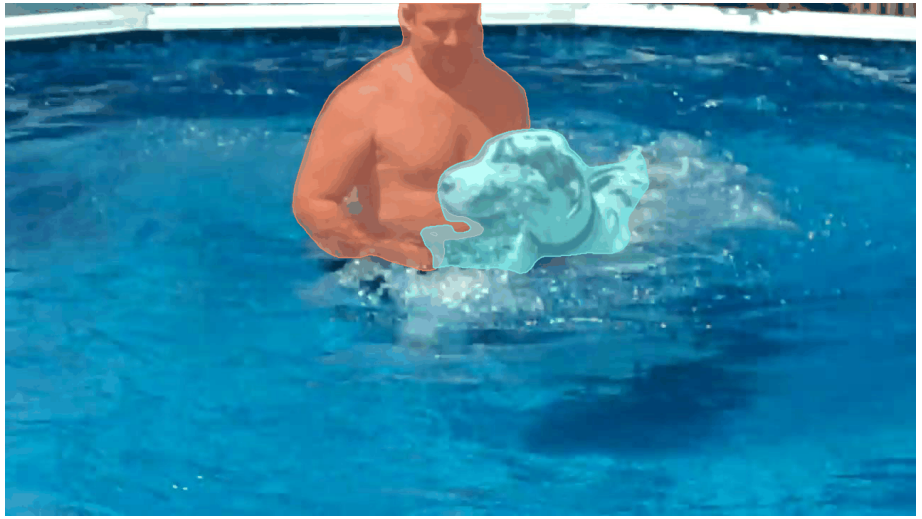


Baseline (Mask2Former + BoxInst)

Oracle (Mask2Former + GT Video Masks Training)

Ours (MaskFreeVIS)

Results on YouTube-VIS **w/o** video masks usage

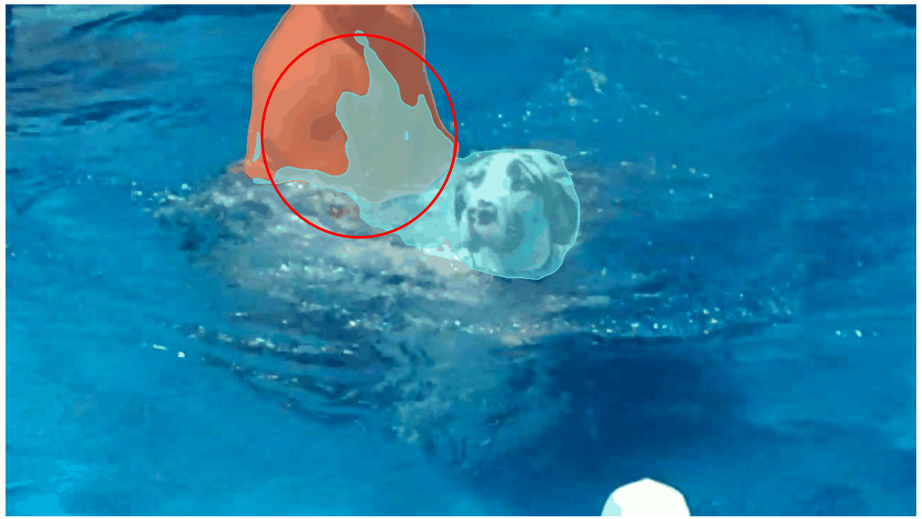


Baseline (Mask2Former + BoxInst)

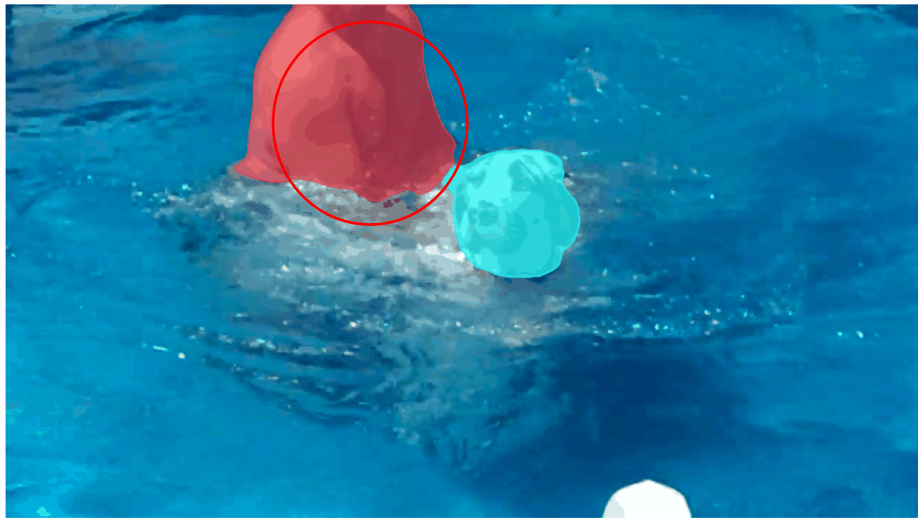
Oracle (Mask2Former + GT Video Masks Training)

Ours (MaskFreeVIS)

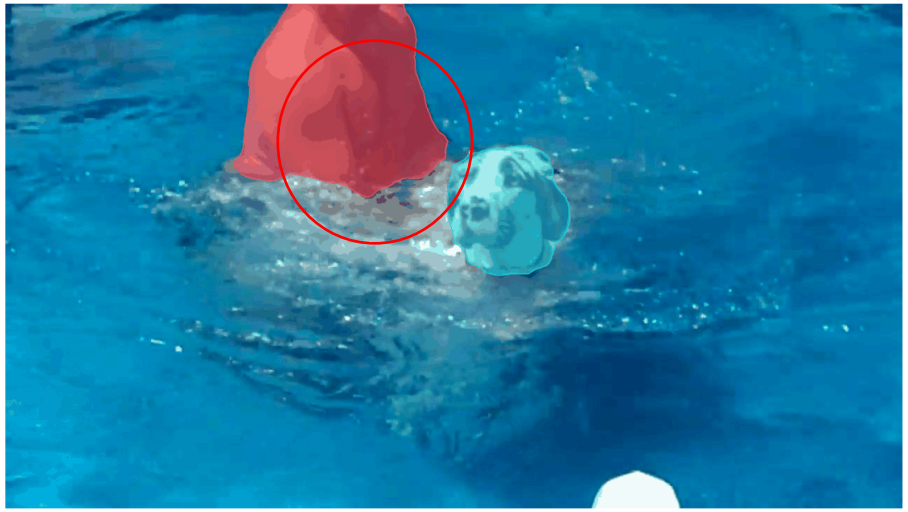
Results on YouTube-VIS **w/o** video masks usage



Baseline (Mask2Former + BoxInst)



Oracle (Mask2Former + GT Video Masks Training)



Ours (MaskFreeVIS)

OVIS Results of MaskFreeVIS

Results of MaskFreeVIS on the Web Videos

Web Video Result of MaskFreeVIS

Summarization

- **Contribution**:
  - The first competitive VIS method that does not need any masks during training.
  - Greatly reducing the gap between fully-supervised and weakly-supervised VIS.
- **Take away**:
  - High-performing VIS can be learned even without any mask annotations.

(github.com/SysCV/MaskFreeVIS)