北京交通大學
BEIJING JIAOTONG UNIVERSITY

MePro 数字媒体信息处理研究中心
Center of Digital Media Information Processing

Meeting of Paper Sharing

Video Instance Segmentation

Qi Tang

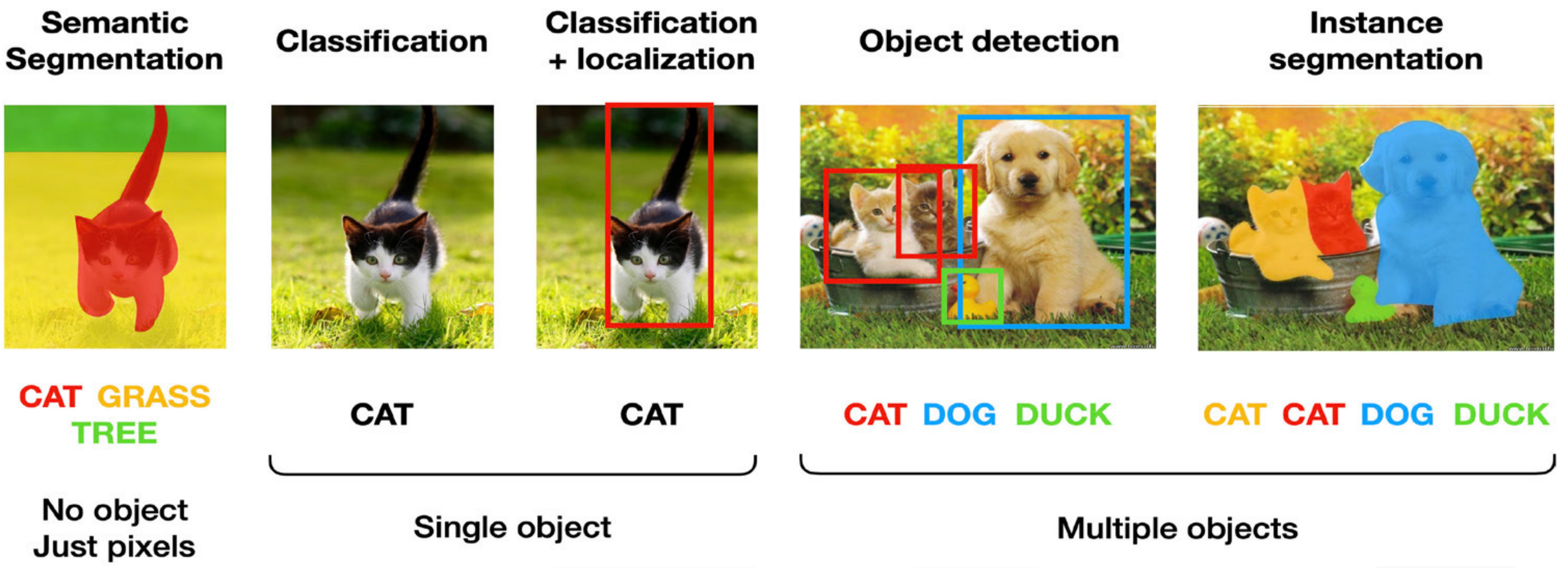
2022/11/25

Video Instance Segmentation

Linjie Yang* Yuchen Fan Ning Xu
ByteDance AI Lab UIUC Adobe Research

Video Instance Segmentation

Computer Vision Tasks



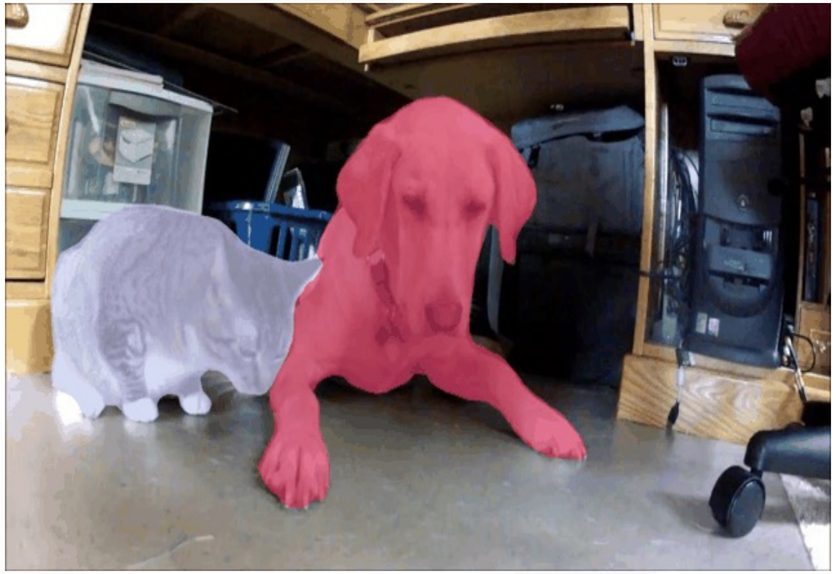
Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Motivation

Image

Video

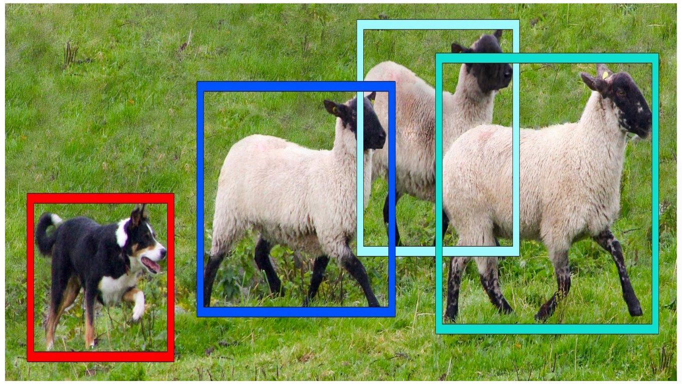


Extend the instance segmentation problem in the image domain to the video domain

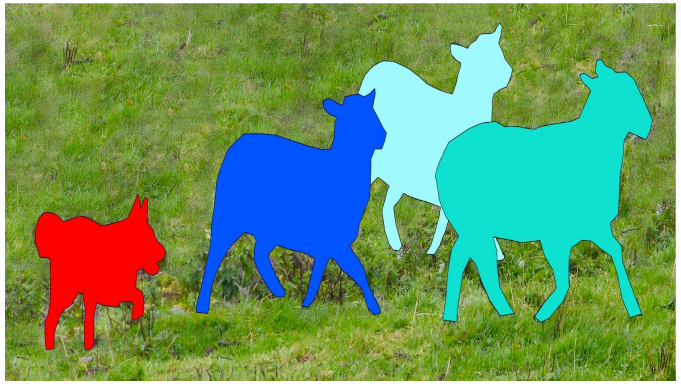
Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Problem Definition



+



+

aims at **simultaneous detection, segmentation and tracking of object instances in videos**

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Problem Definition

In video instance segmentation, we have a predefined category label set

$$\mathcal{C} = \{1, ..., K\}$$

where K is the number of categories.

Given a video sequence with T frames, suppose there are N objects belonging to the category set $\mathcal{C}$ in the video. For each object i, let $c^i \in \mathcal{C}$ denote its category label, and let $\mathbf{m}^i_{p...q}$ denote its binary segmentation masks across the video where $p \in [1, T]$ and $q \in [p, T]$ denote its starting and ending time.

Suppose a video instance segmentation algorithm produces H instance hypotheses. For each hypothese j, it needs to have a predicted category label $\tilde{c}^j \in \mathcal{C}$, a confidence score $s^j \in [0, 1]$ and a sequence of predicted binary masks $\tilde{\mathbf{m}}^j_{\tilde{p}...\tilde{q}}$. The confidence score is used for our evaluation metrics which will be explained shortly.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Problem Definition

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Related Work

- **Image Instance Segmentation**
- Video Object Tracking
- Video Object Detection
- Video Semantic Segmentation
- Video Object Segmentation

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Related Work

- Image Instance Segmentation
- **Video Object Tracking**
- Video Object Detection
- Video Semantic Segmentation
- Video Object Segmentation

One is the detection-based tracking which simultaneously detect and track video objects. Methods under this setting usually take the “tracking-by-detection” strategy. The other setting is the detection-free tracking, which targets at tracking objects given their initial bounding boxes in the first frame.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Related Work

- Image Instance Segmentation
- Video Object Tracking
- **Video Object Detection**
- Video Semantic Segmentation
- Video Object Segmentation

Motion Blur

The evaluation metric is limited to per-frame detection and does not require joint object detection and tracking.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

BEIJING JIAOTONG UNIVERSITY

Related Work

- Image Instance Segmentation
- Video Object Tracking
- Video Object Detection
- **Video Semantic Segmentation**
- Video Object Segmentation

Direct extension of semantic segmentation to videos, where image pixels are predicted as different semantic classes. Temporal information such as optical flow is adopted to improve either accuracy or efficiency of semantic segmentation models. Video semantic segmentation does not require explicit matching of object instances across frames.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Related Work

- Image Instance Segmentation
- Video Object Tracking
- Video Object Detection
- Video Semantic Segmentation
- **Video Object Segmentation**

Semi-supervised video object segmentation targets at tracking and segment a given object with a mask. In unsupervised scenario, a single foreground object is segmented. In both settings, algorithms consider the target objects as general objects and does not care about the semantic categories.

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

与 VIS 相关的任务	定义	区别	检测	分割	跟踪
Image Instance Segmentation	将像素分组为不同的语义类，还将它们分组为不同的对象实例。 通常采用两阶段模式，首先使用区域建议网络 RPN 生成对象建议，然后使用聚集的 ROI 特征预测对象的边界框和 masks	图像级处理 视频实例分割需在每一帧中分割对象实例，还需确定跨帧对象的对应关系	✓	✓	
Video Object Tracking	DBT(Detection by Tracking): 同时进行检测和跟踪 DFT(Detection-Free Tracking): 在第一帧给定初始边界框，无需检测器进行追踪	只进行检测，不进行分割	✓		✓
Video Object Detection	检测视频中的对象，目标身份信息用来提升检测算法的鲁棒性，但评估指标仅限于每帧检测	没有分割和追踪	✓		
Video Semantic Segmentation	在每一帧进行语义分割，采用光流等时间信息来提高语义分割模型的准确性或效率	不需要跨帧显式匹配对象实例		✓	
Video Object Segmentation	半监督：使用一个 mask 跟踪和分割一个给定对象，提取视觉相似性，运动线索和时间一致性，以识别视频中的同一对象。 无监督：不需要给第一帧 mask，不需要区分实例，只需要分割出单个目标即可	没有考虑语义或实例信息	✓	✓	✓

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

YouTube-VIS

Table 1: High level statistics of YouTubeVIS and previous video object segmentation datasets. YTO, YTVOS, and YTVIS stands for YouTubeObjects, YouTubeVOS, and YouTube-VIS respectively.

	YTO	FBMS	DAVIS		YTVOS	YTVIS
Videos	96	59	50	90	4,453	2,883
Categories	10	16	-	-	94	40
Objects	96	139	50	205	7,755	4,883
Masks	1.7k	1.5k	3.4k	13.5k	197k	131k
Exhaustive	✗	✗	✗	✗	✗	✓

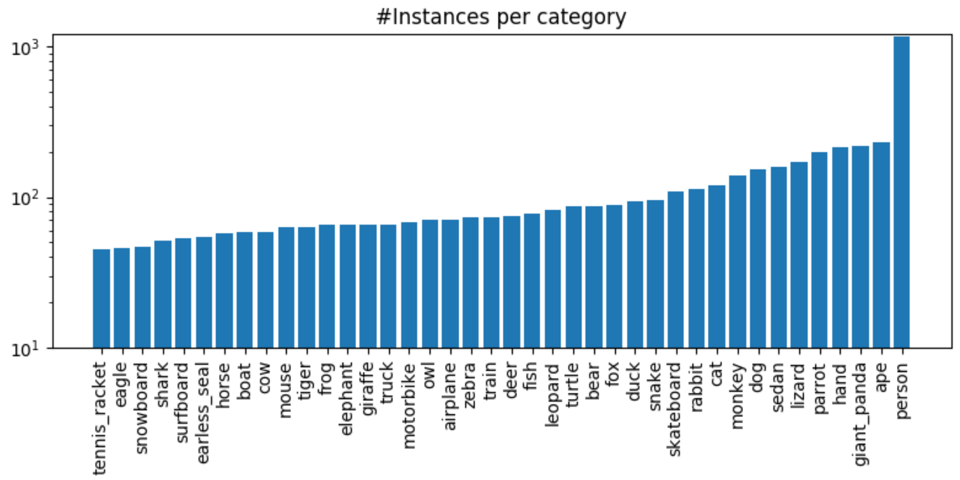
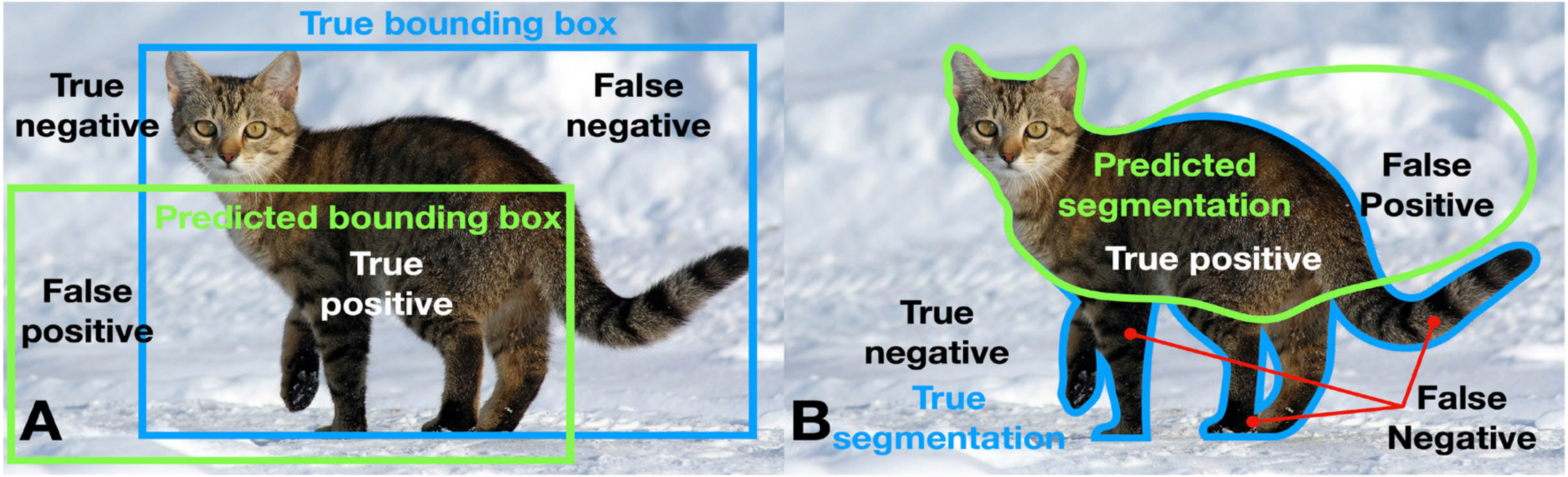


Figure 1: Number of unique video objects for the 40 categories in our dataset.

- **The 2019 version**
  - 2,883 high-resolution YouTube videos, 2,238 training videos, 302 validation videos and 343 test videos
  - A category label set including 40 common objects such as person, animals and vehicles
  - 4,883 unique video instances
  - 131k high-quality manual annotations
- **The 2021 version**
  - 3,859 high-resolution YouTube videos, 2,985 training videos, 421 validation videos and 453 test videos.
  - An improved 40-category label set by merging eagle and owl into bird, ape into monkey, deleting hands, and adding flying disc, squirrel and whale
  - 8,171 unique video instances
  - 232k high-quality manual annotations
- **The 2022 version**
  - 71 additional long videos in validation and 50 additional long videos in test set, with additional separate evaluation
  - 259 additional unique video instances, 9304 high-quality manual annotations

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Evaluation Metrics



- **Average Precision (AP)** is defined as the area under the precision-recall (PR) curve. AP is averaged over multiple intersection-over-union (IoU) thresholds. We follow the COCO evaluation metrics to use 10 IoU thresholds from 50% to 95% at step 5%.
- **Average Recall (AR)** is defined as the maximum recall given some fixed number of segmented instances per video.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$IoU^{st} = \frac{\Sigma_k S(m^k \cap \tilde{m}^k)}{\Sigma_k S(m^k \cup \tilde{m}^k)}$$

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Mask R-CNN

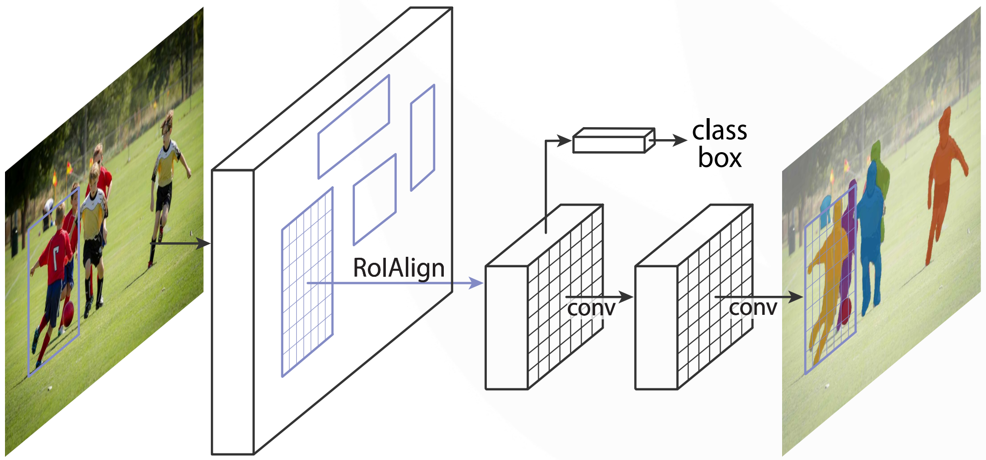


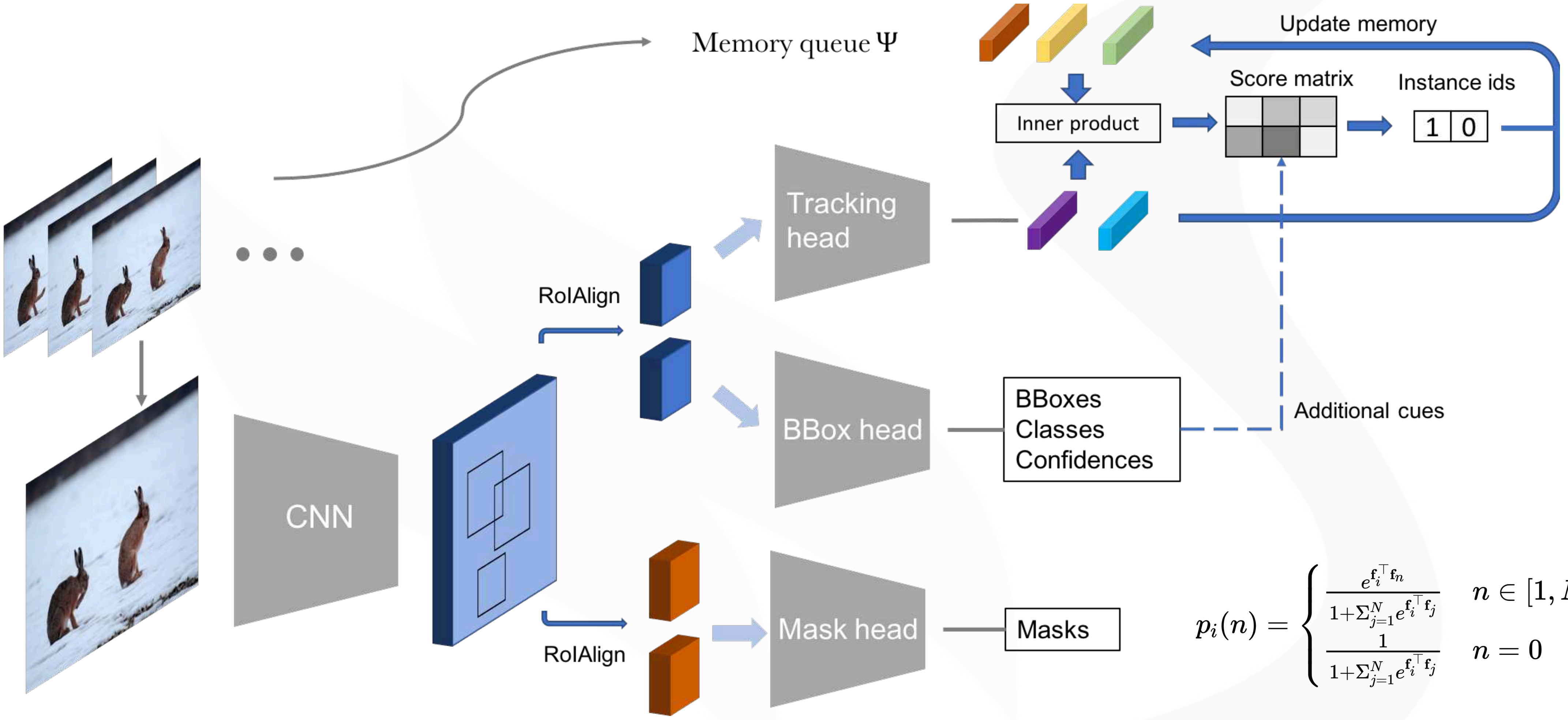
Figure 1: The **Mask R-CNN** framework for instance segmentation.

Mask R-CNN which was a state-of-the-art method for image instance segmentation

Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

MaskTrack R-CNN



$$p_i(n) = \begin{cases} \dfrac{e^{\mathbf{f}_i^\top \mathbf{f}_n}}{1+\Sigma_{j=1}^{N} e^{\mathbf{f}_i^\top \mathbf{f}_j}} & n \in [1, N] \\ \dfrac{1}{1+\Sigma_{j=1}^{N} e^{\mathbf{f}_i^\top \mathbf{f}_j}} & n = 0 \end{cases}$$

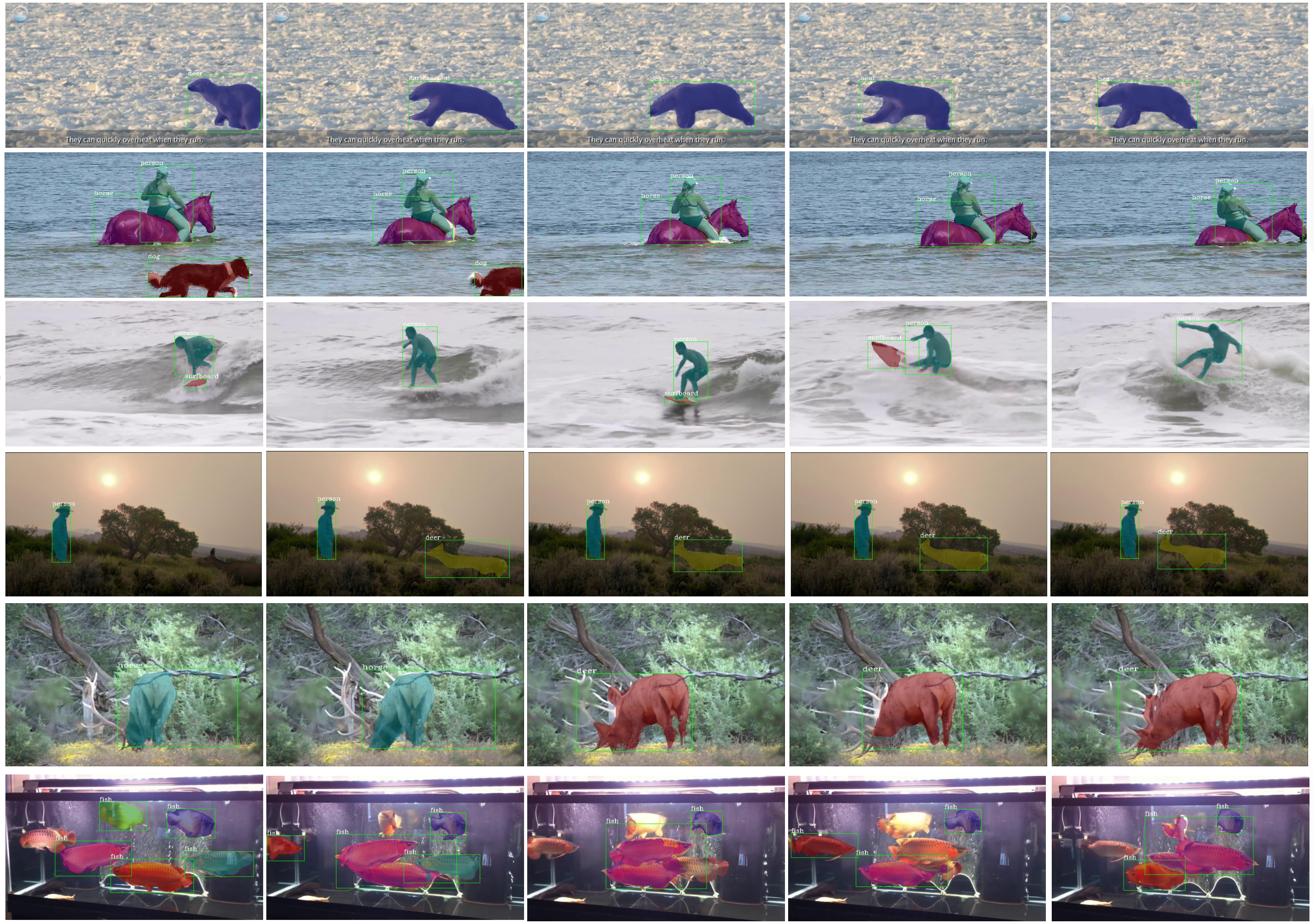
Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

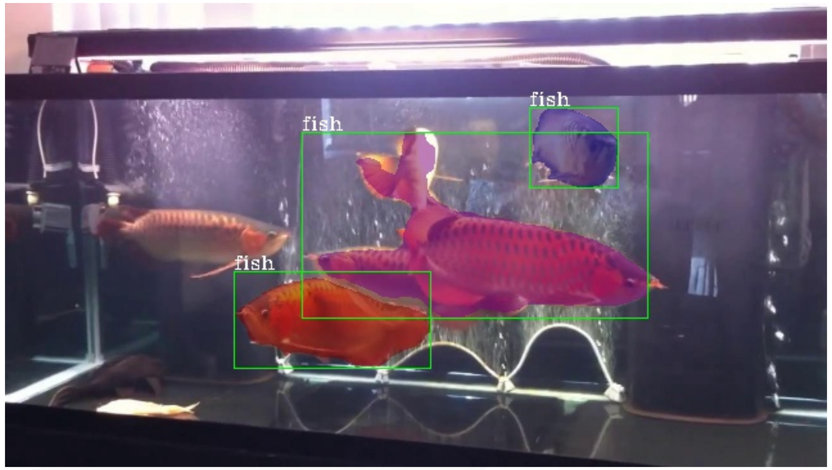
Video Instance Segmentation

BEIJING JIAOTONG UNIVERSITY

Main Results

(a)
(b)
(c)
(d)
(e)
(f)





Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

Video Instance Segmentation

Applications

augmented reality

video editing

autonomous driving

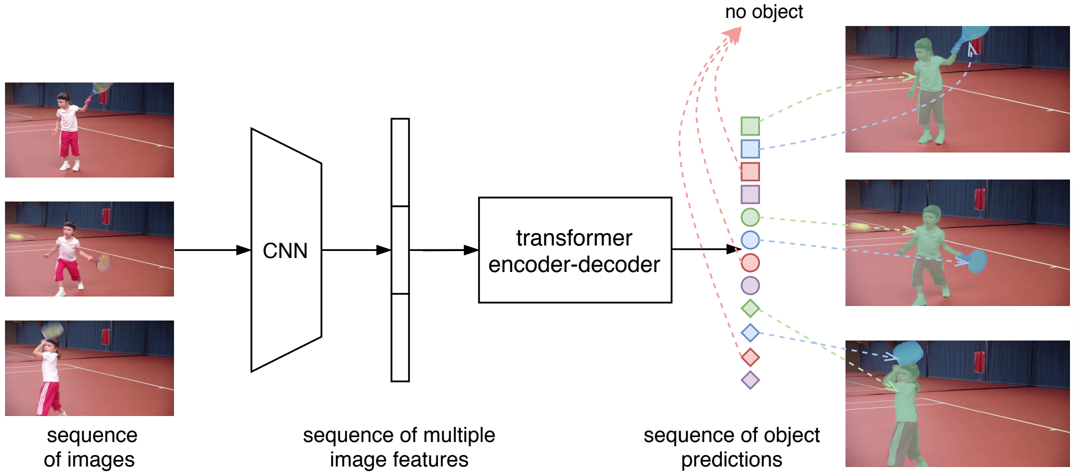
Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In IEEE International Conference on Computer Vision. 5188-5197.

End-to-End Video Instance Segmentation with Transformers

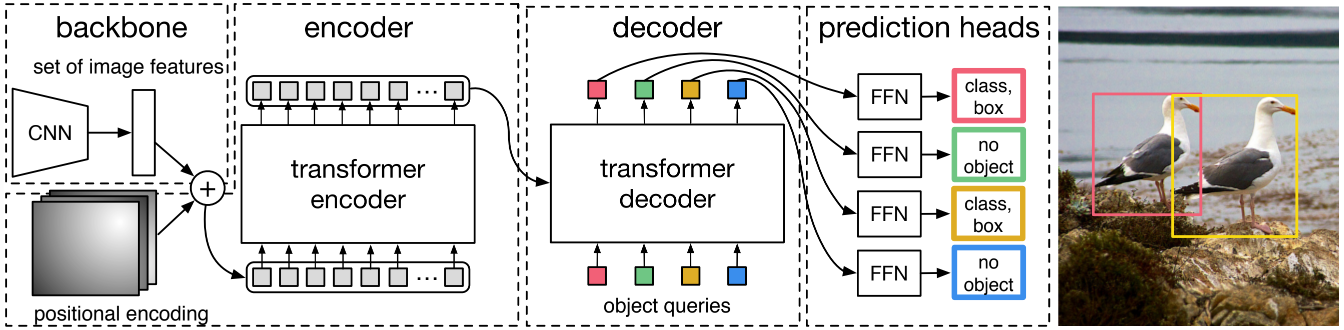
Yuqing Wang[1], Zhaoliang Xu[1], Xinlong Wang[2], Chunhua Shen[2], Baoshan Cheng[1], Hao Shen[1*], Huaxia Xia[1]
[1] Meituan [2] The University of Adelaide, Australia

End-to-End Video Instance Segmentation with Transformers

Motivation



VisTR

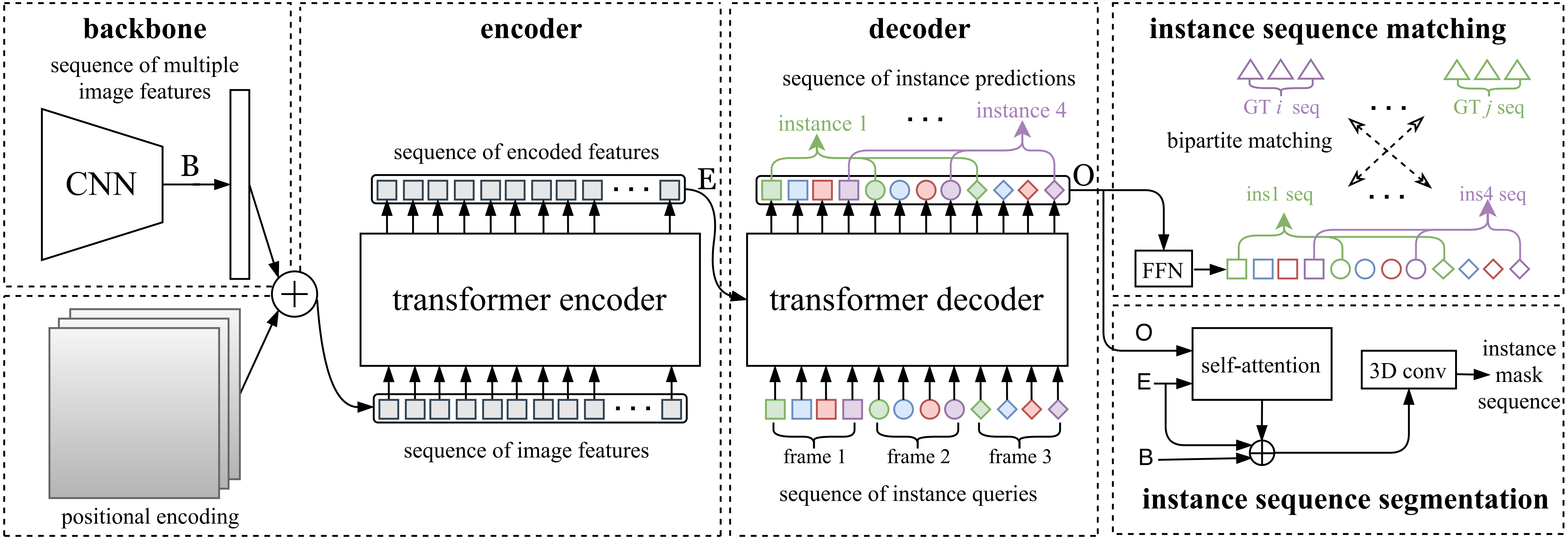


DETR

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-end video instance segmentation with transformers. In IEEE Conference on Computer Vision and Pattern Recognition. 8741–8750

End-to-End Video Instance Segmentation with Transformers

VisTR



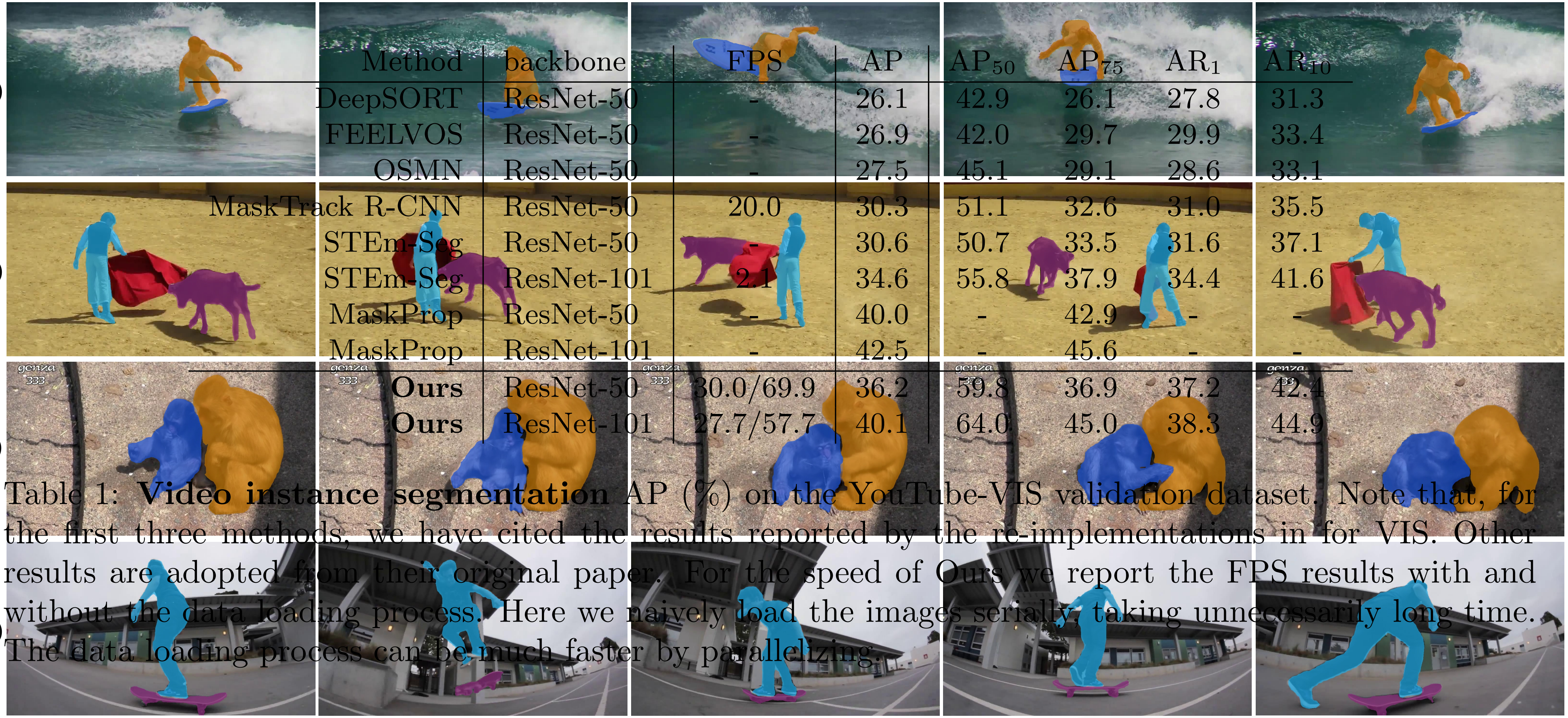
Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-end video instance segmentation with transformers. In IEEE Conference on Computer Vision and Pattern Recognition. 8741–8750

End-to-End Video Instance Segmentation with Transformers

BEIJING JIAOTONG UNIVERSITY

Experiments

(a)
(b)
(c)
(d)



Method	backbone	FPS	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
DeepSORT	ResNet-50	-	26.1	42.9	26.1	27.8	31.3
FEELVOS	ResNet-50	-	26.9	42.0	29.7	29.9	33.4
OSMN	ResNet-50	-	27.5	45.1	29.1	28.6	33.1
MaskTrack R-CNN	ResNet-50	20.0	30.3	51.1	32.6	31.0	35.5
STEm-Seg	ResNet-50	-	30.6	50.7	33.5	31.6	37.1
STEm-Seg	ResNet-101	2.1	34.6	55.8	37.9	34.4	41.6
MaskProp	ResNet-50	-	40.0	-	42.9	-	-
MaskProp	ResNet-101	-	42.5	-	45.6	-	-
Ours	ResNet-50	30.0/69.9	36.2	59.8	36.9	37.2	42.4
Ours	ResNet-101	27.7/57.7	40.1	64.0	45.0	38.3	44.9

Table 1: **Video instance segmentation** AP (%) on the YouTube-VIS validation dataset. Note that, for the first three methods, we have cited the results reported by the re-implementations in for VIS. Other results are adopted from their original paper. For the speed of Ours we report the FPS results with and without the data loading process. Here we naively load the images serially, taking unnecessarily long time. The data loading process can be much faster by parallelizing.

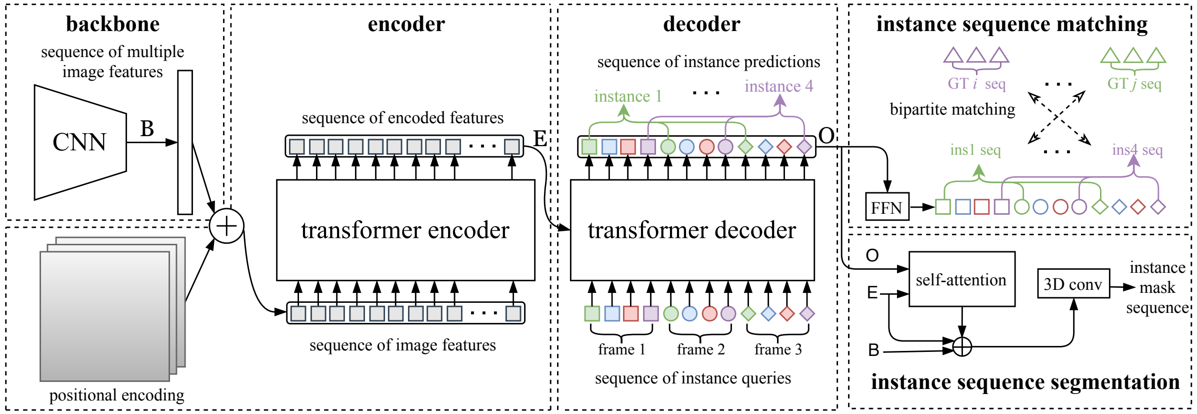
Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-end video instance segmentation with transformers. In IEEE Conference on Computer Vision and Pattern Recognition. 8741–8750

SeqFormer: Sequential Transformer for Video Instance Segmentation

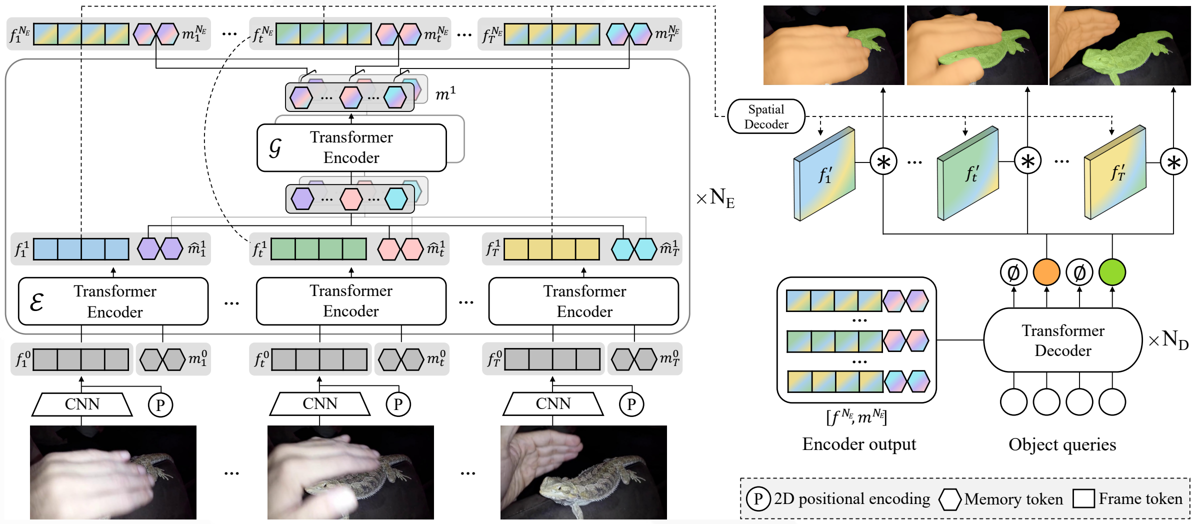
Junfeng Wu[1*], Yi Jiang[2], Song Bai[2], Wenqing Zhang[1*], and Xiang Bai[1†]
[1] Huazhong University of Science and Technology [2] Bytedance

BEIJING JIAOTONG UNIVERSITY

Motivation



VisTR

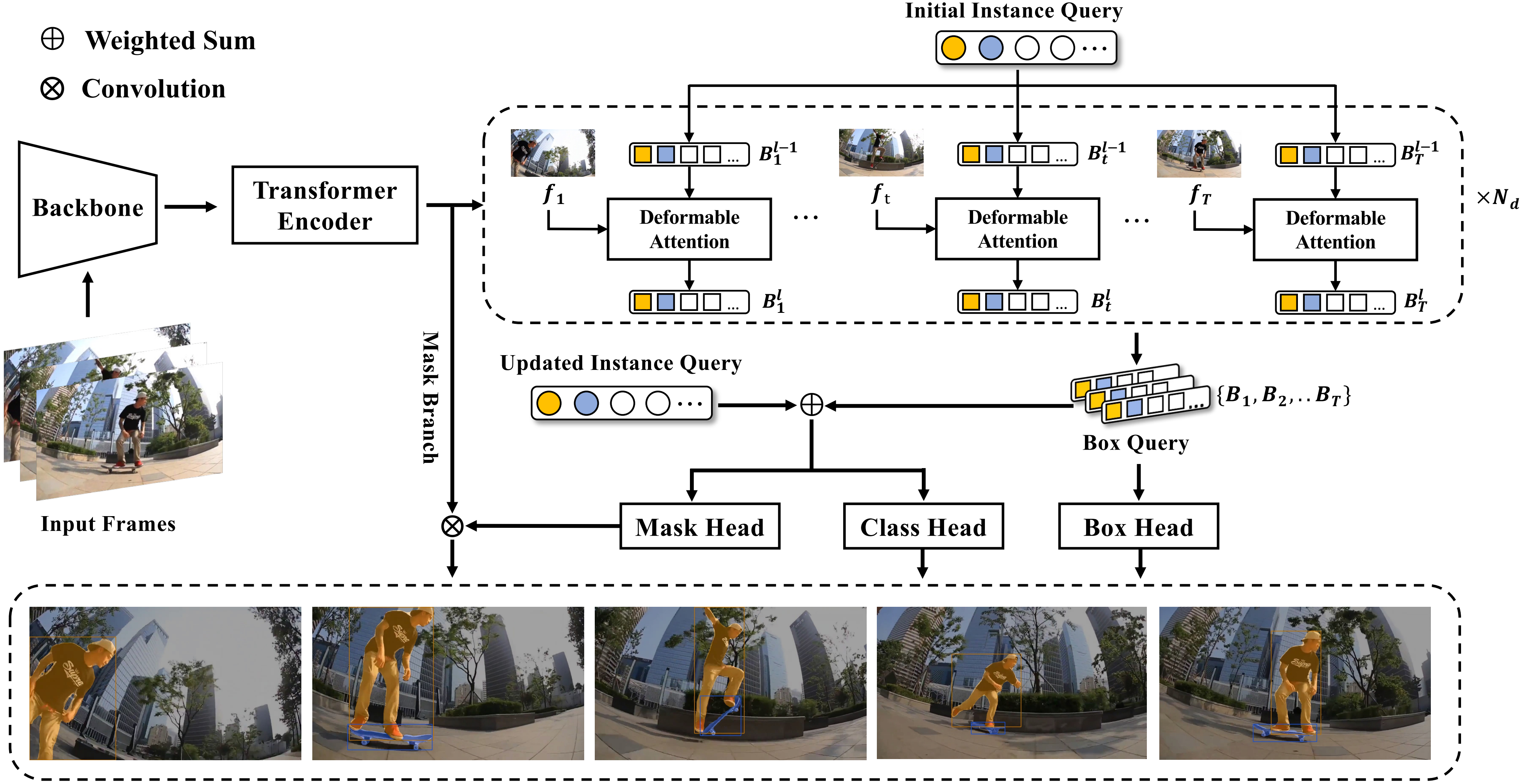


IFC

a stand-alone instance query suffices for capturing a time sequence of instances in a video, but attention mechanisms shall be done with each frame independently

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. 2022. Seqformer: Sequential transformer for video instance segmentation. In European Conference on Computer Vision. 553–569.

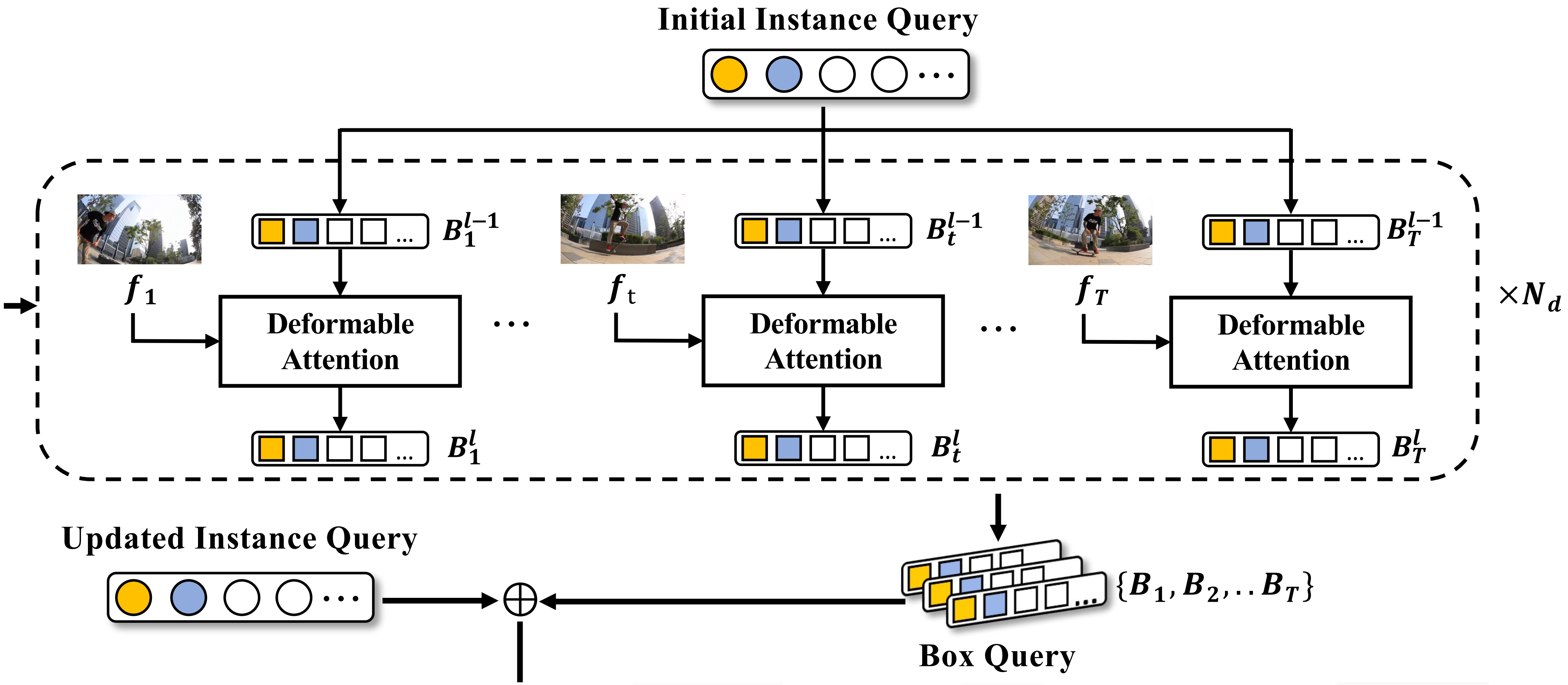
SeqFormer



Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. 2022. Seqformer: Sequential transformer for video instance segmentation. In European Conference on Computer Vision. 553–569.

SeqFormer: Sequential Transformer for Video Instance Segmentation

Query Decompose Transformer Decoder



$$\mathbf{B}_t^1 = \mathrm{DeformAttn}(\mathbf{I}_q, \mathbf{f}_t)$$

$$\mathbf{I}_q^l = \frac{\sum_{t=1}^{T} \mathbf{B}_t^l \times \mathrm{FC}(\mathbf{B}_t^l)}{\sum_{t=1}^{T} \mathrm{FC}(\mathbf{B}_t^l)} + \mathbf{I}_q^{l-1}$$

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. 2022. Seqformer: Sequential transformer for video instance segmentation. In European Conference on Computer Vision. 553–569.

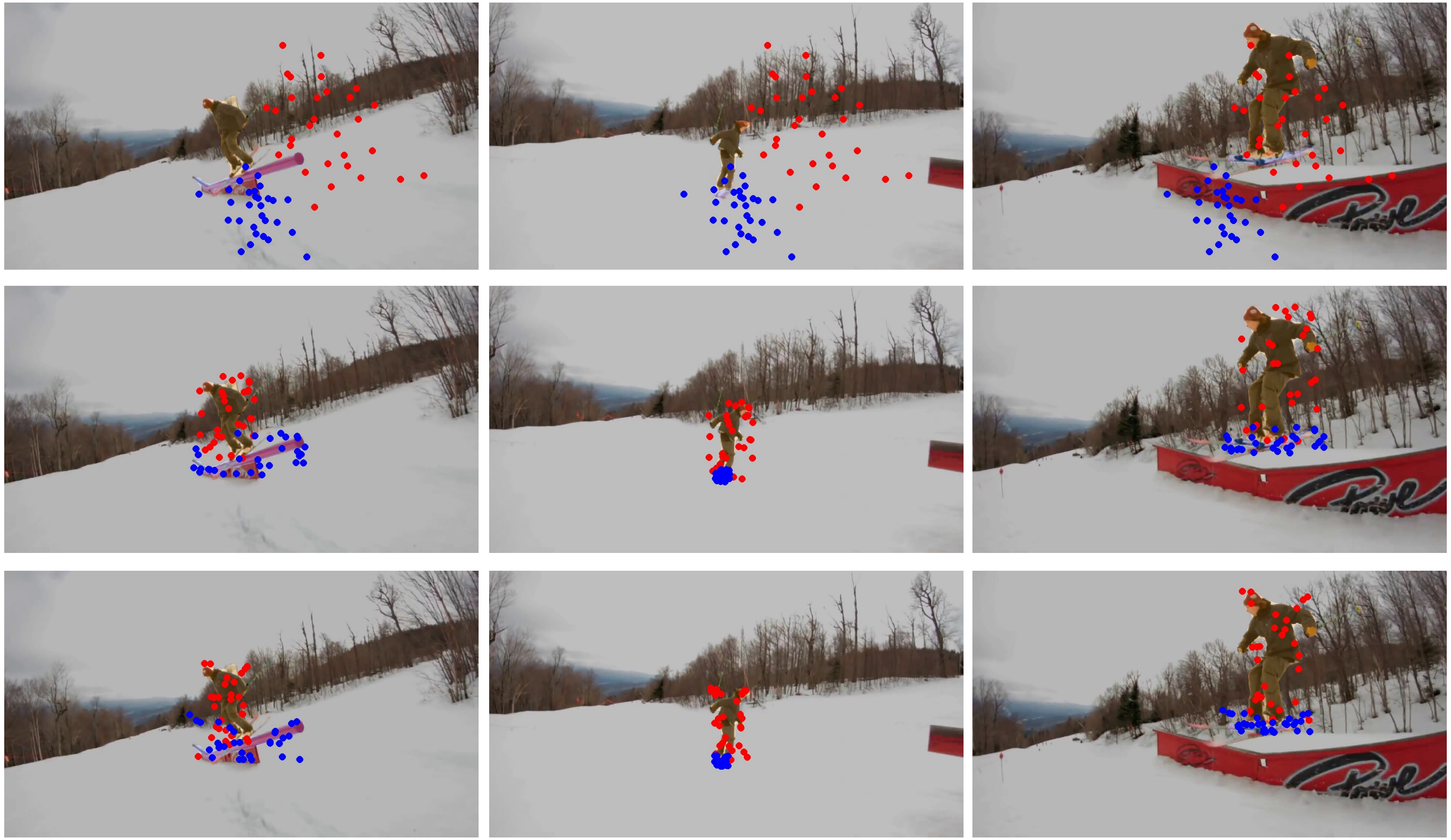
SeqFormer: Sequential Transformer for Video Instance Segmentation

北京交通大學 BEIJING JIAOTONG UNIVERSITY

(a)

(b)

(c)



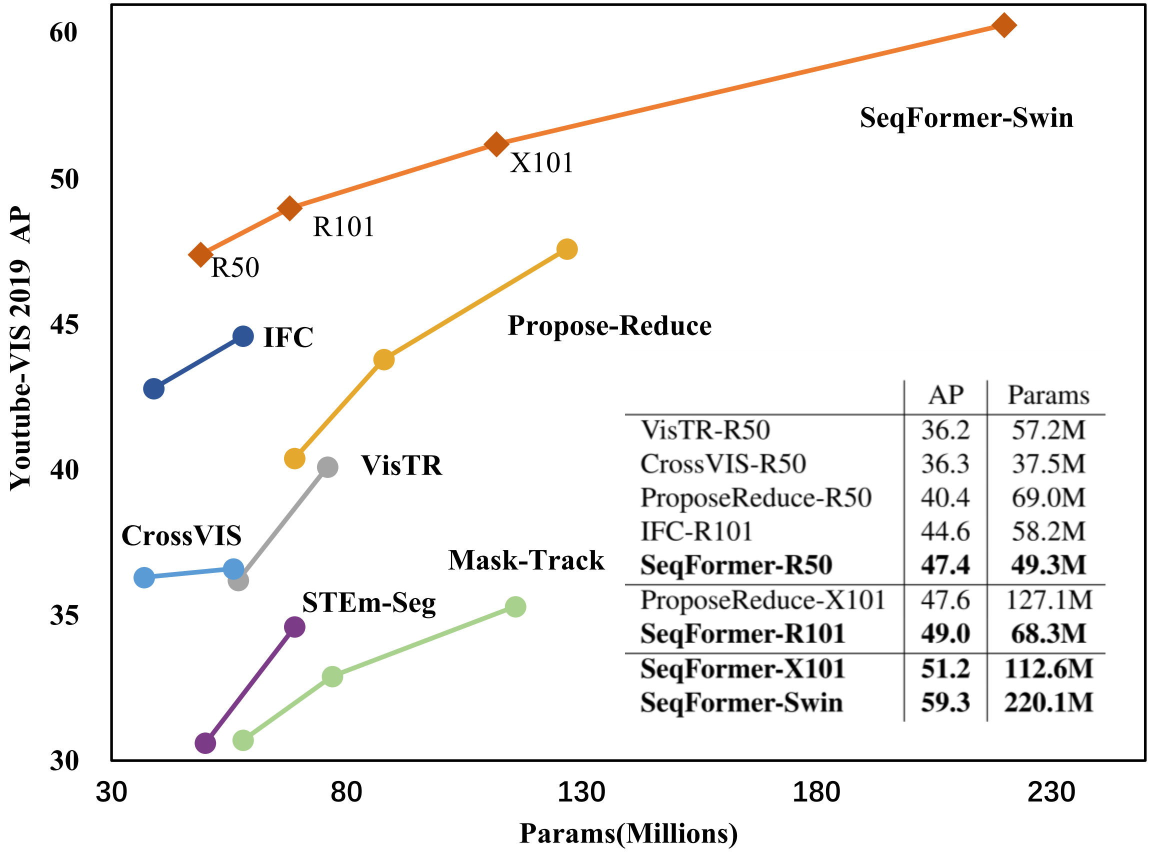
(a) sampling points from the first decoder layer. The refined accurate sampling points from the second and last decoder layer are shown in (b) and (c)

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. 2022. Seqformer: Sequential transformer for video instance segmentation. In European Conference on Computer Vision. 553–569.

北京交通大學 BEIJING JIAOTONG UNIVERSITY

Experiments

Backbone	Method	Params	FPS	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MaskTrack R-CNN	58.1M	20.0	30.3	51.1	32.6	31.0	35.5
	STEm-Seg	50.5M	7.0	30.6	50.7	33.5	37.6	37.1
	SipMask	33.2M	30.0	33.7	54.1	35.8	35.4	40.1
	CompFeat	-	-	35.3	56.0	38.6	33.1	40.3
	SG-Net	-	-	34.8	56.1	36.8	35.8	40.8
	VisTR	57.2M	69.9	36.2	59.8	36.9	37.2	42.4
	MaskProp	-	-	40.0	-	42.9	-	-
	CrossVIS	37.5M	39.8	36.3	56.8	38.9	35.6	40.7
	Propose-Reduce	69.0M	-	40.4	63.0	43.8	41.1	49.7
	IFC [?]	39.3M	107.1	42.8	65.8	46.8	43.8	51.2
	SeqFormer†	49.3M	72.3	45.1	66.9	50.5	**45.6**	54.6
	SeqFormer	49.3M	72.3	**47.4**	**69.8**	**51.8**	45.5	**54.8**
ResNet-101	MaskTrack R-CNN	77.2M	-	31.8	53.0	33.6	33.2	37.6
	STEm-Seg	69.6M	-	34.6	55.8	37.9	34.4	41.6
	SG-Net	-	-	36.3	57.1	39.6	35.9	43.0
	VisTR	76.3M	57.7	40.1	64.0	45.0	38.3	44.9
	MaskProp	-	-	42.5	-	45.6	-	-
	CrossVIS	56.6	35.6	36.6	57.3	39.7	36.0	42.0
	Propose-Reduce	88.1M	-	43.8	65.5	47.4	43.0	53.2
	IFC	58.3M	89.4	44.6	69.2	49.5	44.0	52.1
	SeqFormer	68.4M	64.6	**49.0**	**71.1**	**55.7**	**46.8**	**56.9**
ResNeXt-101	MaskProp	-	-	44.3	-	48.3	-	-
	Propose-Reduce	127.1M	-	47.6	71.6	51.8	46.3	56.0
	SeqFormer	112.7M	30.7	**51.2**	**75.3**	**58.0**	**46.5**	**57.3**
Swin-L	**SeqFormer**	220.0M	27.7	**59.3**	**82.1**	**66.4**	**51.7**	**64.4**



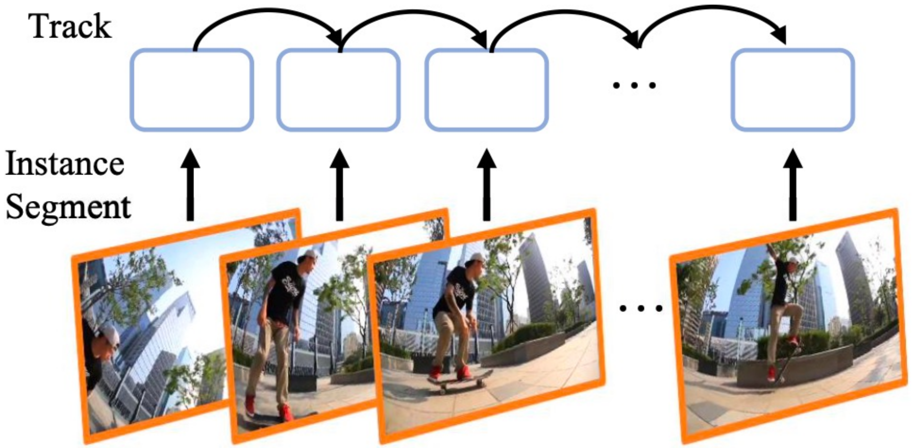
	AP	Params
VisTR-R50	36.2	57.2M
CrossVIS-R50	36.3	37.5M
ProposeReduce-R50	40.4	69.0M
IFC-R101	44.6	58.2M
SeqFormer-R50	**47.4**	**49.3M**
ProposeReduce-X101	47.6	127.1M
SeqFormer-R101	**49.0**	**68.3M**
SeqFormer-X101	**51.2**	**112.6M**
SeqFormer-Swin	**59.3**	**220.1M**

Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. 2022. Seqformer: Sequential transformer for video instance segmentation. In European Conference on Computer Vision. 553–569.

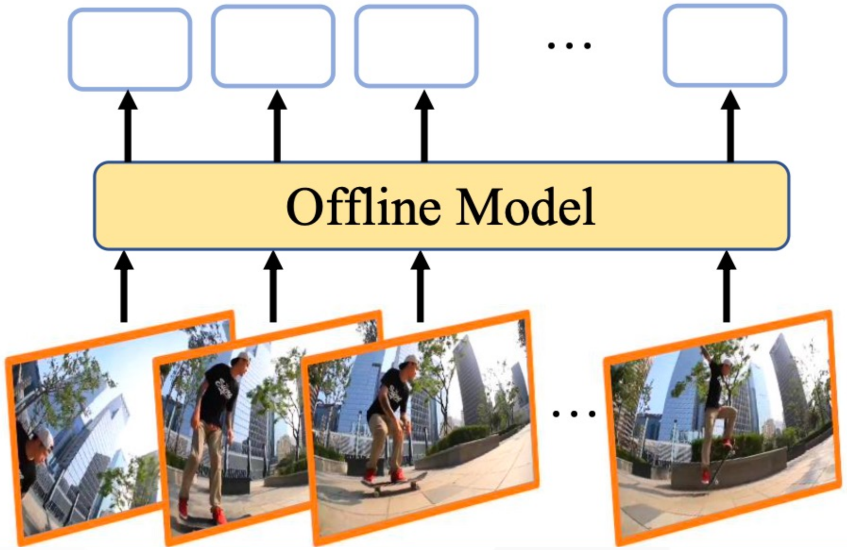
In Defense of Online Models for Video Instance Segmentation

Junfeng Wu[1*] , Qihao Liu2[*], Yi Jiang[3], Song Bai[3†], Alan Yuille[2], Xiang Bai[1]
[1] Huazhong University of Science and Technology
[2] Johns Hopkins University [3] Bytedance

Online Methods



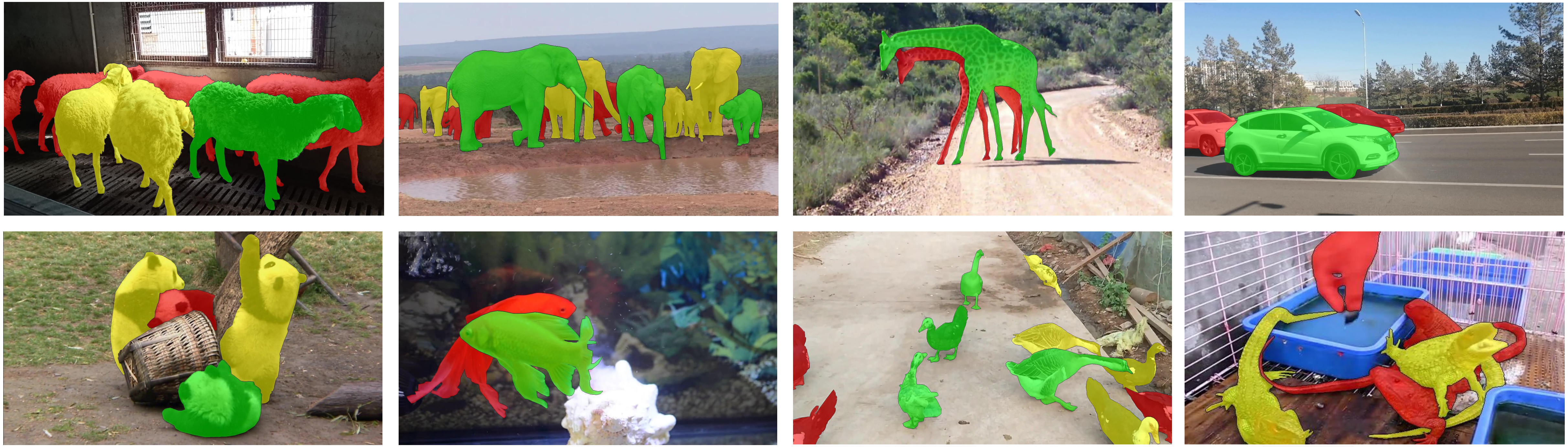
Offline Methods



Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

In Defense of Online Models for Video Instance Segmentation

OVIS

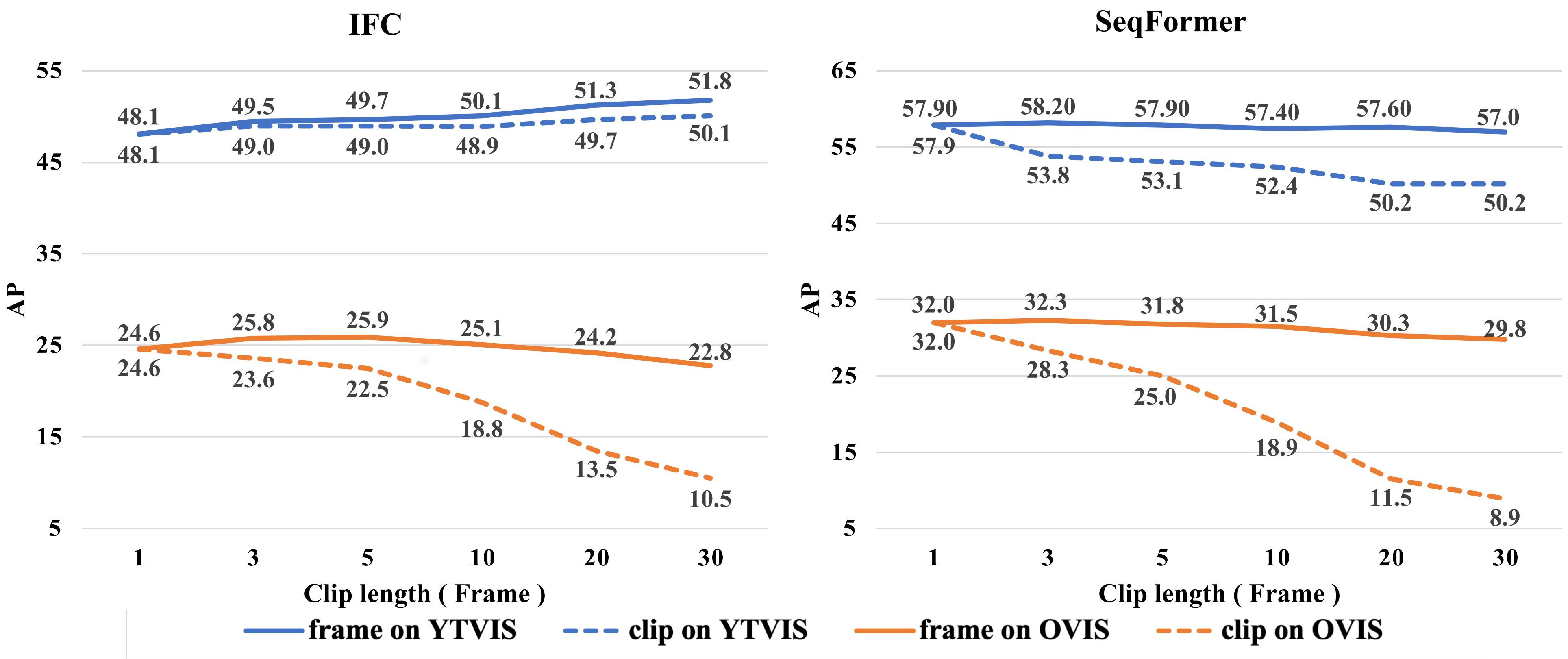


complex and occluded scenarios

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

BEIJING JIAOTONG UNIVERSITY

Oracle experiments on SOTA offline methods & motivation



For frame oracles, we provide the ground-truth instance ID both within each clip and between adjacent clips.
For clip oracles, we only provide the ground-truth instance ID between adjacent clips, and the method is required to do association within the clips by itself.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

BEIJING JIAOTONG UNIVERSITY

Online v.s. Offline

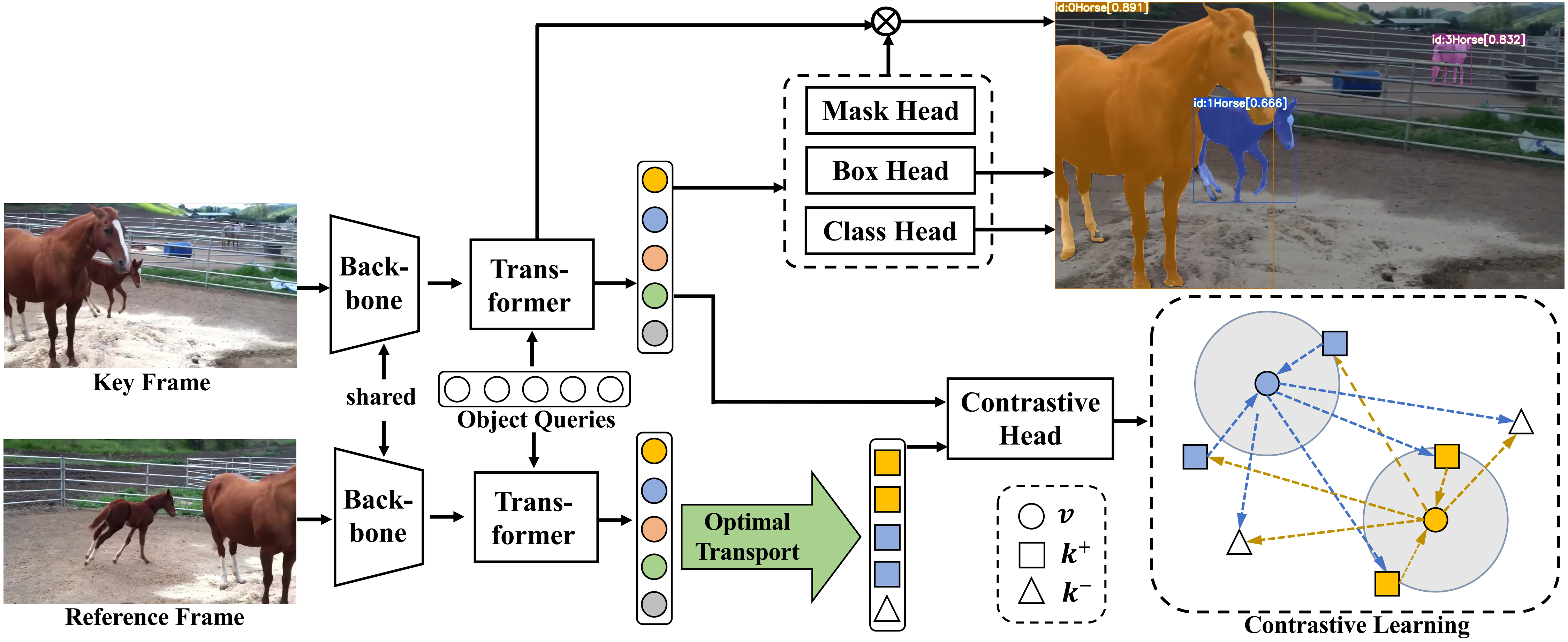
Dataset	Method	Publish	Predicted	Frame Oracle
YouTube-VIS	CrossVIS	ICCV 2021	43.4	52.8
	IFC	NeurIPS 2021	46.8	50.1
OVIS	CrossVIS	ICCV 2021	10.1	29.9
	IFC	NeurIPS 2021	8.7	25.1

Key insight: matching/association is the main reasoning for the performance gap

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

In Defense of Online Models for Video Instance Segmentation

IDOL

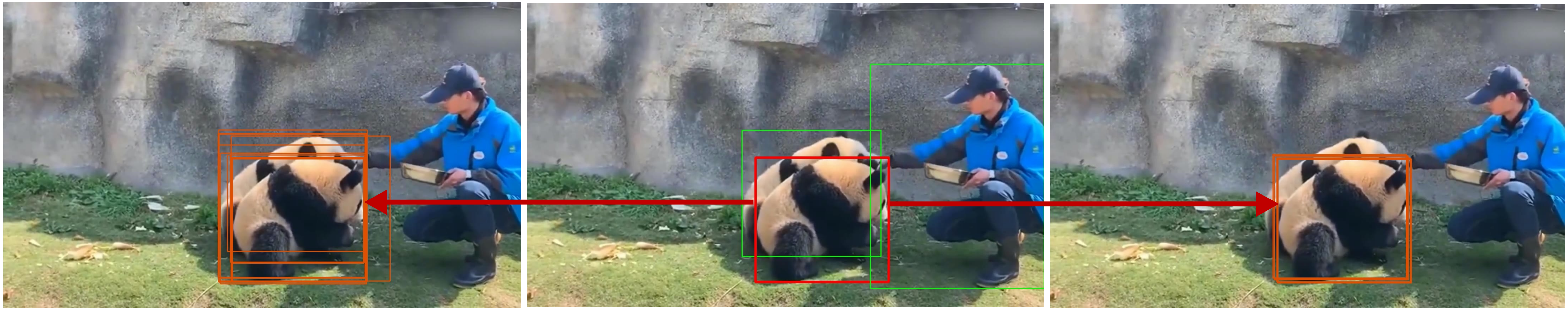


Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

In Defense of Online Models for Video Instance Segmentation

BEIJING JIAOTONG UNIVERSITY

Contrastive Learning



(a) IoU-Based (b) GT Box (c) Optimal Transport

The panda with red bounding box in (b) is the key instance. The positive samples selected by the IoU-based method are shown in (a), which causes false positives

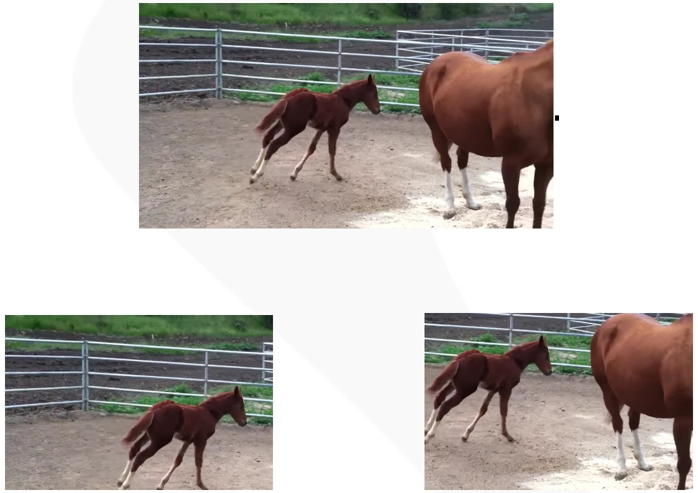
Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

In Defense of Online Models for Video Instance Segmentation

BEIJING JIAOTONG UNIVERSITY

Contrastive Learning

Method	mAP	ΔmAP	mAP_S	mAP_L
ResNet-50	42.4	-	45.5	39.2
Swin-L	53.0	10.6	57.6	48.4
+pseudo frame	55.2	2.2	59.7	50.7
+multi-scale	56.6	1.4	61.2	52.0
+multi-model	57.6	1.0	61.7	53.6



Swin-L: Integrated with the Swin Transformer backbone.
pseudo frame: Randomly crop a image from COCO twice to form a pseudo key-reference frame pair.
multi-scale testing: The shortest side is at [480, 640, 800].
multi-model: Ensemble Swin-L and ConvNext-L.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

In Defense of Online Models for Video Instance Segmentation

Experiments

Backbone	Method	Type	FPS	Data	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MaskTrack R-CNN [45]	online	20.0	V	30.3	51.1	32.6	31.0	35.5
	SipMask [4]	online	30.0	V	33.7	54.1	35.8	35.4	40.1
	CompFeat [8]	online	-	V	35.3	56.0	38.6	33.1	40.3
	CrossVIS [46]	online	39.8	V	36.3	56.8	38.9	35.6	40.7
	PCAN [18]	online	-	V	36.1	54.9	39.4	36.3	41.6
	STEm-Seg [1]	offline	7.0	V+I	30.6	50.7	33.5	37.6	37.1
	VisTR [37]	offline	69.9	V	36.2	59.8	36.9	37.2	42.4
	MaskProp [3]	offline	-	V	40.0	-	42.9	-	-
	Propose-Reduce [21]	offline	-	V+I	40.4	63.0	43.8	41.1	49.7
	IFC [16]	offline	107.1	V	42.8	65.8	46.8	43.8	51.2
	SeqFormer [40]	offline	72.3	V	45.1	66.9	50.5	45.6	54.6
	SeqFormer [40]	offline	72.3	V+I	47.4	69.8	51.8	45.5	54.8
	IDOL(ours)	online	30.6	V	46.4	70.7	51.9	44.8	54.9
	IDOL(ours)†	online	30.6	V	**49.5**	**74.0**	**52.9**	**47.7**	**58.7**
ResNet-101	MaskTrack R-CNN [45]	online	-	V	31.8	53.0	33.6	33.2	37.6
	CrossVIS [46]	online	35.6	V	36.6	57.3	39.7	36.0	42.0
	PCAN [18]	online	-	V	37.6	57.2	41.3	37.2	43.9
	STEm-Seg [1]	offline	-	V+I	34.6	55.8	37.9	34.4	41.6
	VisTR [37]	offline	57.7	V	40.1	64.0	45.0	38.3	44.9
	MaskProp [3]	offline	-	V	42.5	-	45.6	-	-
	Propose-Reduce [21]	offline	-	V+I	43.8	65.5	47.4	43.0	53.2
	IFC [16]	offline	89.4	V	44.6	69.2	49.5	44.0	52.1
	SeqFormer [40]	offline	64.6	V+I	49.0	71.1	55.7	46.8	56.9
	IDOL(ours)	online	26.0	V	48.2	**73.6**	52.5	45.6	55.5
	IDOL(ours)†	online	26.0	V	**50.1**	73.1	**56.1**	**47.0**	**57.9**
Swin-L	SeqFormer [40]	offline	27.7	V+I	59.3	82.1	66.4	51.7	64.4
	IDOL(ours)	online	17.6	V	61.5	84.2	**69.3**	53.3	65.6
	IDOL(ours)†	online	17.6	V	**62.2**	**86.5**	69.2	**54.6**	**68.1**

YouTube-VIS 2019

Backbone	Method	Type	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MaskTrack R-CNN [45]	online	28.6	48.9	29.6	26.5	33.8
	SipMask [4]	online	31.7	52.5	34.0	30.8	37.8
	STMask [20]	online	31.1	50.4	33.5	26.9	35.6
	CrossVIS [46]	online	34.2	54.4	37.9	30.4	38.2
	IFC [16]	offline	36.6	57.9	39.3	-	-
	SeqFormer [40]	offline	40.5	62.4	43.7	36.1	48.1
	IDOL(ours)	online	**43.9**	**68.0**	**49.6**	**38.0**	**50.9**
Swin-L	SeqFormer [40]	offline	51.8	74.6	58.2	42.8	58.1
	IDOL(ours)	online	**56.1**	**80.8**	**63.5**	**45.0**	**60.1**

YouTube-VIS 2021

Backbone	Method	Type	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	MaskTrack R-CNN [45]	online	10.8	25.3	8.5	7.9	14.9
	SipMask [4]	online	10.2	24.7	7.8	7.9	15.8
	CMaskTrack R-CNN [30]	online	15.4	33.9	13.1	9.3	20.0
	CrossVIS [46]	online	14.9	32.7	12.1	10.3	19.8
	STEm-Seg [1]	offline	13.8	32.1	11.9	9.1	20.0
	IFC† [16]	offline	13.1	27.8	11.6	9.4	23.9
	SeqFormer† [40]	offline	15.1	31.9	13.8	10.4	27.1
	IDOL(ours)	online	**30.2**	**51.3**	**30.0**	**15.0**	**37.5**
Swin-L	**IDOL(ours)**	online	**42.6**	**65.7**	**45.2**	**17.9**	**49.6**

OVIS

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

In Defense of Online Models for Video Instance Segmentation

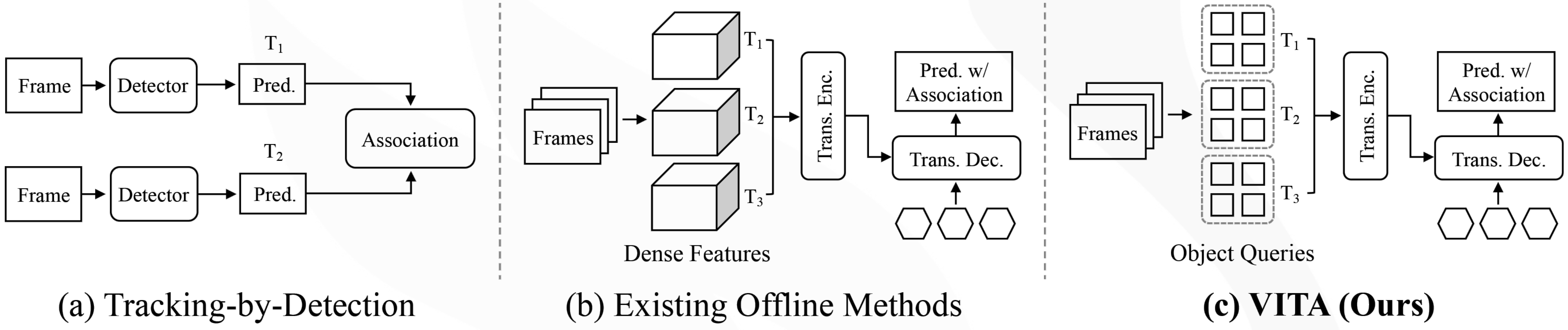
Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan L. Yuille, and Xiang Bai. 2022. In Defense of Online Models for Video Instance Segmentation. In European Conference on Computer Vision. 588–605.

VITA: Video Instance Segmentation via Object Token Association

Miran Heo[1*] Sukjun Hwang[1*] Seoung Wug Oh[2] Joon-Young Lee[2] Seon Joo Kim[1]
[1] Yonsei University [2] Adobe Research

VITA: Video Instance Segmentation via Object Token Association

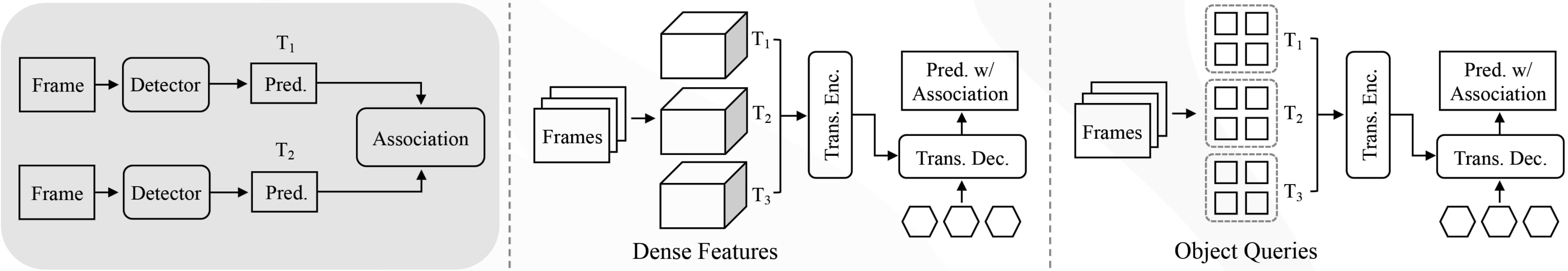
Motivation



Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

VITA: Video Instance Segmentation via Object Token Association

Motivation



(a) Tracking-by-Detection (b) Existing Offline Methods **(c) VITA (Ours)**

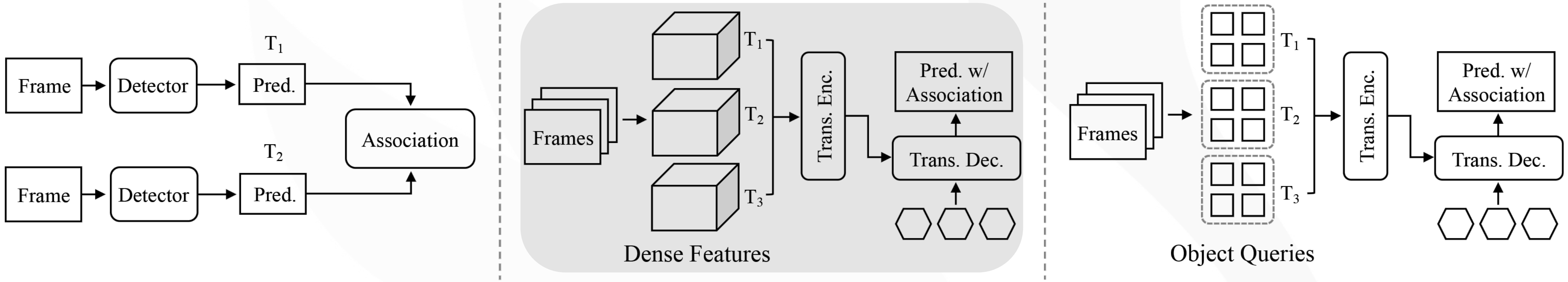
Link Per-frame Predictions using Heuristics
- Temporal Locality
- Equal Category, etc.

Limitations
- Low accuracy
- Hand-crafted tracking algorithms

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

VITA: Video Instance Segmentation via Object Token Association

Motivation



(a) Tracking-by-Detection (b) Existing Offline Methods **(c) VITA (Ours)**

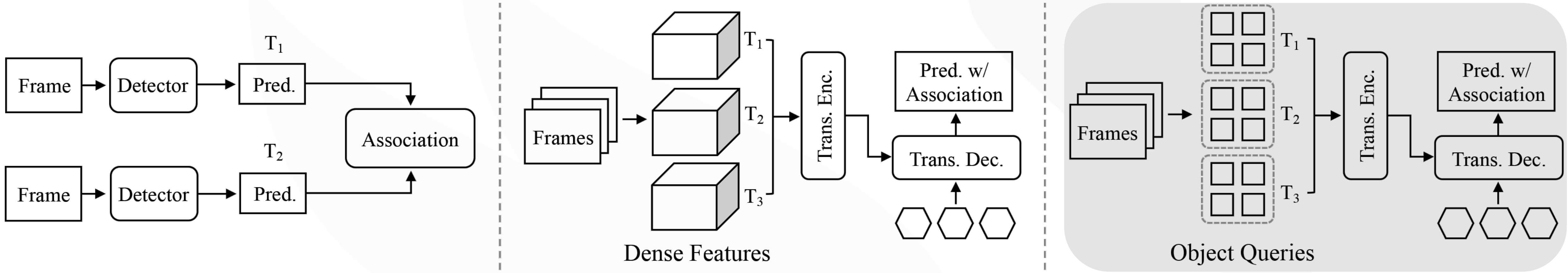
What if T = 300 ?

Limitations

- Heavy GPU memory consumption
- Numerous spatio-temporal tokens

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

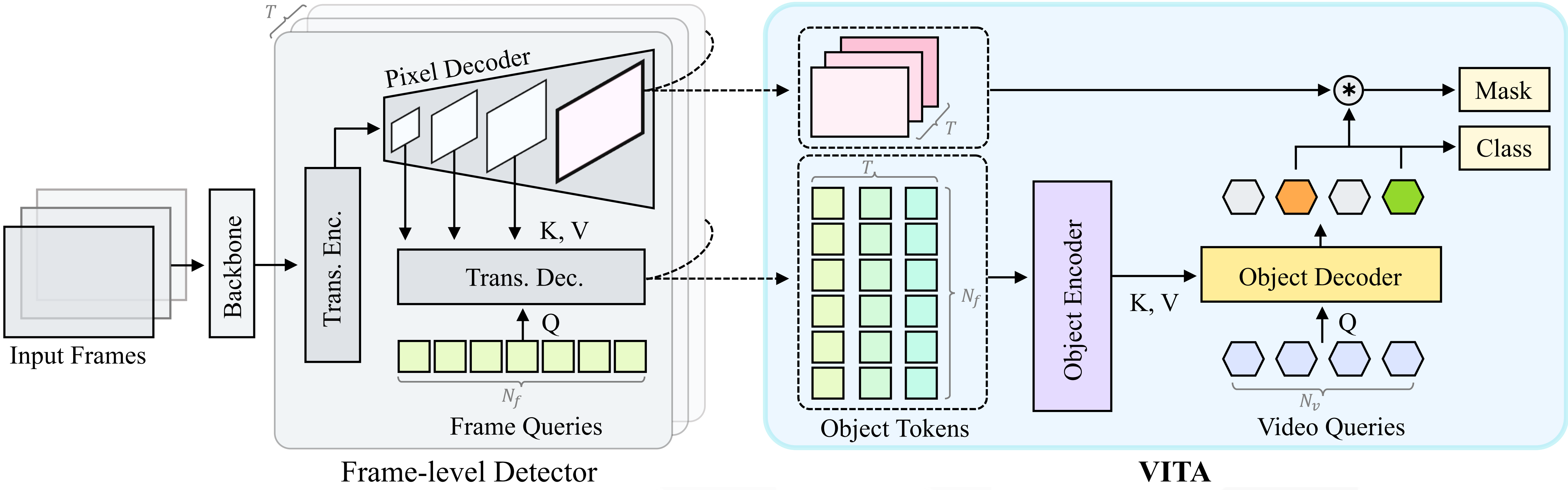
Motivation



(a) Tracking-by-Detection (b) Existing Offline Methods **(c) VITA (Ours)**

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

VITA



Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

Experiments

Table 1: Comparisons on YouTube-VIS 2019.

	Method	Backbone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
(Near) Online	MaskTrack R-CNN	ResNet-50	30.3	51.1	32.6	31.0	35.5
	MaskTrack R-CNN	ResNet-101	31.8	53.0	33.6	33.2	37.6
	CrossVIS	ResNet-50	36.3	56.8	38.9	35.6	40.7
	CrossVIS	ResNet-101	36.6	57.3	39.7	36.0	42.0
	PCAN	ResNet-50	36.1	54.9	39.4	36.3	41.6
	PCAN	ResNet-101	37.6	57.2	41.3	37.2	43.9
	EfficientVIS	ResNet-50	37.9	59.7	43.0	40.3	46.6
	EfficientVIS	ResNet-101	39.8	61.8	44.7	42.1	49.8
	VISOLO	ResNet-50	38.6	56.3	43.7	35.7	42.5
Offline	VisTR	ResNet-50	35.6	56.8	37.0	35.2	40.2
	VisTR	ResNet-101	38.6	61.3	42.3	37.6	44.2
	IFC	ResNet-50	41.2	65.1	44.6	42.3	49.6
	IFC	ResNet-101	42.6	66.6	46.3	43.5	51.4
	TeViT	MsgShifT	46.6	71.3	51.6	44.9	54.3
	SeqFormer	ResNet-50	47.4	69.8	51.8	45.5	54.8
	SeqFormer	ResNet-101	49.0	71.1	55.7	46.8	56.9
	SeqFormer	Swin-L	59.3	82.1	66.4	51.7	64.4
	Mask2Former-VIS	ResNet-50	46.4	68.0	50.0	-	-
	Mask2Former-VIS	ResNet-101	49.2	72.8	54.2	-	-
	Mask2Former-VIS	Swin-L	60.4	84.4	67.0	-	-
	VITA (Ours)	ResNet-50	49.8	72.6	54.5	49.4	61.0
		ResNet-101	51.9	75.4	57.0	49.6	59.1
		Swin-L	63.0	86.9	67.9	56.3	68.1

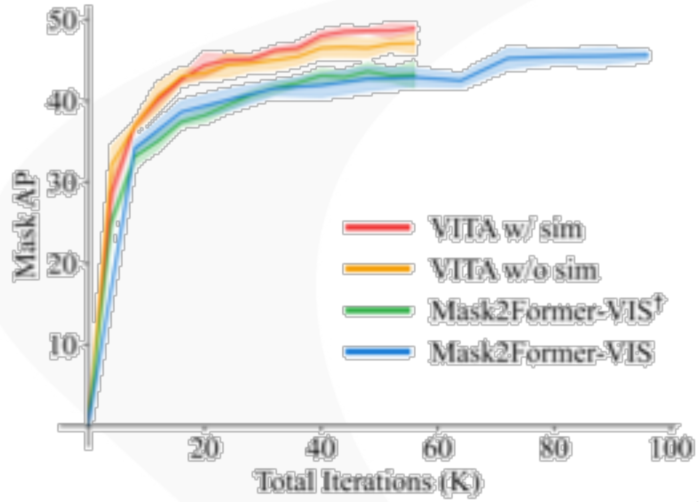
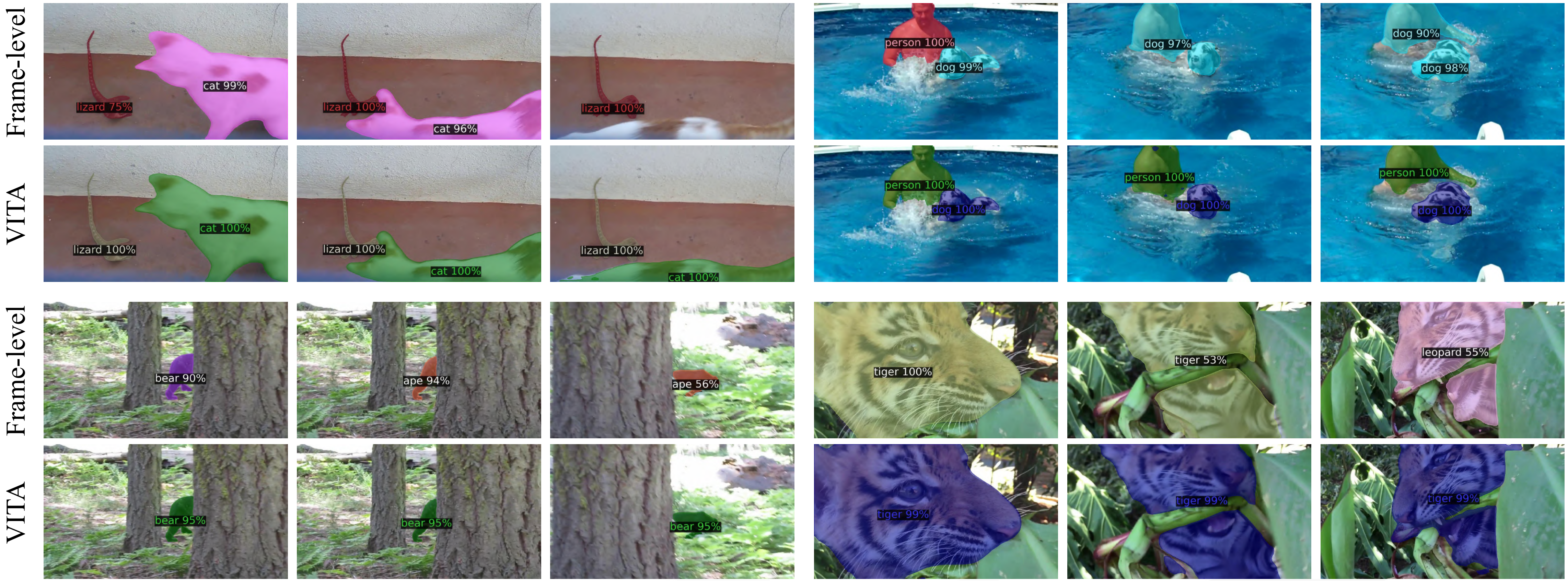


Table 1: Comparisons with ResNet-50 backbone on YouTube-VIS 2021 and OVIS. † indicates using MsgShifT backbone. ‡ indicates using Swin-L backbone.

Method	YouTube-VIS 2021 AP	AP_{50}	AP_{75}	AR_1	AR_{10}	OVIS AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN	28.6	48.9	29.6	26.5	33.8	10.8	25.3	8.5	7.9	14.9
CMaskTrack R-CNN	-	-	-	-	-	15.4	33.9	13.1	9.3	20.0
STMask	31.1	50.4	33.5	26.9	35.6	15.4	33.8	12.5	8.9	21.3
CrossVIS	34.2	54.4	37.9	30.4	38.2	14.9	32.7	12.1	10.3	19.8
IFC	35.2	55.9	37.7	32.6	42.9	-	-	-	-	-
VISOLO	36.9	54.7	40.2	30.6	40.9	15.3	31.0	13.8	11.1	21.7
TeViT†	37.9	61.2	42.1	35.1	44.6	17.4	34.9	15.0	11.2	21.8
SeqFormer	40.5	62.4	43.7	36.1	48.1	-	-	-	-	-
Mask2Former-VIS	40.6	60.9	41.8	-	-	-	-	-	-	-
VITA (Ours)	**45.7**	**67.4**	**49.5**	**40.9**	**53.6**	**19.6**	**41.2**	**17.4**	**11.7**	**26.0**
SeqFormer‡	51.8	74.6	58.2	42.8	58.1	-	-	-	-	-
Mask2Former-VIS‡	52.6	76.4	57.2	-	-	-	-	-	-	-
VITA (Ours)‡	**57.5**	**80.6**	**61.0**	**47.7**	**62.6**	**27.7**	**51.9**	**24.9**	**14.9**	**33.0**

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

Experiments



Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2022. VITA: Video Instance Segmentation via Object Token Association. arXiv preprint arXiv:2206.04403 (2022)

Reference

- **MaskTrack R-CNN**
  - The Ultimate Guide to Object Detection
  - 一文看懂视频实例分割任务VIS和VOS MOTS等的区别
  - 目标检测中评估指标mAP详解和计算方式
  - 图像实例分割评价指标
  - 目标检测-语义分割-实例分割模型常用性能评价指标
  - 一文读懂Faster RCNN
  - 实例分割算法（mask rcnn）总结
  - mask-rcnn 解读
  - Mask-RCNN 算法及其实现详解
- **VisTR**
  - 二分图和匈牙利算法
  - DETR 论文精读【论文精读】
  - 大白话用Transformer做object detection（上）
- **SeqFormer & IDOL**
  - 极市直播第100期丨ECCV2022 Oral-吴俊峰：视频实例分割新SOTA：SeqFormer&IDOL
  - Deformable DETR 详解
  - 论文笔记-DETR and Deformable DETR
  - Deformable DETR: 基于稀疏空间采样的注意力机制，让DCN与Transformer一起玩！
  - 目标检测: 一文读懂 OTA 标签分配
  - 论文阅读《OTA:Optimal Transport Assignment for Object Detection》

Thank you