JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

TarViS: A Unified Approach for Target-based Video Segmentation

Ali Athar[1], Alexander Hermans[1], Jonathon Luiten[1,2], Deva Ramanan[2], Bastian Leibe[1]

[1] RWTH Aachen University (Germany)   [2] Carnegie Mellon University (USA)



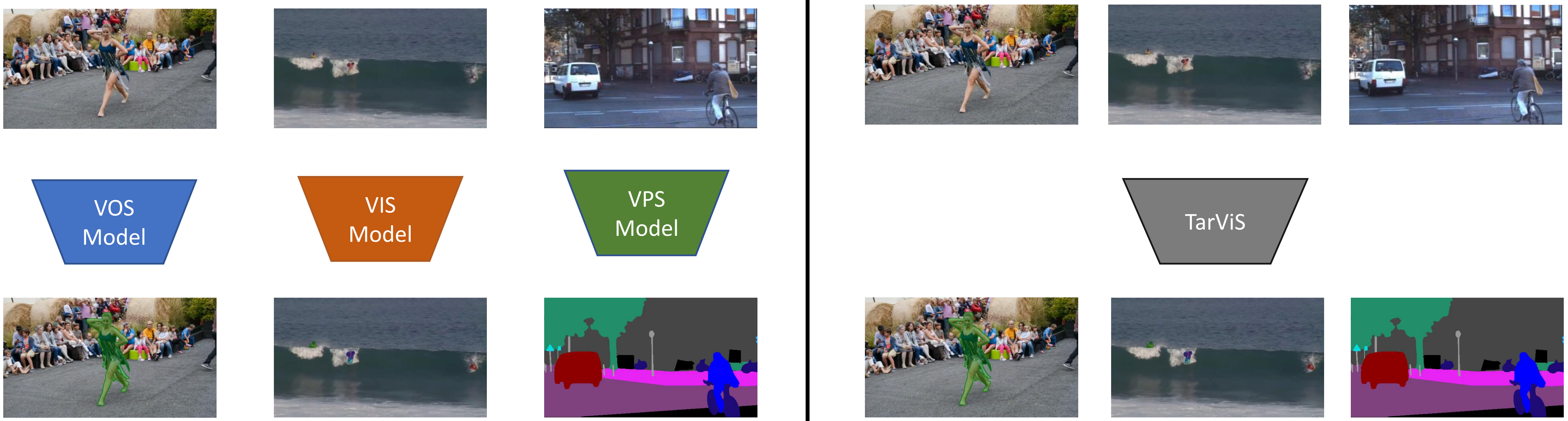


THU-AM-216

RWTH AACHEN UNIVERSITY

Carnegie Mellon University

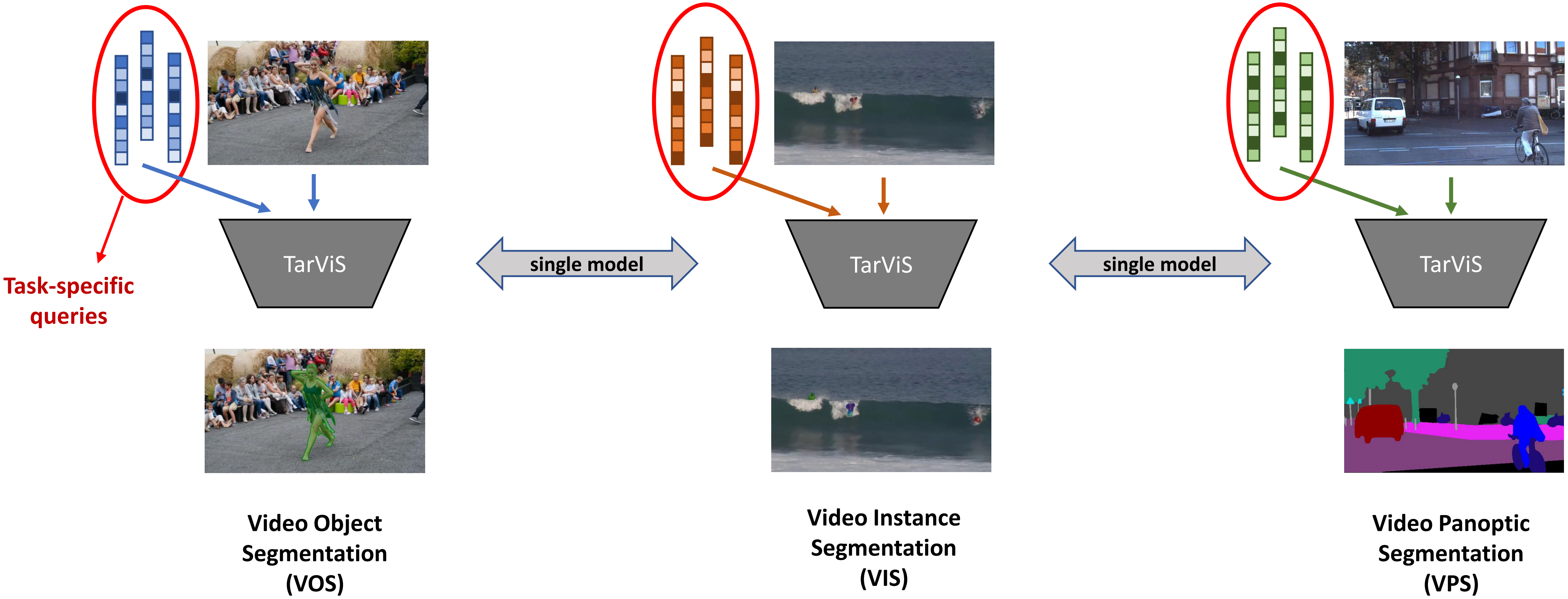
BEFORE ☹   NOW ☺



BEFORE: Video Object Segmentation (VOS) | Video Instance Segmentation (VIS) | Video Panoptic Segmentation (VPS)

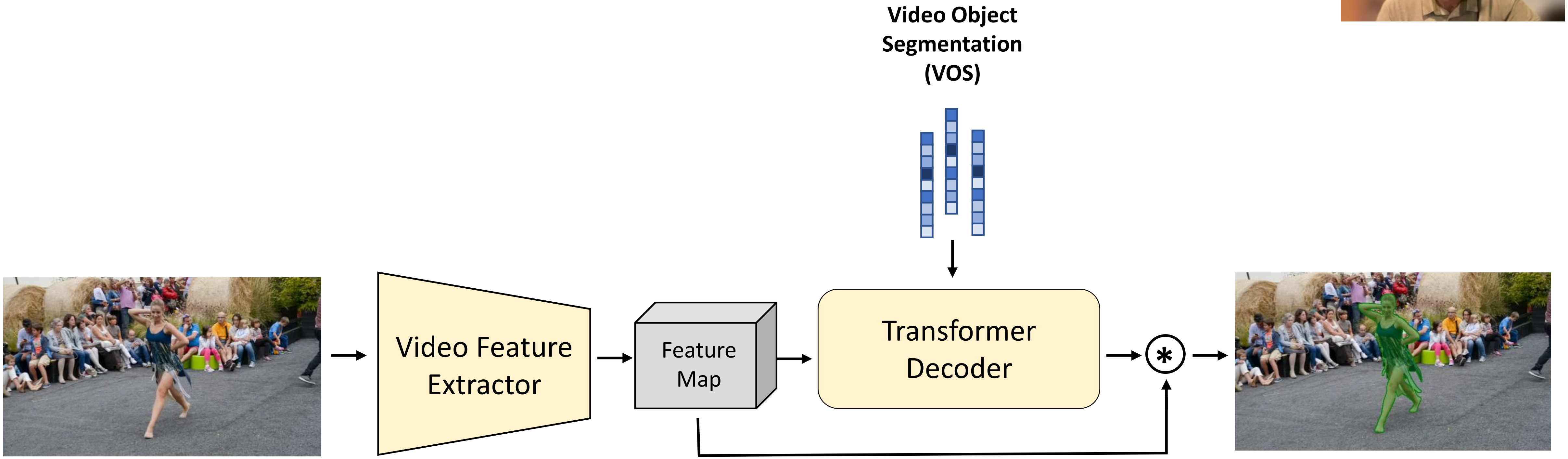
NOW: Video Object Segmentation (VOS) | Video Instance Segmentation (VIS) | Video Panoptic Segmentation (VPS)

Overview

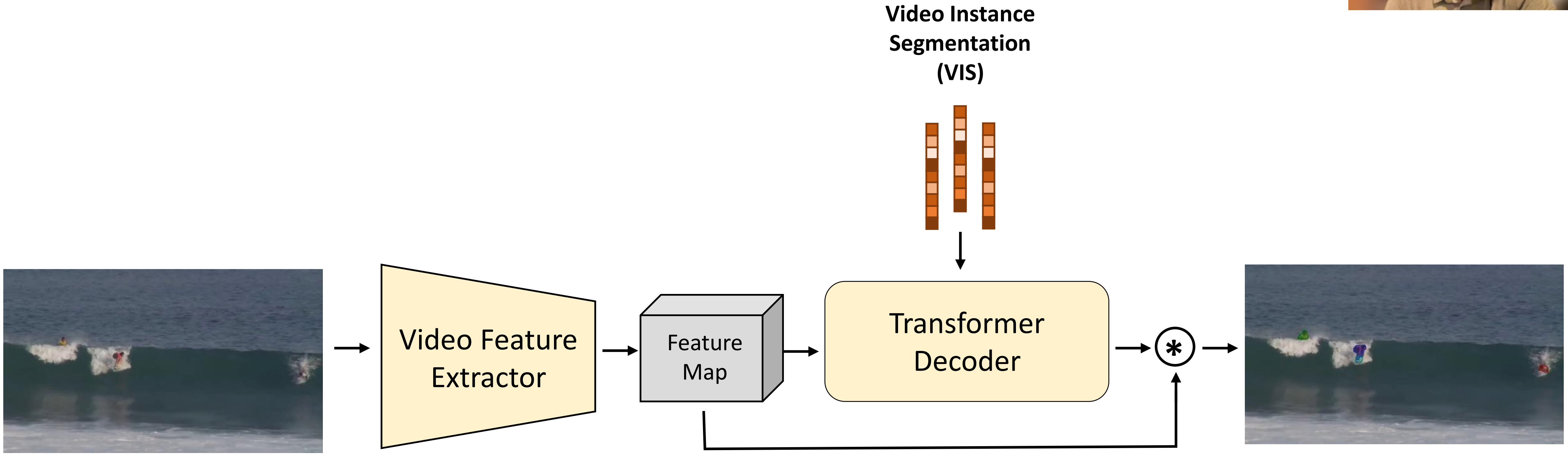


Task-specific queries
TarViS
single model
TarViS
single model
TarViS
Video Object Segmentation (VOS)
Video Instance Segmentation (VIS)
Video Panoptic Segmentation (VPS)

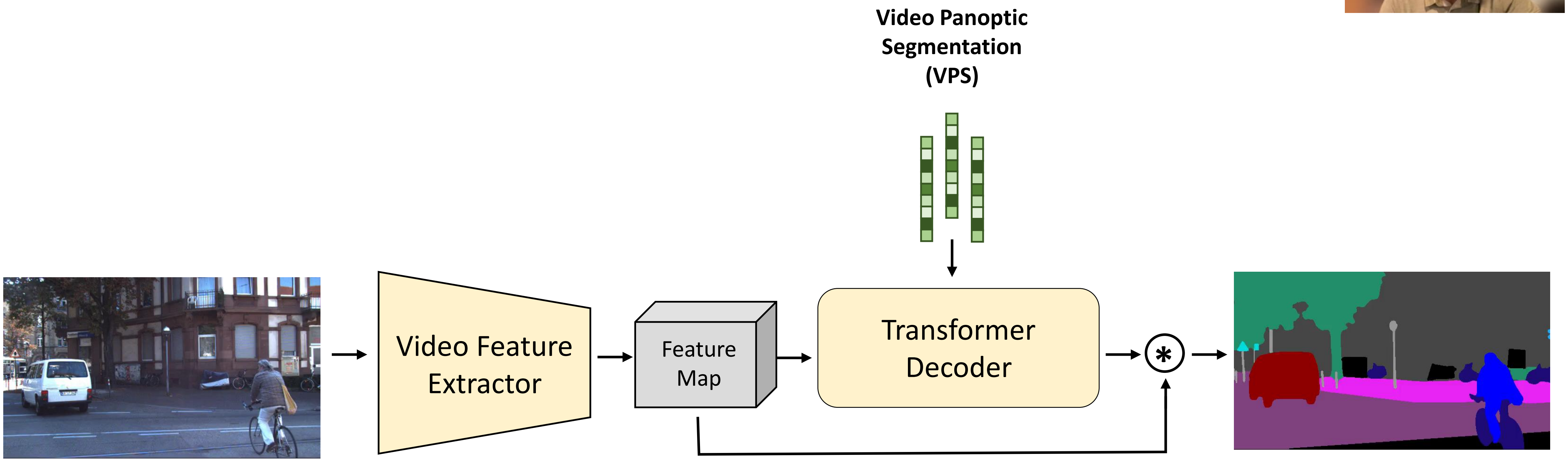
Overview


Video Object
Segmentation
(VOS)
Video Feature
Extractor
Feature
Map
Transformer
Decoder
*

Overview


Video Instance
Segmentation
(VIS)
Video Feature
Extractor
Feature
Map
Transformer
Decoder
*

Overview



Video Panoptic
Segmentation
(VPS)
Video Feature
Extractor
Feature
Map
Transformer
Decoder

Overview

Video Object Segmentation (VOS)

1. DAVIS

Video Instance Segmentation (VIS)

3. YouTube-VIS
4. OVIS

Point Exemplar-guided Tracking (PET)

2. BURST

Video Panoptic Segmentation (VPS)

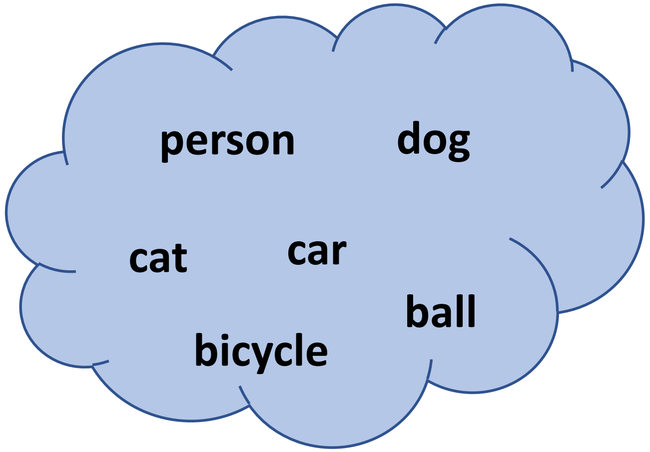
5. KITTI-STEP
6. CityscapesVPS
7. VIPSeg

Unified Task Definition

- Video segmentation tasks can be conceptually unified
- All of them require segmenting a set of ‘targets’ from the input video

- <u>Video Instance Segmentation:</u>

All objects belonging to these classes…

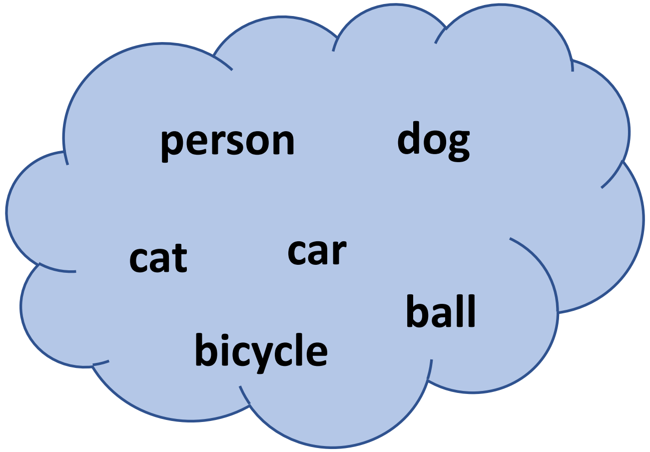


Unified Task Definition

- Video segmentation tasks can be conceptually unified
- All of them require segmenting a set of 'targets' from the input video

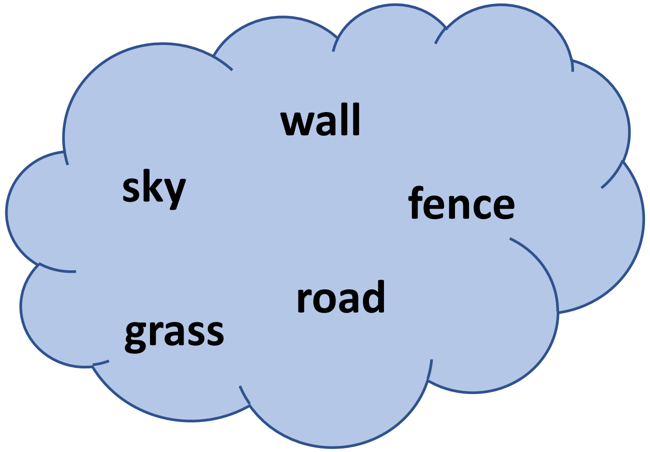
- <u>Video Panoptic Segmentation:</u>

All objects belonging to these classes…



thing classes

All parts of the…



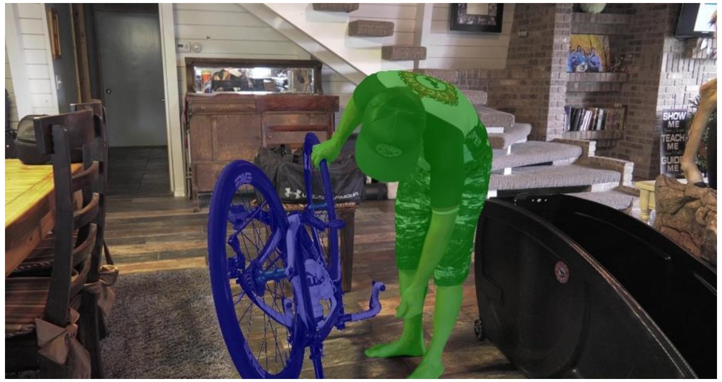
stuff classes

Unified Task Definition

- Video segmentation tasks can be conceptually unified
- All of them require segmenting a set of ‘targets’ from the input video

- <u>Video Object Segmentation (VOS):</u>

These specific objects with the first-frame masks…

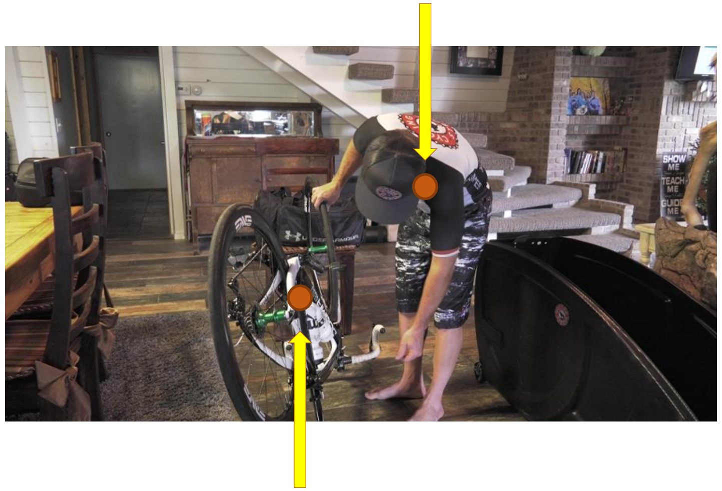


Unified Task Definition

- Video segmentation tasks can be conceptually unified
- All of them require segmenting a set of ‘targets’ from the input video

- <u>Point Exemplar-guided Tracking:</u>

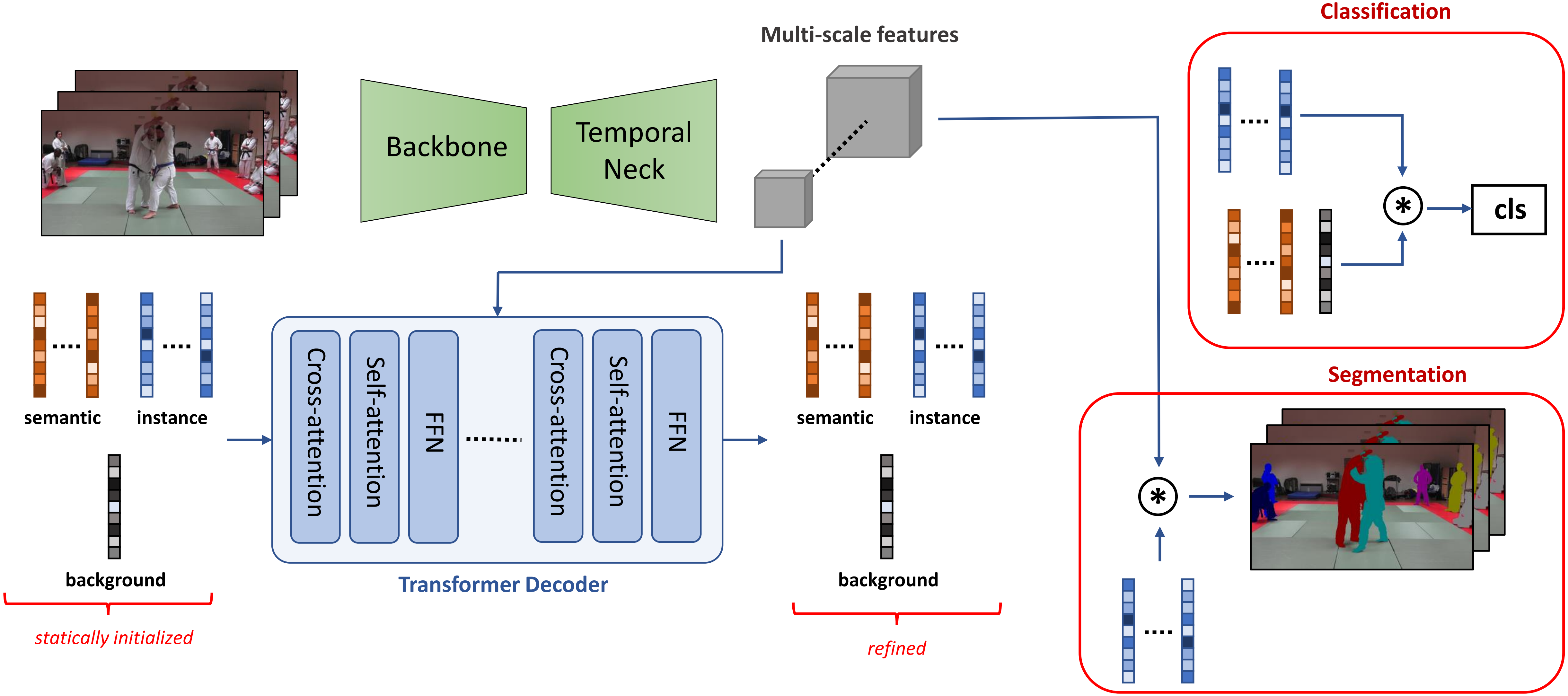
These specific objects with the first-frame points…



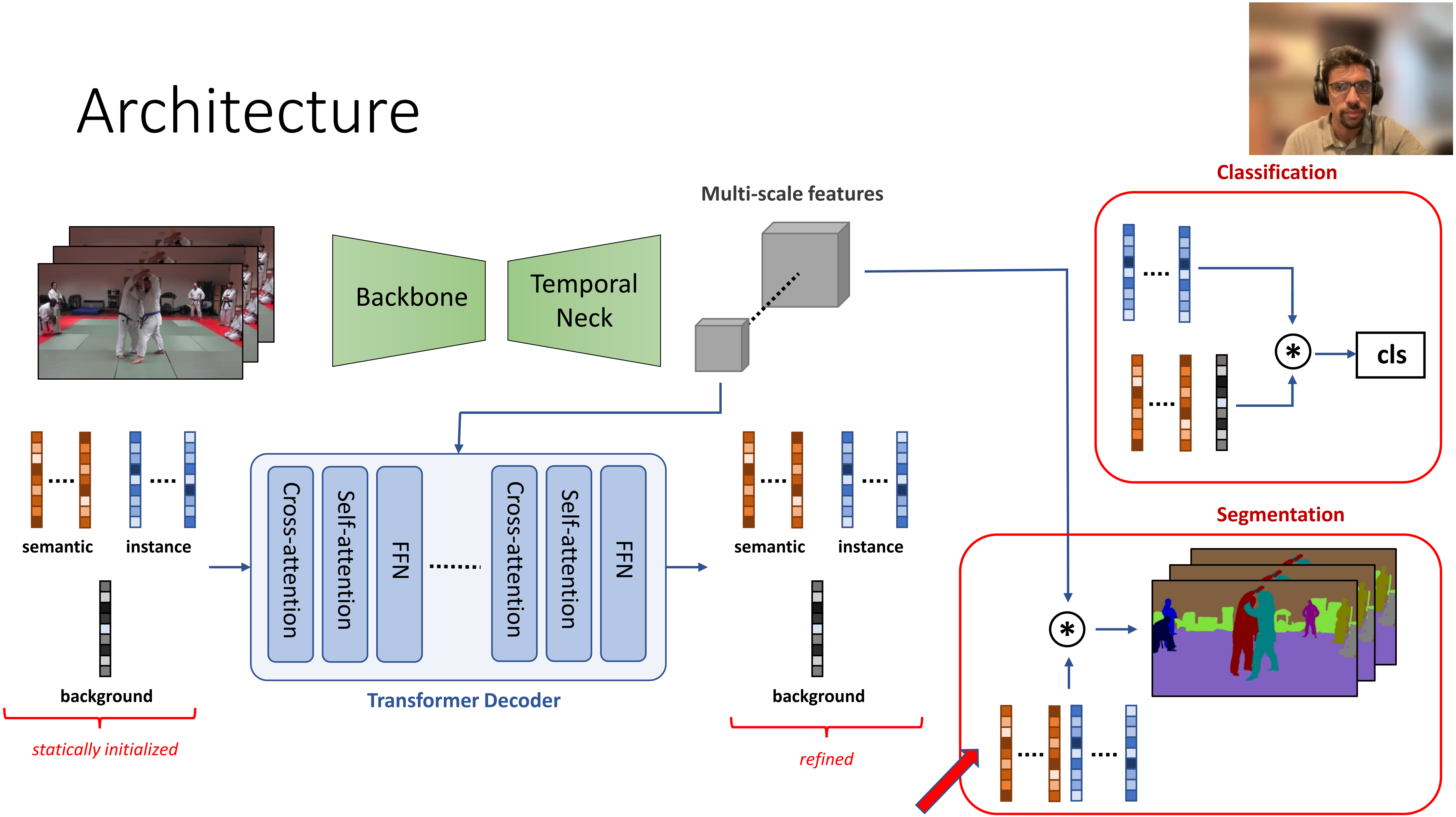
Unified Task Definition

- Architecture is largely task-agnostic
- Encode the task-specific targets as dynamic network inputs (queries)
- Can theoretically tackle any segmentation task

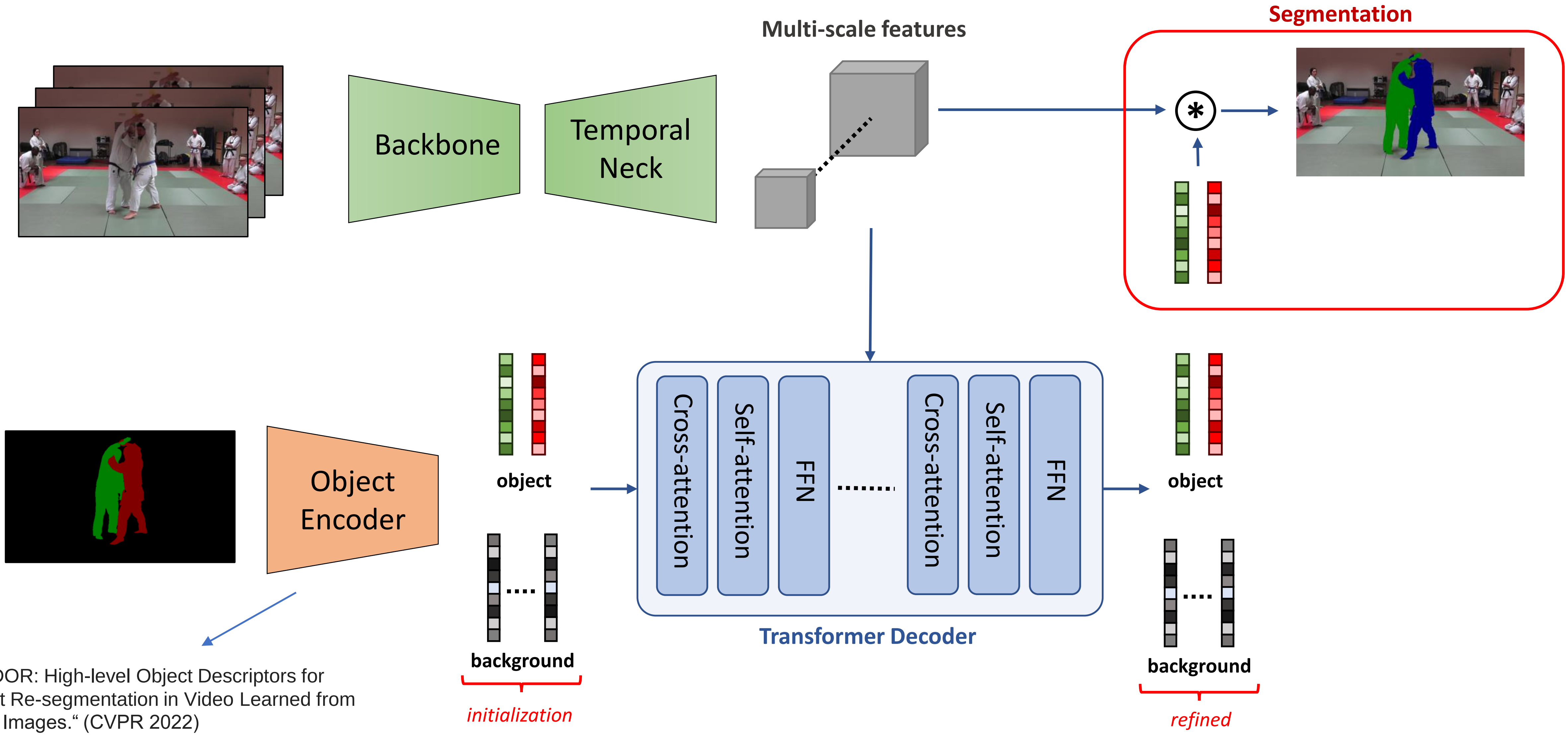
Architecture



Architecture

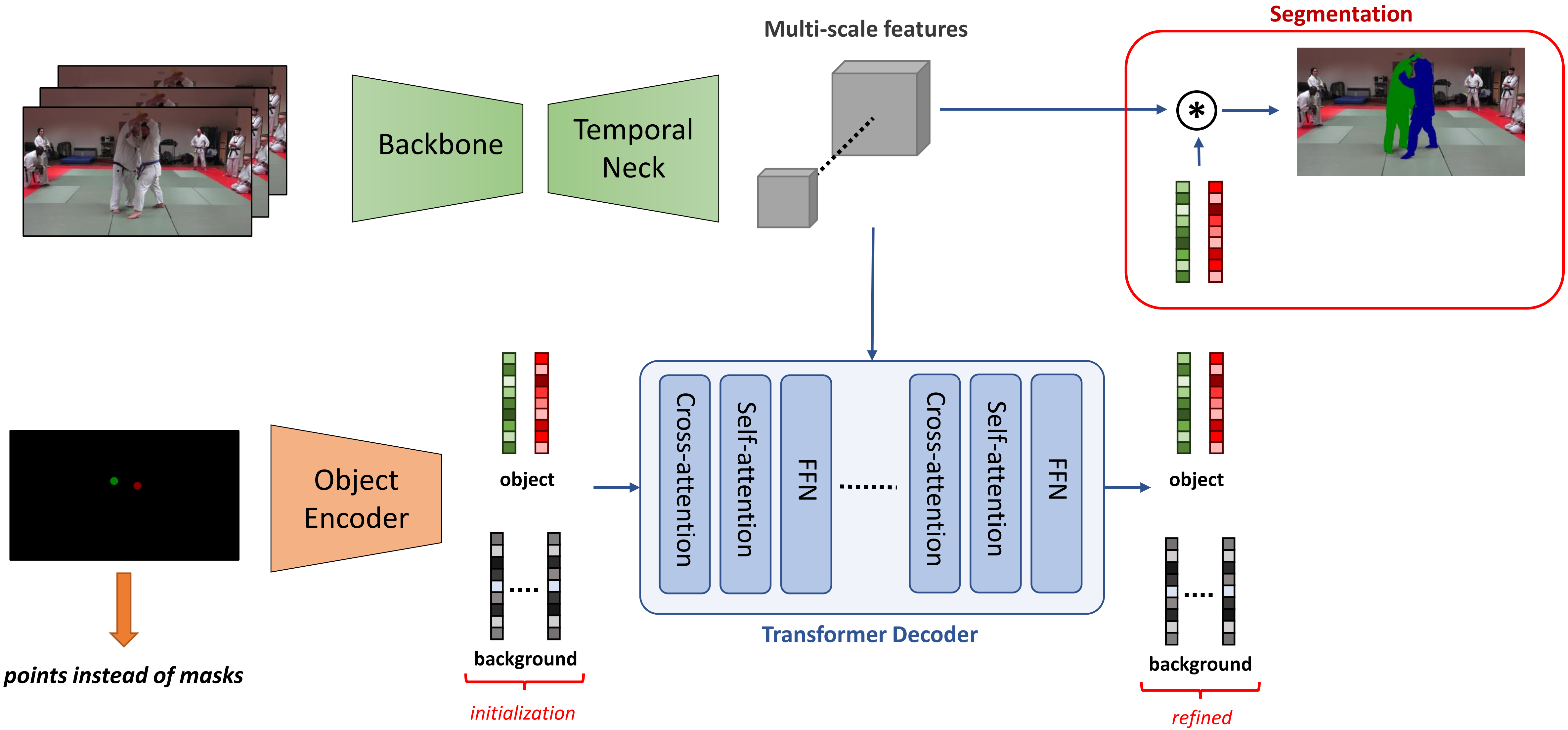


Architecture



"HODOR: High-level Object Descriptors for Object Re-segmentation in Video Learned from Static Images." (CVPR 2022)

Architecture



Temporal Neck

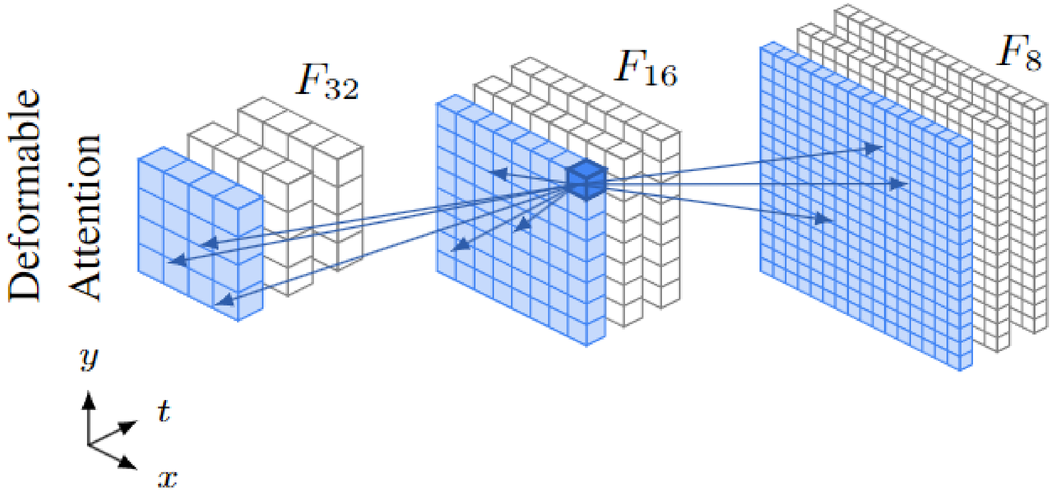
- Masks generated by computing dot-product
- Good mask quality conditioned on consistent video features
- Backbone: Per-image network e.g. ResNet or Swin
- Motivation for temporal neck: incorporate temporal context in video features

Temporal Neck

- Based on deformable deformable attention encoder
- Multi-scale features from backbone are iteratively refined
- Contains 6 layers. Each layer contains two parts:

1. Deformable attention separately within each image frame

2. Self-attention within grid-like cells across all frames

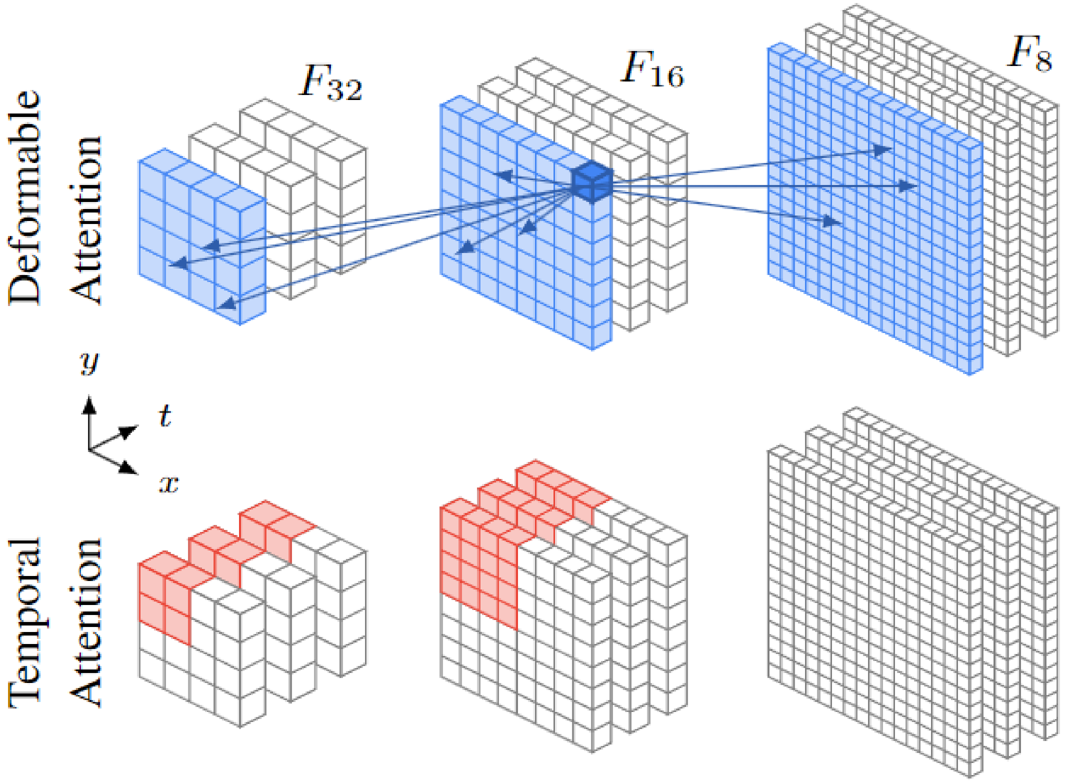


Temporal Neck

- Based on deformable deformable attention encoder
- Multi-scale features from backbone are iteratively refined
- Contains 6 layers. Each layer contains two parts:

1. Deformable attention separately within each image frame

2. Self-attention within grid-like cells across all frames

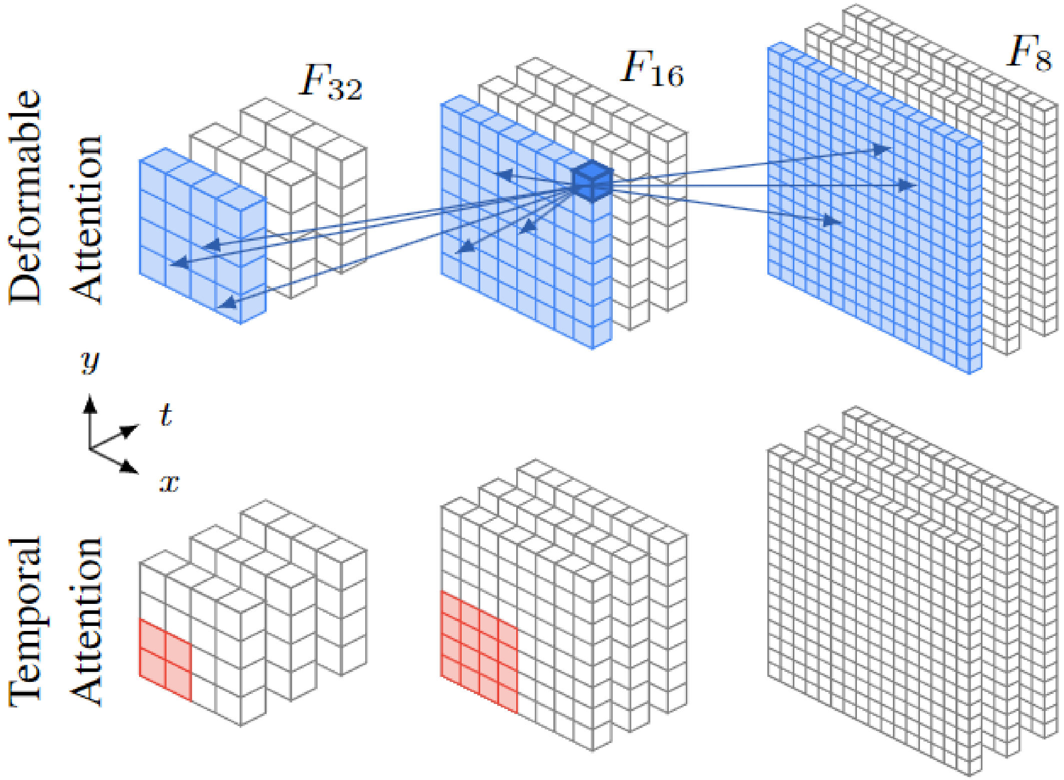


Temporal Neck

- Based on deformable deformable attention encoder
- Multi-scale features from backbone are iteratively refined
- Contains 6 layers. Each layer contains two parts:

1. Deformable attention separately within each image frame

2. Self-attention within grid-like cells across all frames

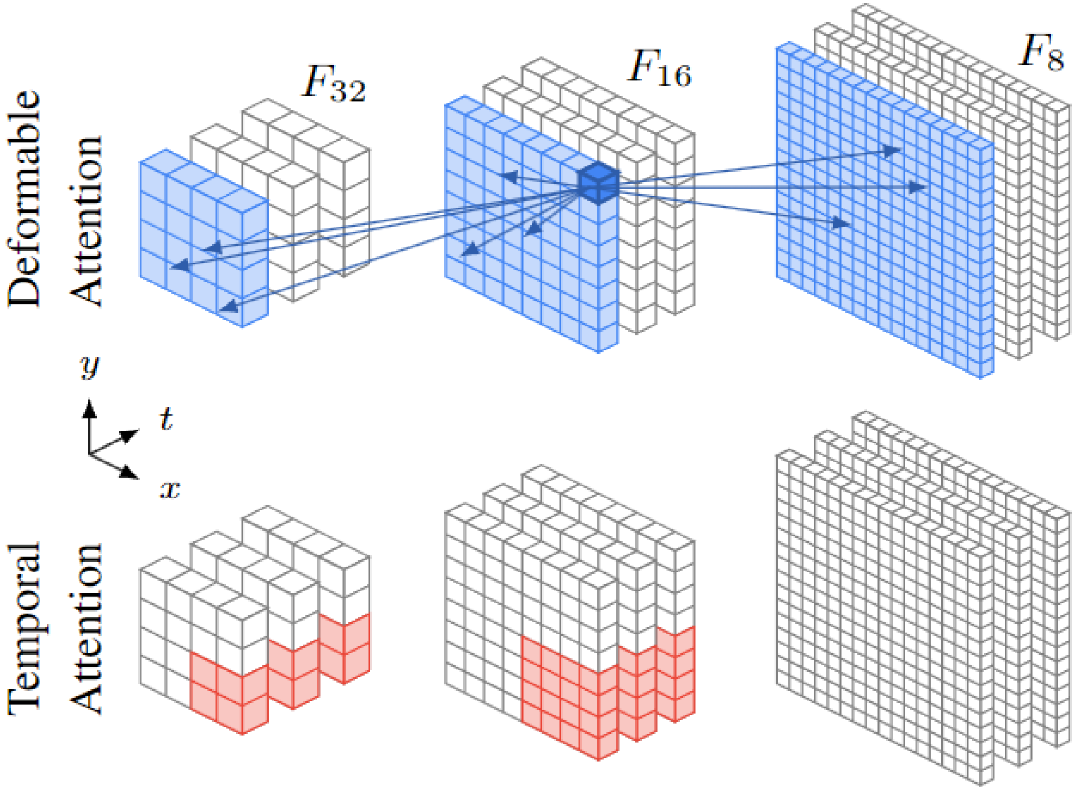


Temporal Neck

- Based on deformable deformable attention encoder
- Multi-scale features from backbone are iteratively refined
- Contains 6 layers. Each layer contains two parts:

1. Deformable attention separately within each image frame

2. Self-attention within grid-like cells across all frames



Benchmark Results

Video Instance Segmentation (VIS)

YouTube-VIS 2021 (val)

Method	AP	AP50	AP75	AR1	AR10
VITA	57.5	80.6	61.0	47.7	62.6
TarViS	60.2	81.4	67.6	47.6	64.8
Difference	+2.7	+0.8	+6.6	-0.1	+1.8

OVIS (val)

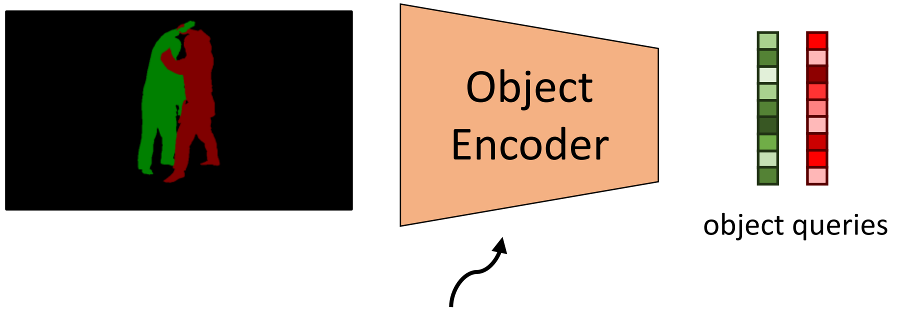
Method	AP	AP50	AP75	AR1	AR10
IDOL	42.6	65.7	45.2	17.9	49.6
TarViS	43.2	67.8	44.6	18.0	50.4
Difference	+0.6	+2.1	-0.8	+0.1	+0.8

Benchmark Results

Video Object Segmentation (VOS)

DAVIS (val)

Method	J&F	J	F
XMem	86.2	82.9	89.5
TarViS	85.3	81.7	88.5
Difference	-0.9	-1.2	-1.0



Benchmark Results

Point Exemplar-guided Tracking (PET)

BURST (val)

Method	$HOTA_{all}$	$HOTA_{com}$	$HOTA_{unc}$
STCN+M2F	24.4	44.0	19.5
TarViS	37.5	51.7	34.0
Difference	+12.9	+7.7	+14.5

BURST (test)

Method	$HOTA_{all}$	$HOTA_{com}$	$HOTA_{unc}$
STCN+M2F	24.9	39.5	22.0
TarViS	36.1	47.1	33.8
Difference	+11.2	+7.6	+11.8

Benchmark Results

Video Panoptic Segmentation (VPS)

KITTI-STEP (val)

Method	STQ	AQ	SQ
Mask Propagation	67.0	63.0	71.0
TarViS	72.0	72.0	73.0
Difference	+5.0	+9.0	+2.0

CityscapesVPS (val)

Method	VPQ	VPQ^{th}	VPQ^{st}
VIP-DeepLab	63.1	49.5	73.0
TarViS	58.9	43.7	69.9
Difference	-4.2	-5.8	-3.1

Benchmark Results

Video Panoptic Segmentation (VPS)

VIPSeg (val)

Method	VPQ	VPQ^{th}	VPQ^{st}	STQ
Clip-PanoFCN	22.9	25.0	20.8	31.5
TarViS	48.0	58.2	39.0	52.9
Difference	+25.1	+33.2	+18.2	+21.4

Qualitative Results (OVIS)

Qualitative Results (KITTI-STEP)

Qualitative Results (DAVIS)

Qualitative Results (BURST)

Conclusion

- TarViS: A unified approach for video segmentation tasks
- Network is task-agnostic: formulate task as queries
- High quality results on 7 benchmarks spanning 4 different tasks
- Pre-trained models + source code available on GitHub

https://github.com/Ali2500/TarViS