Towards Unified and Efficient Pixel-wised Video Perception

Xiangtai Li

2023/8/30

Research Fellow in MMLab@NTU

Understanding Video

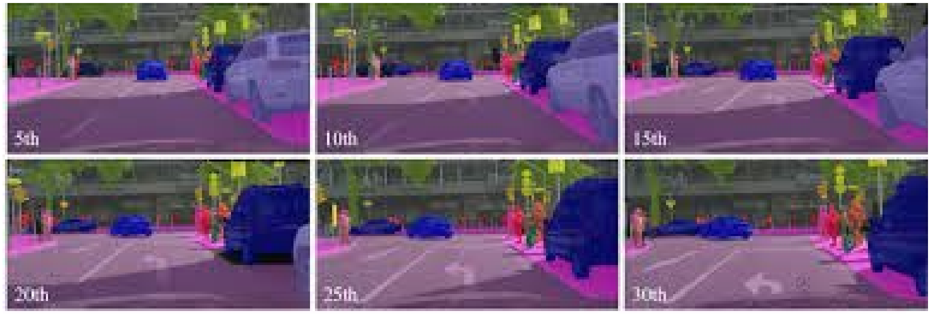
1, Video data are increasing today!

2, Understanding the video contents and mining the instance-wised information are the core computer vision tasks.

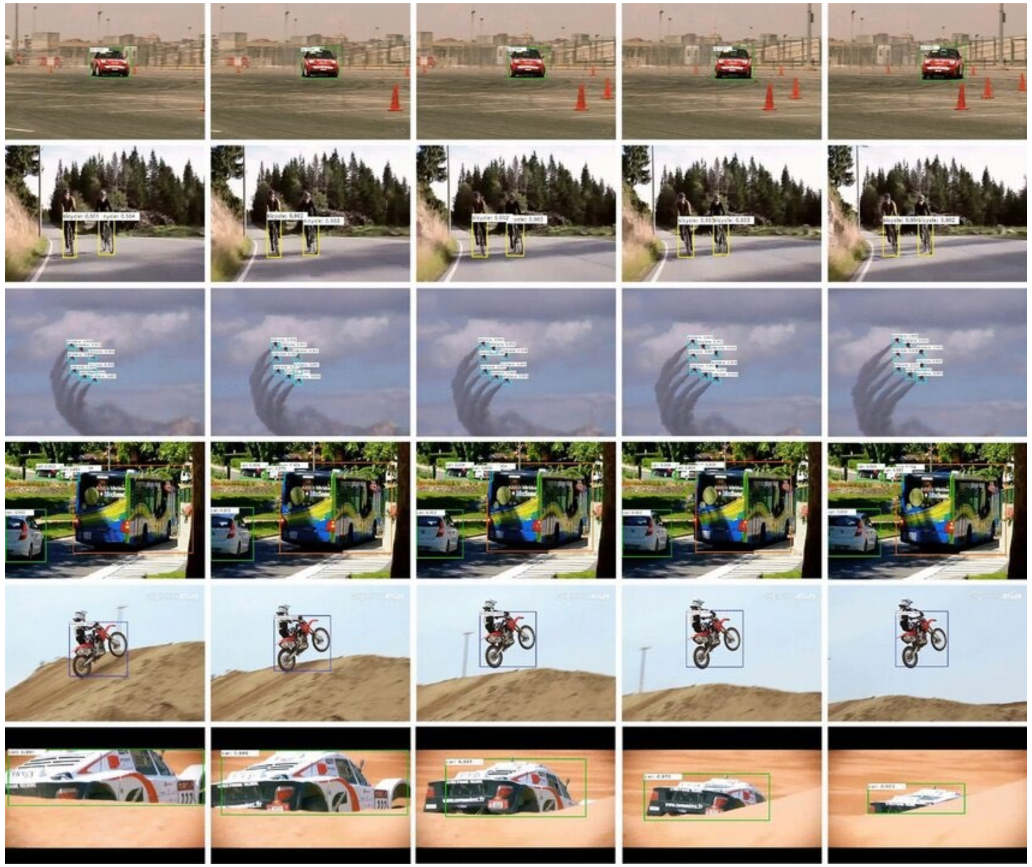
Understanding Video

3, Compare to image understanding task, video tasks have more problems, including motion/occlusion/video consistency.

4, Rather action recognition, We focus on pixel-level video scene understanding tasks, including segmentation, detection and tracking.



Understanding Video



imagenet VID dataset

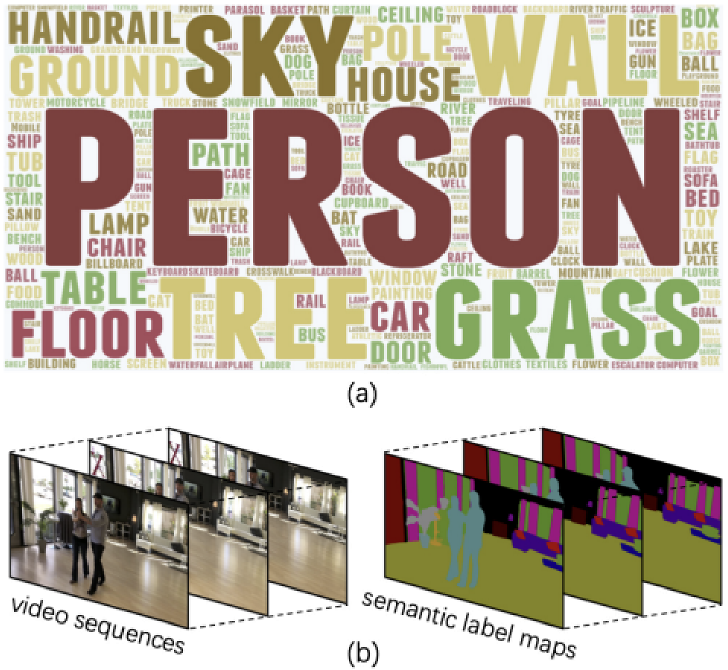
视频目标检测 Video Object Detection:

1. detect each object in the video.

2. explore temporal detection consistency without considering their ID.

Understanding Video

Cityscape dataset



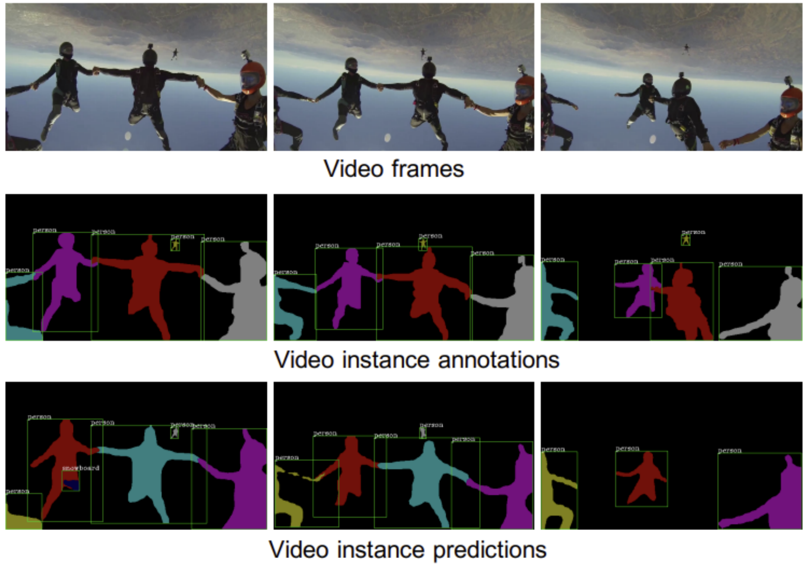
VSPW dataset

视频语义分割 Video Semantic Segmentation:

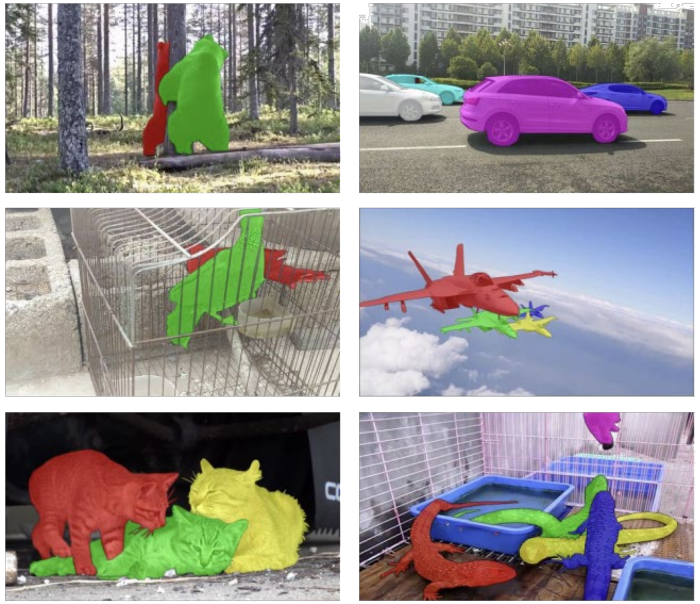
1. classifies each pixel in an image into a certain class along the video.

2. explore the temporal segmentation consistency.

Understanding Video



Youtube-VIS



OVIS

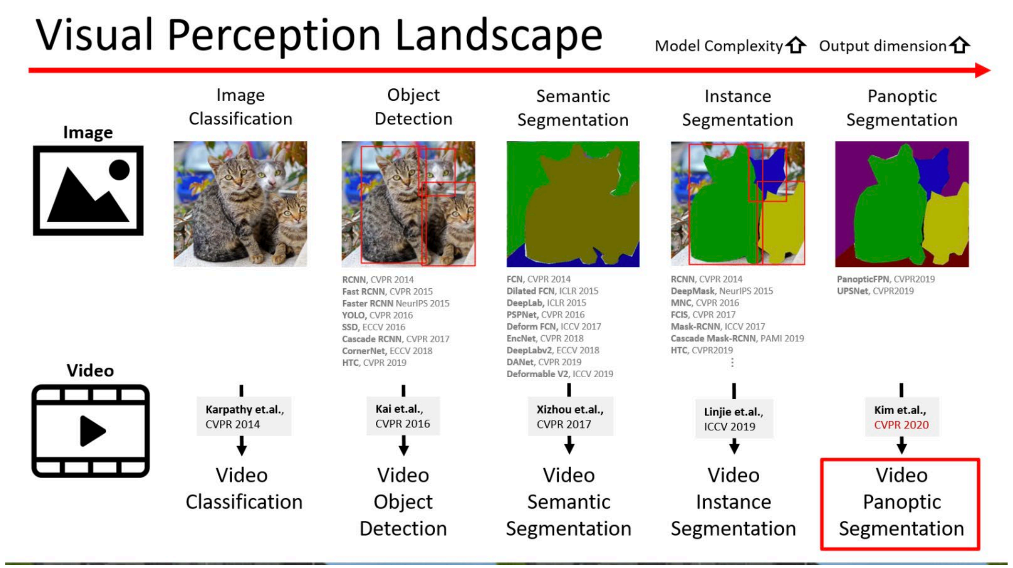
视频实例分割 Video Instance Segmentation:

1. segment and track foreground object in pixel level.

2. explore the temporal consistency and instance ID consistency.

3. more complex than image instance segmentation because it needs to handle object motion, changes in appearance, occlusions, and the temporal consistency of labels across frames.

Understanding Video



VPSNet

KITTI-STEP

Figure 1. Examples of our large-scale VIdeo Panoptic Segmentation in the Wild (VIPSeg) dataset.

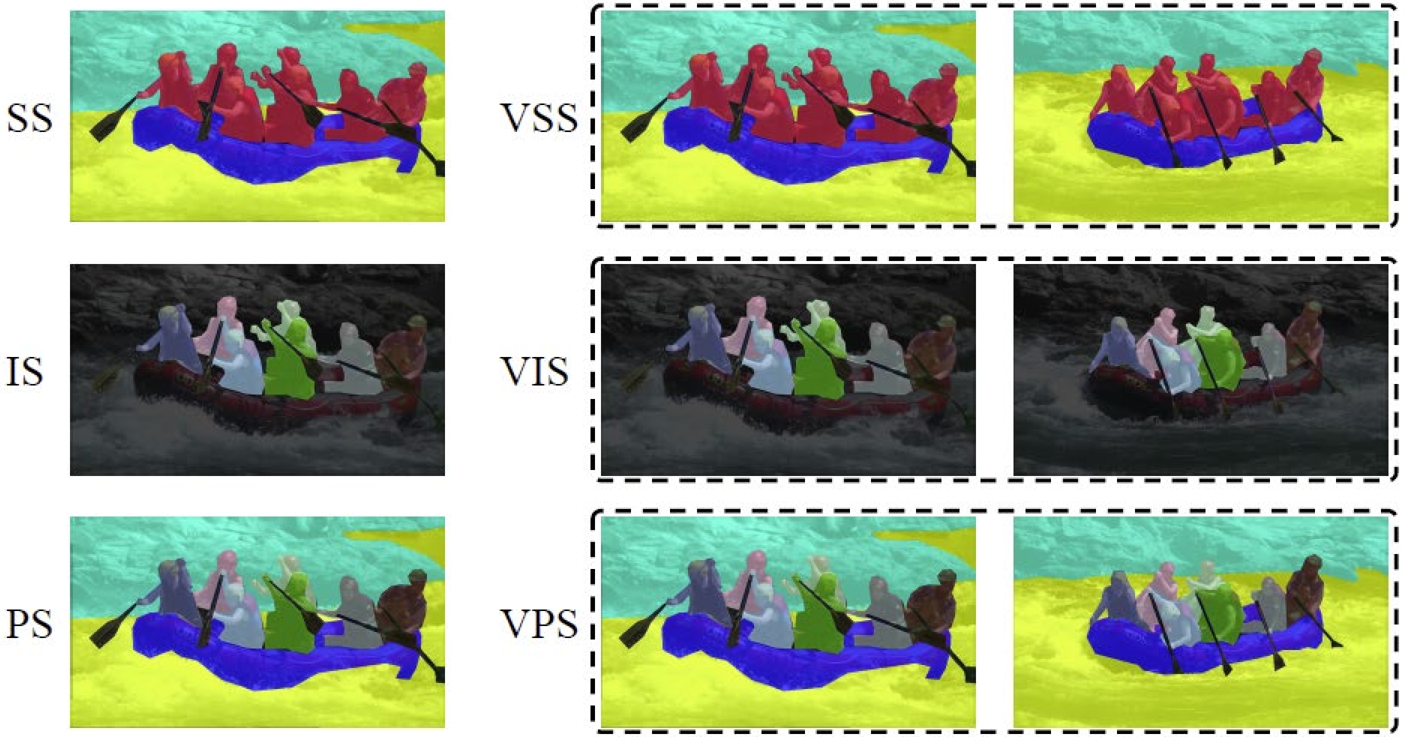
VIP-Seg

视频全景分割 Video Panoptic Segmentation

1. identify and classify every pixel in every frame of a video sequence and maintain the identities of instances (individual objects) across the different frames.

2. is a complex task due to the need to handle motion, changes in appearance, occlusions, and maintaining temporal consistency of labels across frames.

Understanding Video



(a) Image Segmentation (b) Video Segmentation

Semantic Segmentation -> Instance Segmentation -> Panoptic Segmentation.

SS -> IS -> PS

Video Segmentation -> Video Instance Segmentation -> Video Panoptic Segmentation

VSS -> VIS -> VPS

1. Transformer-Based Visual Segmentation: A Survey, arxiv, 2023.
2. Largescale video panoptic segmentation in the wild: A benchmark, CVPR-2022.

Four Research Works

1, **TransVOD**: End-to-End Video Object Detection with Spatial-Temporal Transformers. (Video Object Detection), TPAMI-2022

2, **PolyphonicFormer** : Unified Query Learning for Depth-aware Video Panoptic Segmentation, ECCV-2022

3, **Video K-Net**: A Simple, Strong, and Unified Baseline for Video Segmentation (Video Panoptic Segmentation, online), CVPR-2022

4, **Tube-Link**: A Flexible Cross Tube Baseline for Universal Video Segmentation

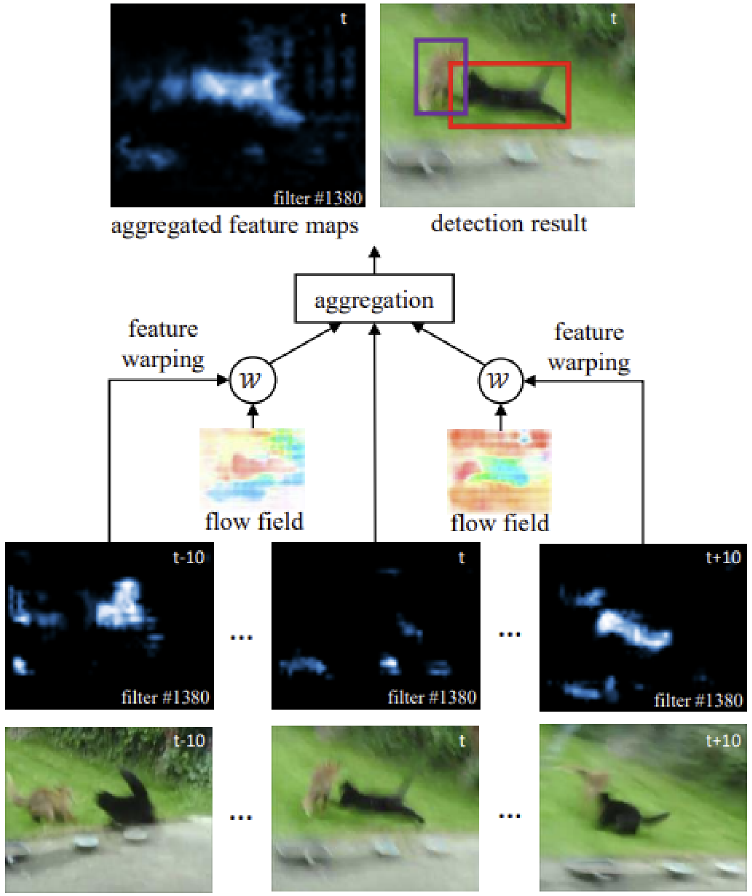
(Universal Video Segmentation, semi-online), ICCV-2023

TransVOD:End-to-End Video Object Detection with Spatial-Temporal Transformers

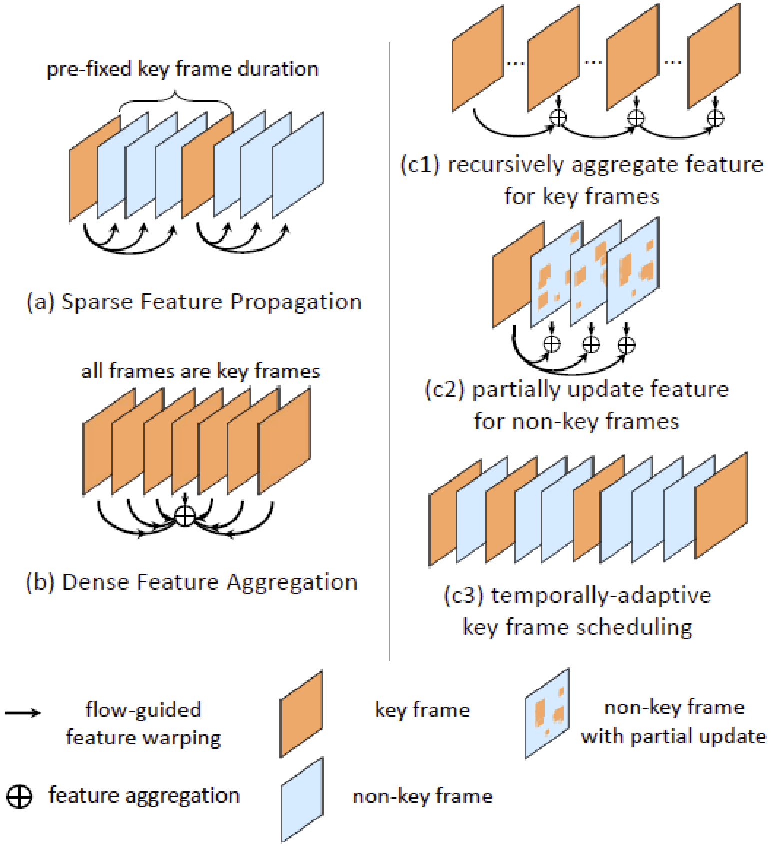
Qianyu Zhou[1*], **Xiangtai Li**[2*(project leader)], Lu He[1*], Yibo Yang[3], Guangliang Cheng[4], Yunhai Tong[2], Shouhong Ding[1], Lizhuang Ma[1], Dacheng Tao

1. Shanghai Jiao Tong University, China; 2. Peking University, China;
3. Sensetime Research, China 4. JD Explore Academy, China;

Previous VOD Methods

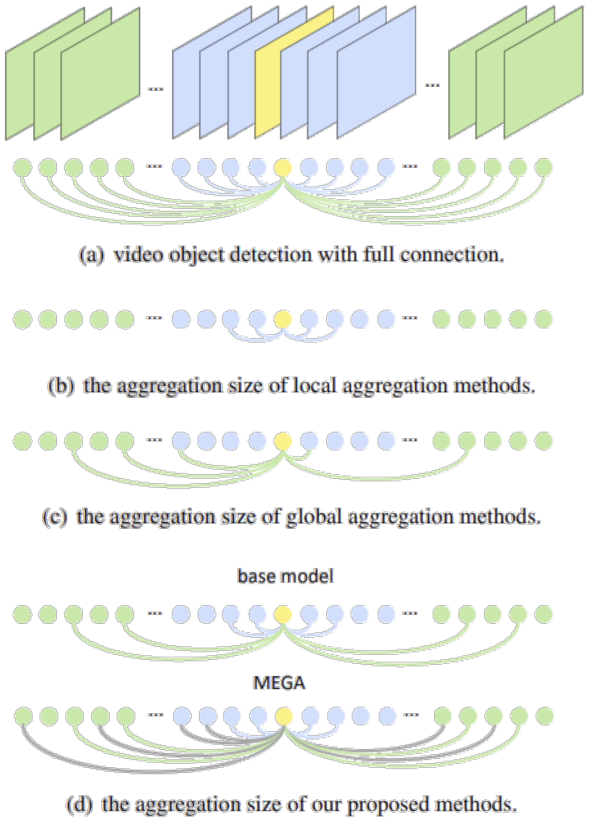


FGFA (ICCV-2017)

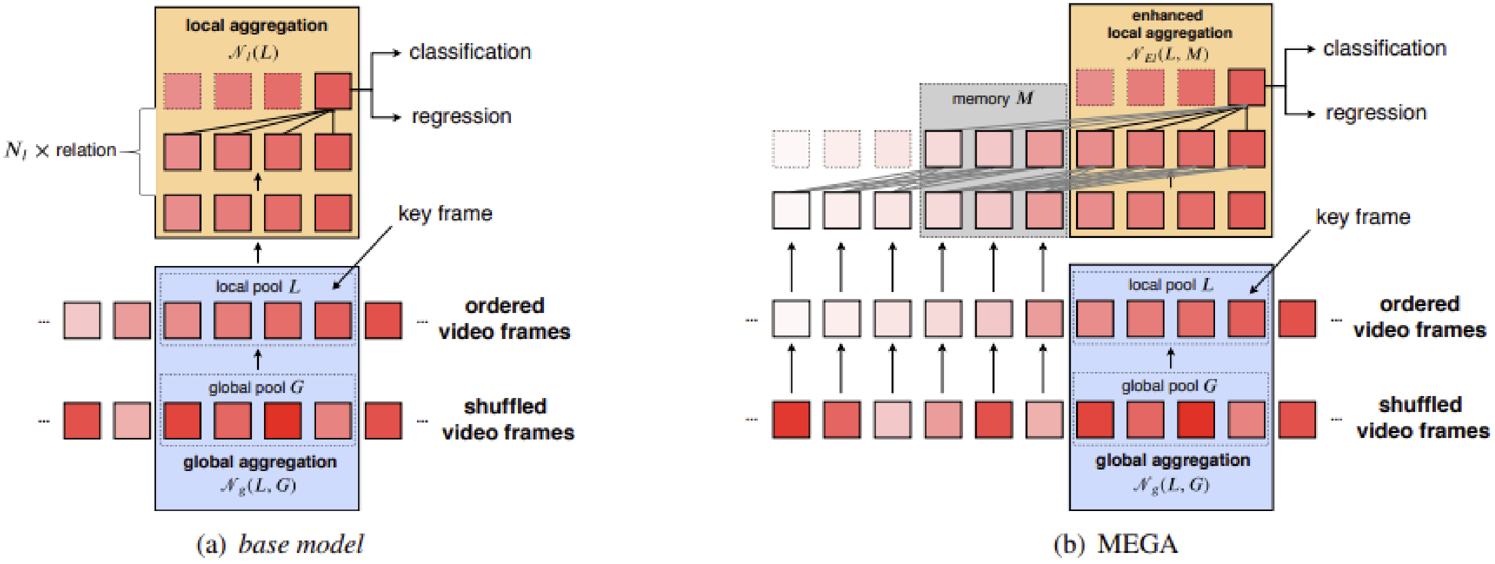


High Performance VOD (CVPR-2018)

Previous VOD Methods



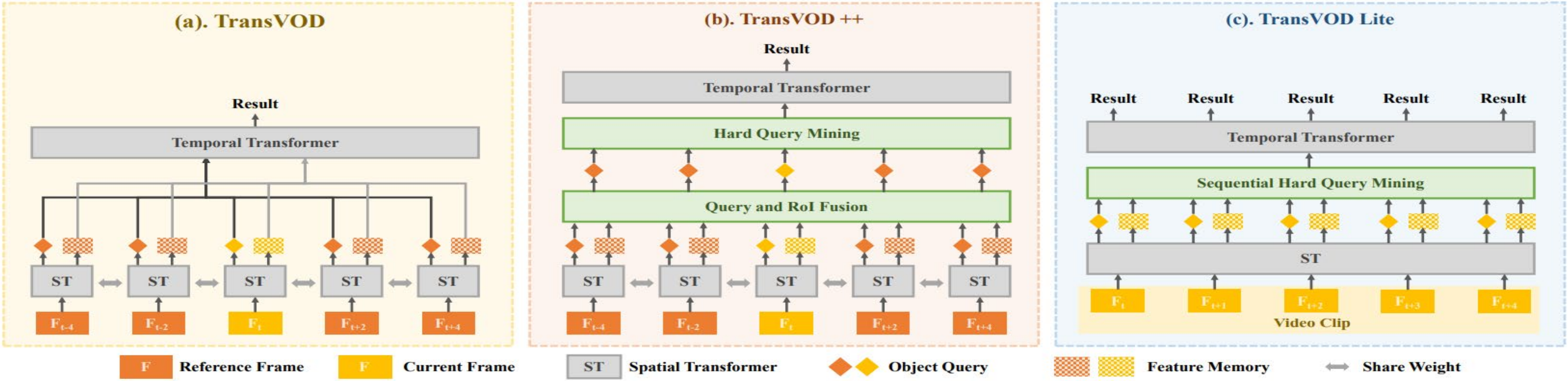
MEGA (CVPR-2020)



MEGA (CVPR-2020)

Motivation

TransVOD, TransVOD++, TransVOD Lite (TPAMI-2023)



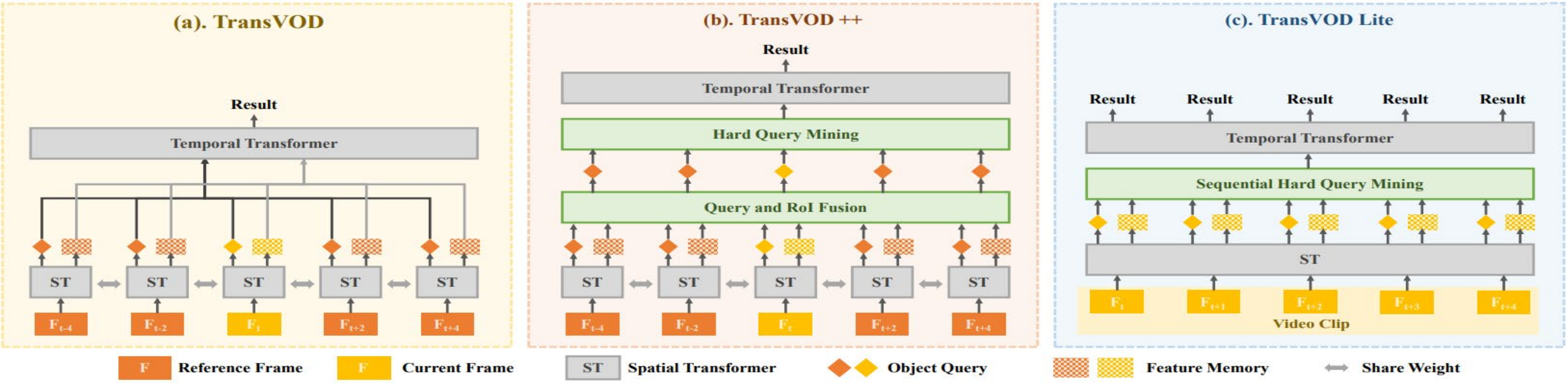
Motivation and Goals:

1, Current solutions for VOD contains multiple components, including sequential NMS and multiple frame fusing.

2, Extending simple DTER-like detectors to the video domain is necessary.

Motivation

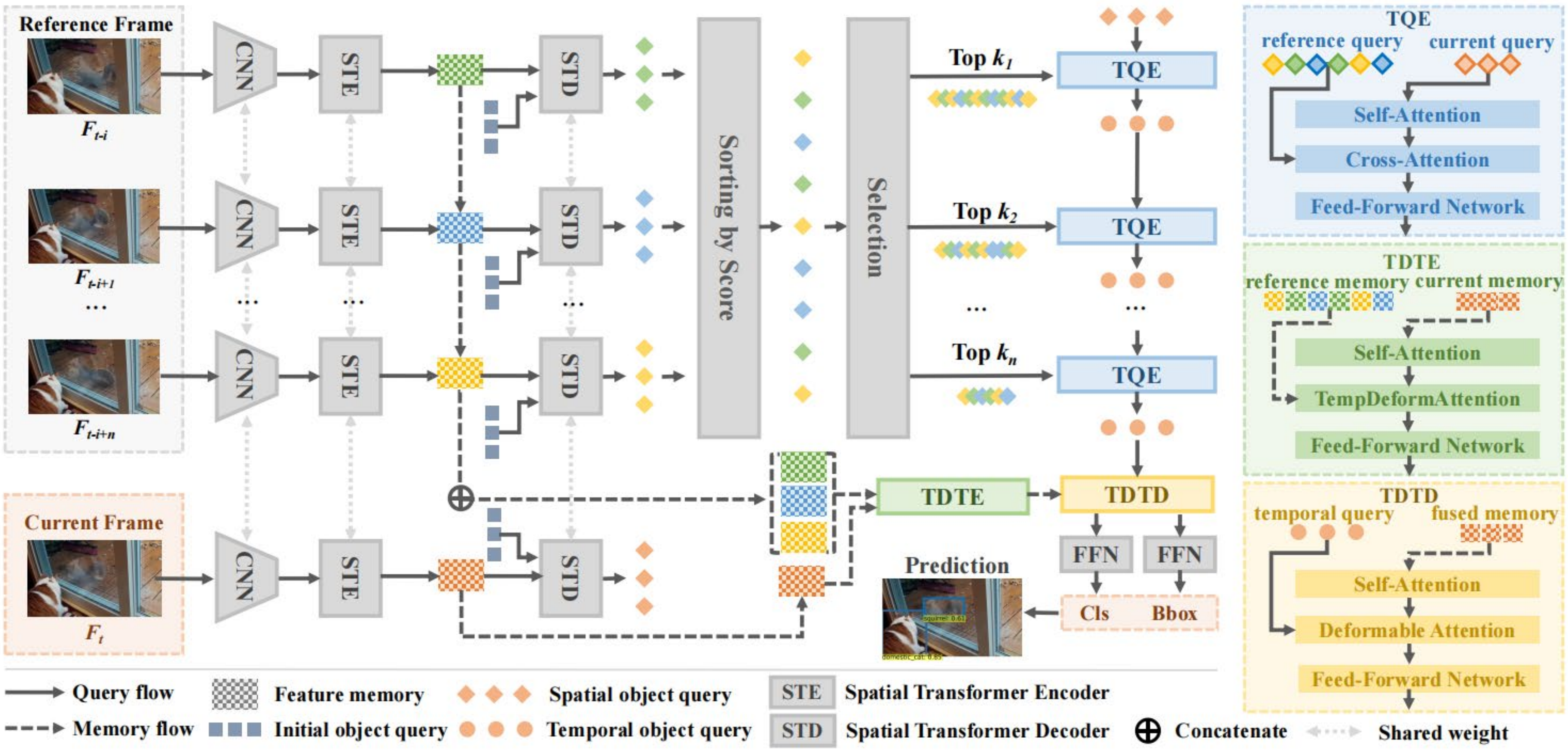
TransVOD, TransVOD++, TransVOD Lite (TPAMI-2023)



3, Streamline the pipeline of VOD based on spatial-temporal Transformers.

4, View VOD as a sequence-to-sequence task with Transformers.

Method



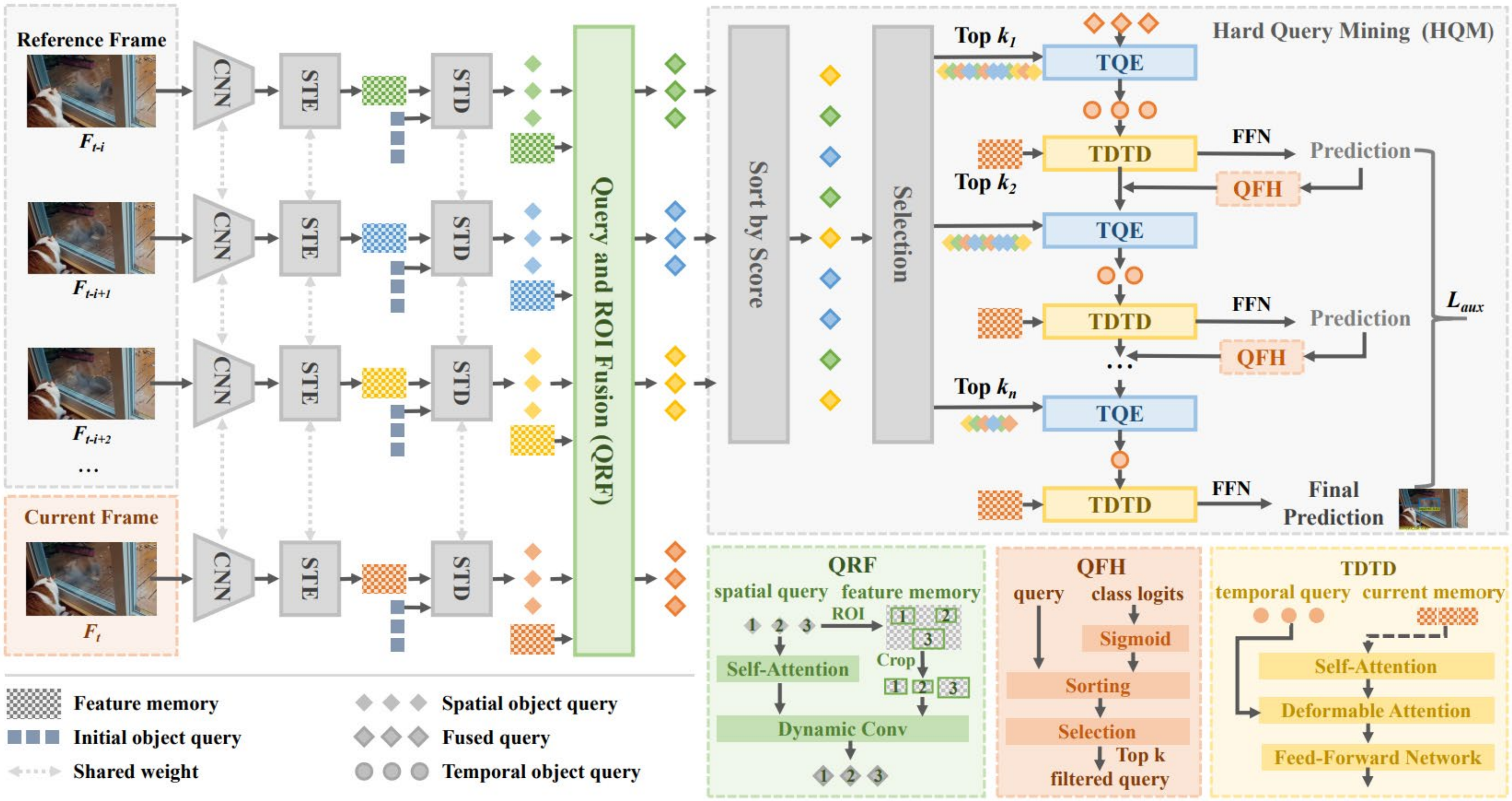
TransVOD:

1, Temporal Query Enocder: Encode temporal object information in the encoder.

2, Temporal Deformable Transformer Encoder: Encode feature level information in the encoder.

3, Temporal Deformable Transformer Decoder: Fuse the multiple object in the decoder.

Method

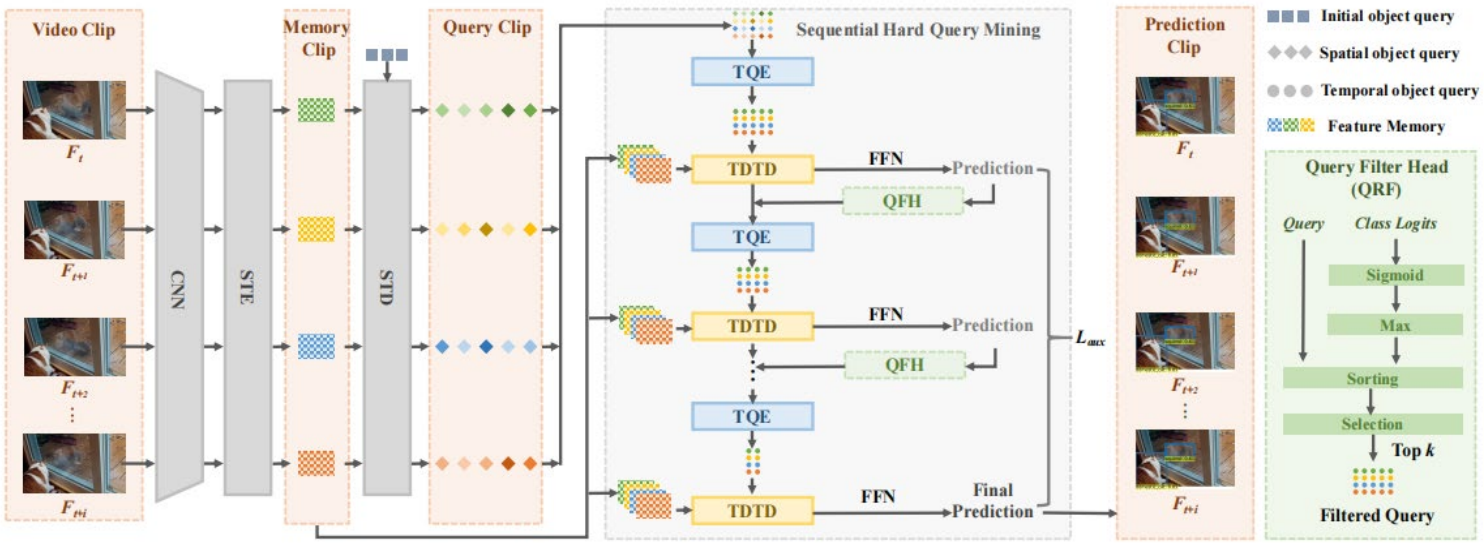


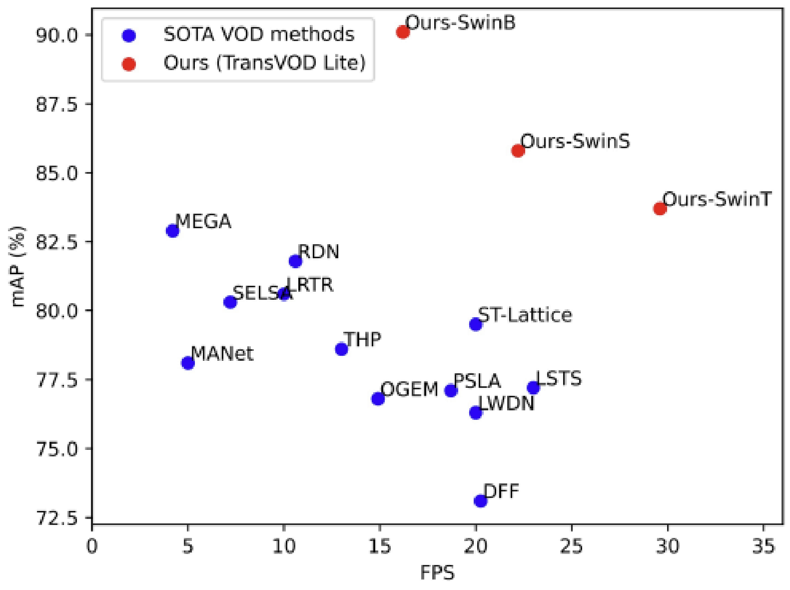
TransVOD++:

1, Query and RoI Fusion. Fuse object-level information into object query

2, Hard Query Mining: Mining the hardest query via auxiliary loss in each TDTD.

Method





Video object detection (VOD) results on ImageNet VID

TransVOD Lite:

1, Direct Multiple Frame Prediction: take multiple frames as inputs and obtain multiple frame results simultaneously in a temporal window.

2, Sequential Hard Query Mining: to mine the hardest query for a video clip.

Experiment Results

1, Considerable performance using ResNet50.

2, The first method achieves over 90% mAP on ImageNet-VID dataset.

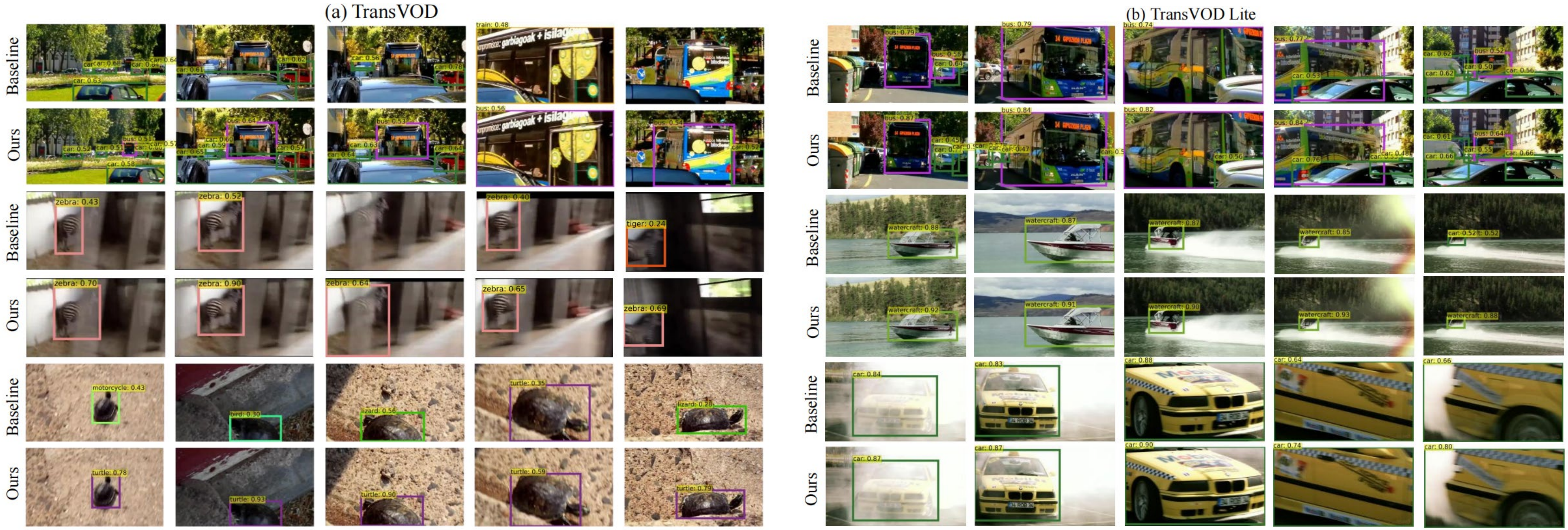
Methods	Base Detector	mAP (%)
Single Frame Baseline [1]	Faster-RCNN	71.8
DFF [19]	Faster-RCNN	70.4
FGFA [20]	Faster-RCNN	74.0
RDN [27]	Faster-RCNN	76.7
MEGA [16]	Faster-RCNN	77.3
Single Frame Baseline† [1]	Faster-RCNN†	72.7
DFF† [19]	Faster-RCNN†	71.6
FGFA† [20]	Faster-RCNN†	75.1
RDN† [27]	Faster-RCNN†	77.6
MEGA† [16]	Faster-RCNN†	78.3
Single Frame Baseline [34]	Deformable DETR	76.0
TransVOD	Deformable DETR	**79.9**
TransVOD++	Deformable DETR	**80.5**

Single Frame Baseline [1]	Faster RCNN	76.7
ST-Lattice [22]	Faster RCNN	79.0
BFAN [70]	Faster RCNN	79.1
STCA [71]	Faster RCNN	80.3
SELSA [49]	Faster RCNN	80.3
MINet [72]	Faster RCNN	80.6
LRTR [35]	Faster RCNN	81.0
RDN [34]	Faster RCNN	81.8
TROI [73]	Faster RCNN	82.0
MEGA [23]	Faster RCNN	82.9
HVRNet [18]	Faster RCNN	83.2
TF-Blender [74]	Faster RCNN	83.8
DSFNet [20]	Faster RCNN	84.1
MAMBA [24]	Faster RCNN	84.6
EBFA [19]	Faster RCNN	84.8
CFA-Net [75]	Faster RCNN	85.0
Single Frame Baseline [76]	CenterNet	73.6
CHP [77]	CenterNet	76.7
Single Frame Baseline [11]	Deformable DETR	78.3
TransVOD Lite	Deformable DETR	80.5
TransVOD++	Deformable DETR	82.0
TransVOD++*	Deformable DETR	**90.0**

Experiment Results

1, Perform well on the cases with motion blur, occlusion.

2, Fast inference on GPU. Detect the missing objects in the video.



What are the Nexts?

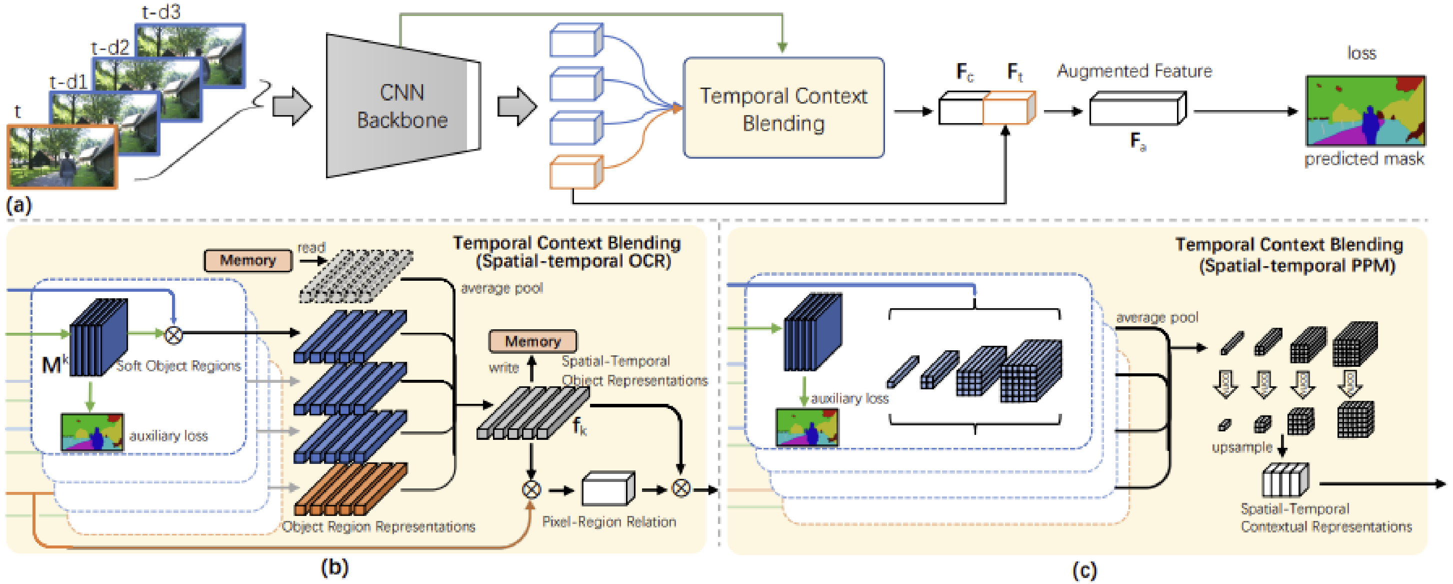
1, Boxes are coarse representations of objects.

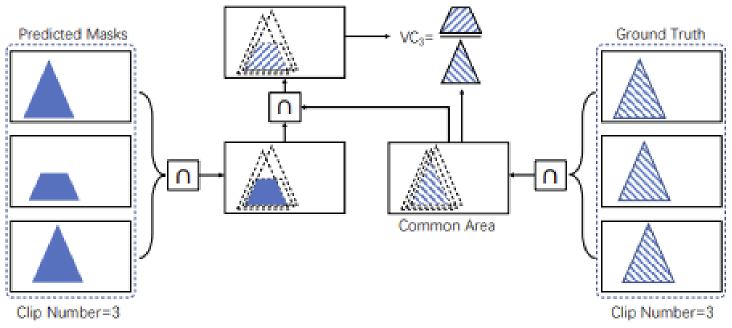
2, Missing background context.

3, ImageNet VID datasets are less challenging with introduce of vision transformer.

Next, we focus on video segmentation and tracking tasks are more challenging.

VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild CVPR-2021





Video Consistency Metric

TCBNet:

1. Focus on spatial-Temporal global context modeling.

2. Design two different temporal fusion methods: Spatial-temporal OCR and Spatial-temporal PPM from two different image segmentation baseline, OCR-Net and PSPNet.

Coarse-to-Fine Feature Mining for Video Semantic Segmentation, CVPR-2022

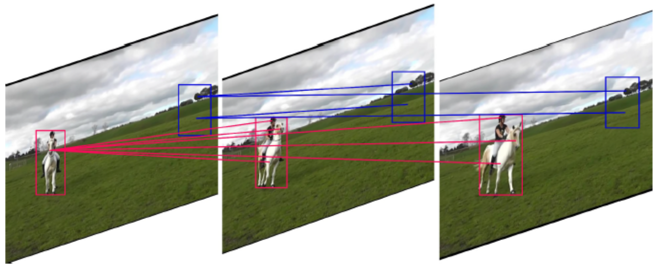
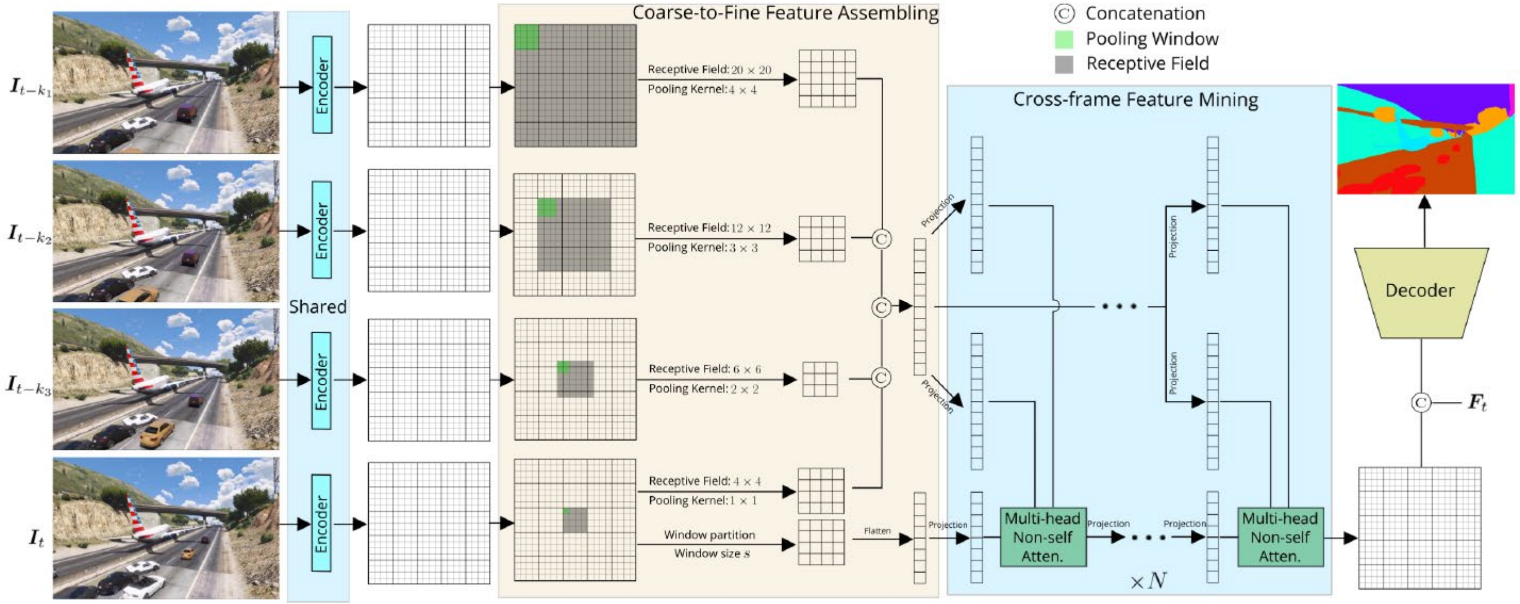


Figure 1. Illustration of *static contexts* (in blue) and *motional contexts* (in red) across neighbouring video frames. The human and horse are moving objects, while the grassland and sky are static background. Note that the static stuff is helpful for the recognition of moving objects, *i.e.*, a human is riding a horse on the grassland.



CFFM:

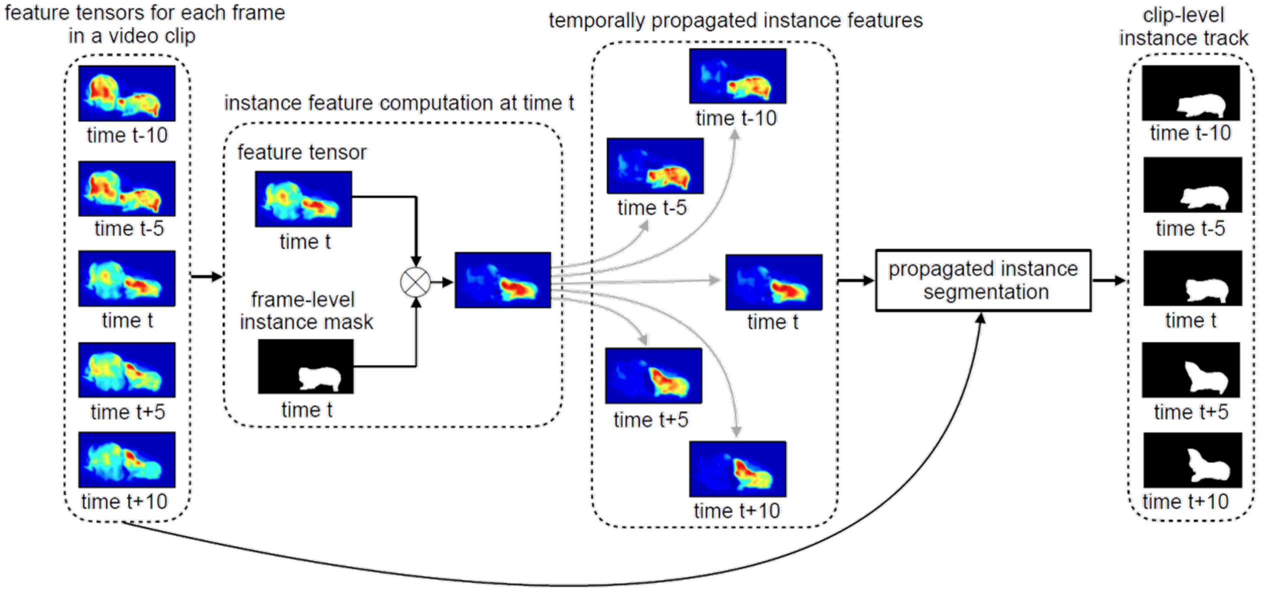
1. Coarse-to-Fine Feature Assembling: assemble local frame features via different kernels.

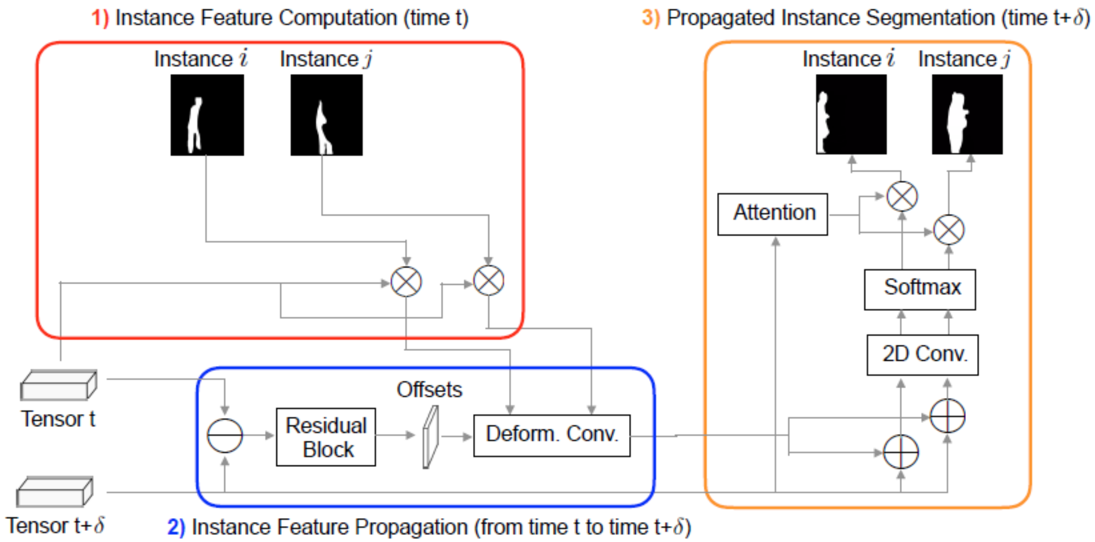
2. Cross-Frame Feature Mining: fuse the cross frame features via multi-head non-self attention.

Classifying, Segmenting, and Tracking Object Instances in Video with Mask Propagation

CVPR-2020

Gedas Bertasius, Lorenzo Torresani
Facebook AI



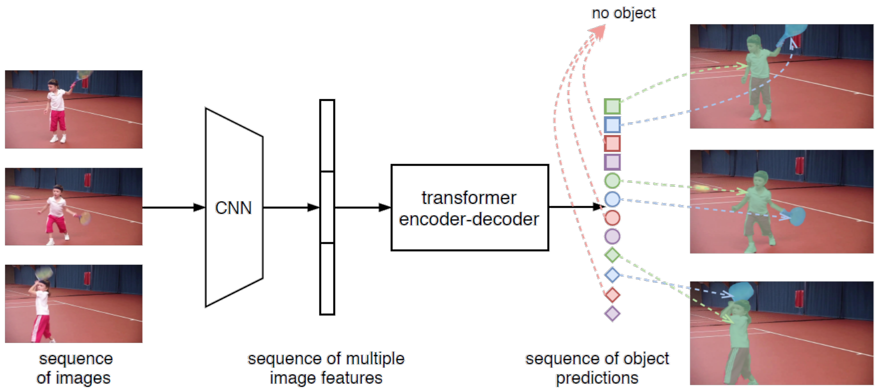


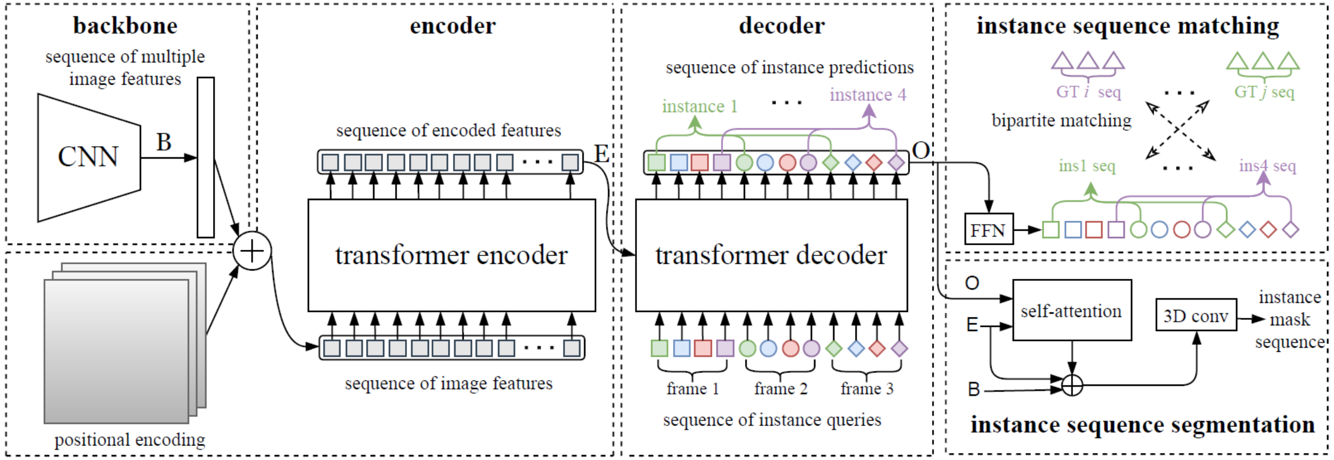
Mask Propagation:

1, use DCN + instance-wised attention to fuse instance-wised feature in each frame.

2, Design the High Resolution Refinement to generate fine-grained masks in each frame.

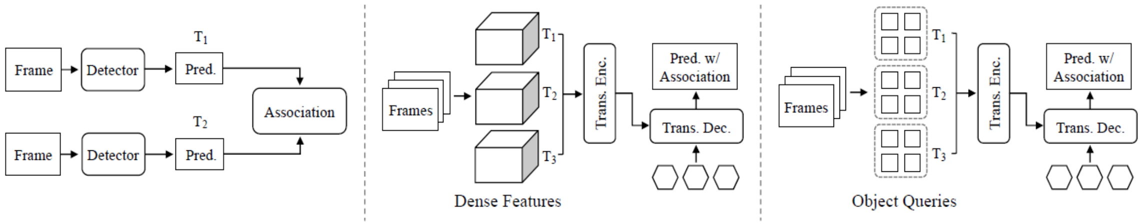
End-to-End Video Instance Segmentation with Transformers, CVPR-2021





1. First transformer-based video instance segmentation.
2. Treat each instance as tracked query. No extra tracking process is needed.
3. Use 3D conv and DCN for post-process each tube instance masks.

Vita: Video instance segmentation via object token association, NeurIPS-2022



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

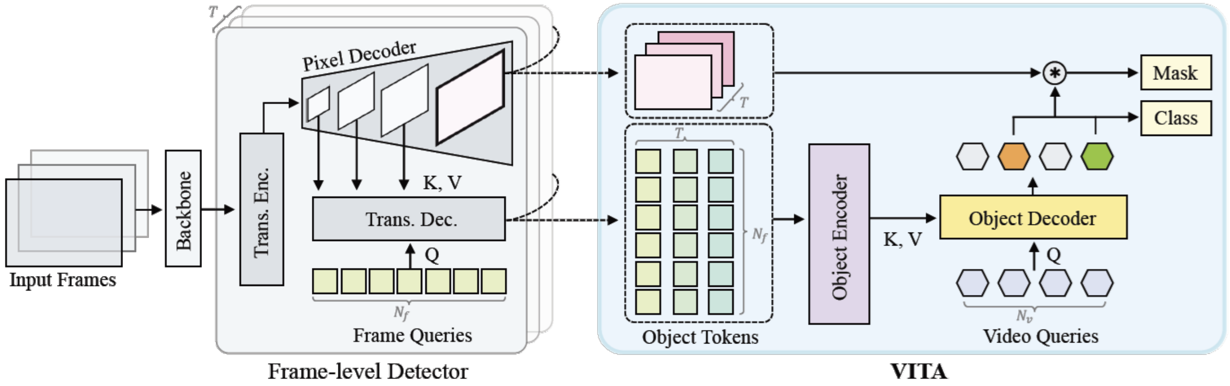


Figure 2: VITA takes only mask features and frame queries that are independently decoded by the frame-level detector for entire video sequence. By directly constructing temporal interactions between frame queries that encapsulate rich object-aware knowledge in spatial scenes, VITA yields mask trajectories with corresponding categories in an end-to-end manner.

1, An extra offline fusion method to fuse image segmenter results in each frame.

2, Introduce the video to associate and fuse object tokens from each frame.

3, VITA plays as post-process for VIS.

Video Panoptic Segmentation CVPR-2020

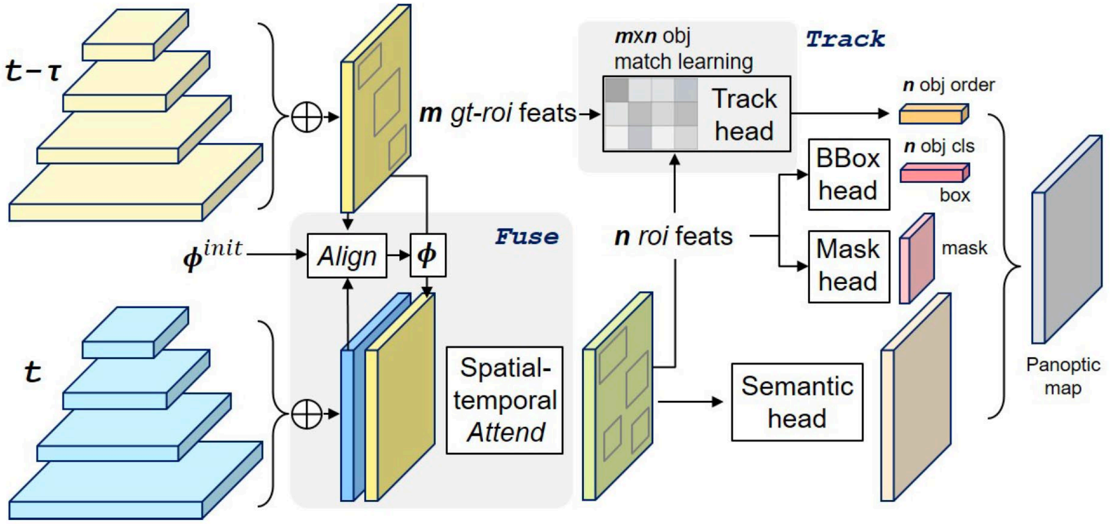


Figure 3: **Overall architecture of our VPSNet.**

stoa baseline:

UPSNet + masktrack rcnn head
+ Balanced FPN

Fuse at Pixel Level

Extra Neck:
Align and Atten across the balanced feature pyramid

1, Align: Flownet2 to warp feature
and refine via deep feature flow (inner lite flow net)

2, Atten:
spatial-temporal attention to reweight the features
(high computation cost)

Track at Object Level

Training:
Difference: ROI features are enhanced via temporal fusion module.

Inference:
additional cue from the panoptic head: the IoU of things logits.

CVPR-2022 (oral)

Video K-Net: A Simple, Strong, and Unified Baseline for Video Segmentation

[1]Peking University, [2]S-Lab, Nanyang Technological University, [3]The Chinese University of Hong Kong, [4]SenseTime Research, [5]Shanghai AI Laboratory

Motivation

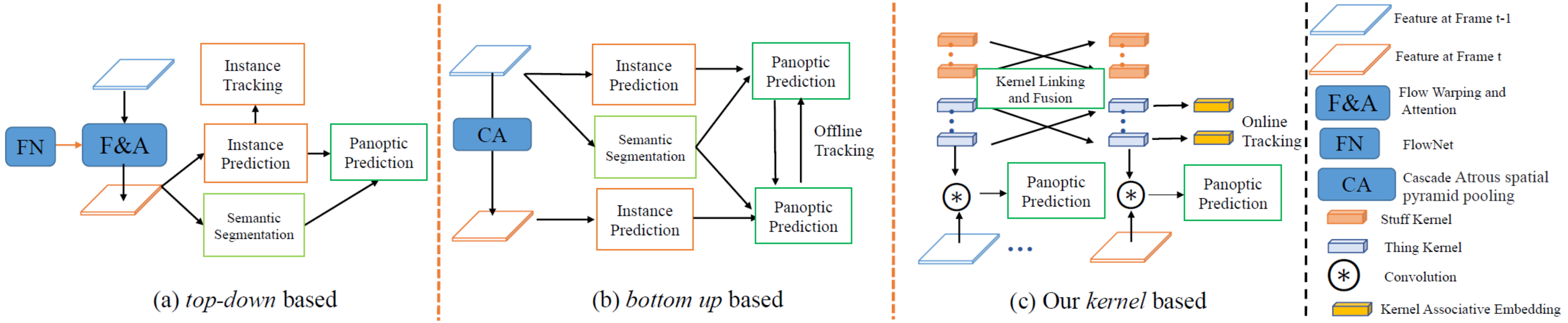


Fig.1 Current Solution For Video Panoptic Segmentation (VPS)

The problems of current Video Panoptic Segmentation (VPS):

1. Complex and hand-crafted Pipeline for VPS.
2. Need the post process or offline tracking.
3. Need optical flow learning and warping.
4. Tackle the segmentation and tracking with specific task head.

Is there any simpler solution for VPS?

Fig.2 Current Video Segmentation Tasks. (Fig from A Survey on Deep Learning Technique for Video Segmentation)

The problems of other Video Segmentation Tasks:

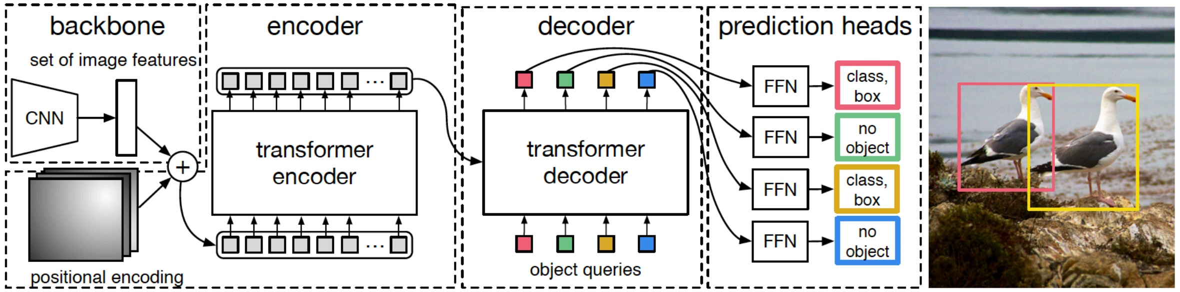
1. Different Tasks have different solutions including specific design. Such as Optical Flow Warping or Clip-Level Transformer.

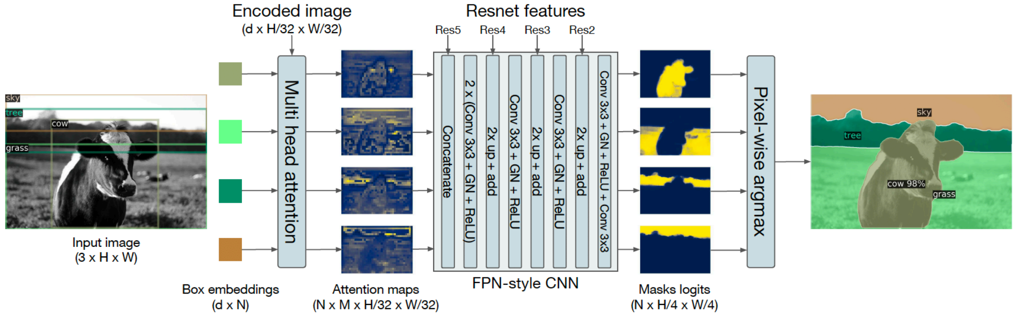
2. Video Semantic Segmentation (VSS): no instance tracking.

3. Video Instance Segmentation (VIS): no background context.

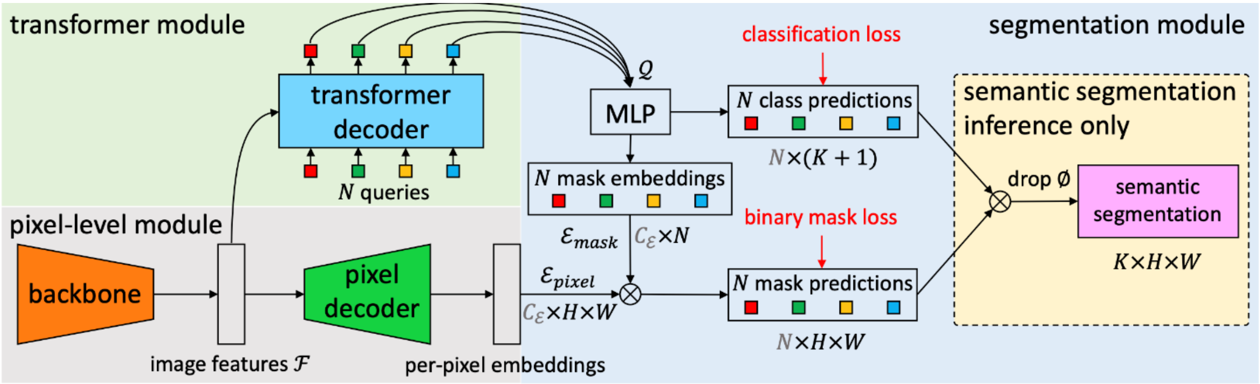
Is there one general model to solve all video segmentation tasks including VPS, VIS and VSS?

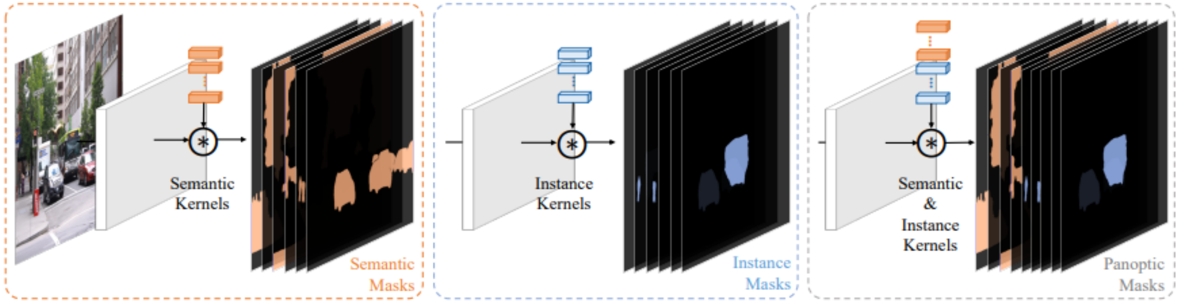
MaskFormer and K-Net Unified Segmentation Framework.





DETR (ECCV-2020)





(a) FCN for Semantic Segmentation Framework (b) K-Net for Instance Segmentation (c) K-Net for Panoptic Segmentation

Why K-Net?

1. More Computation efficiency.
2. Faster Convergence.
3. Stronger Results.

Revisiting K-Net

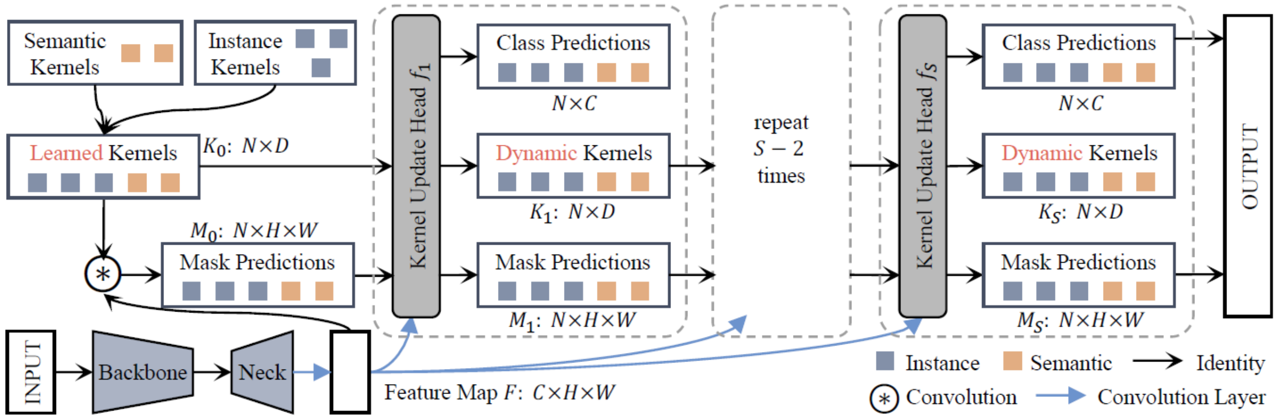


Figure 3: **K-Net for panoptic segmentaion.** A set of learned kernels first performs convolution with the feature map F to predict masks M_0. Then the kernel update head takes the mask predictions M_0, learned kernels K_0, and feature map F as input and produce class predictions, group-aware (dynamic) kernels, and mask predictions. The produced mask prediction, dynamic kernels, and feature map F are sent to the next kernel update head. This process is performed iteratively to progressively refine the kernels and the mask predictions.

1, Universal segmenter.

2, Use learned kernel for updating corresponding features.

3, Propose the kernel update head to dynamically update learned kernel.

Revisiting K-Net

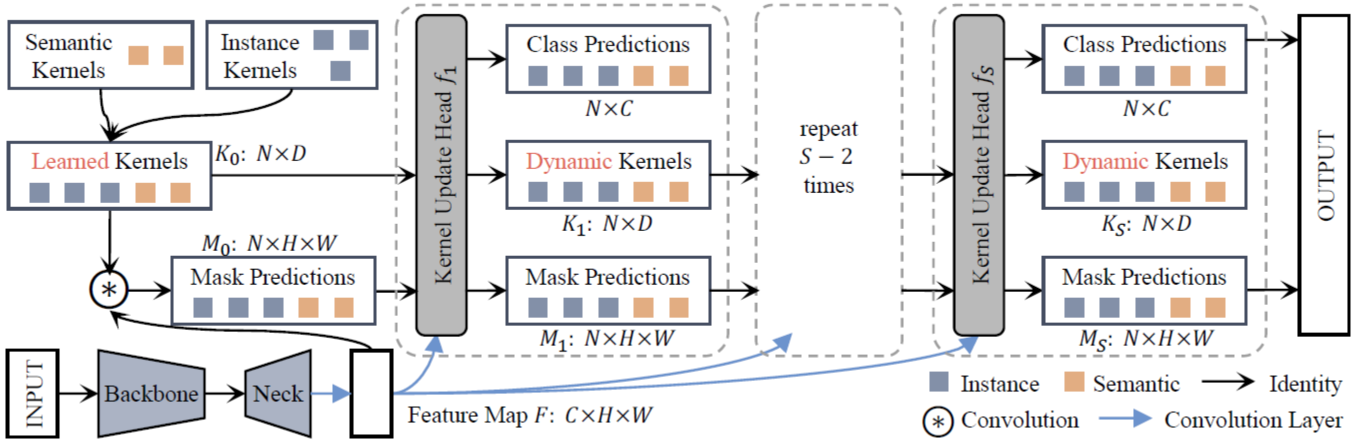


Figure 3: **K-Net for panoptic segmentaion.** A set of learned kernels first performs convolution with the feature map F to predict masks M_0. Then the kernel update head takes the mask predictions M_0, learned kernels K_0, and feature map F as input and produce class predictions, group-aware (dynamic) kernels, and mask predictions. The produced mask prediction, dynamic kernels, and feature map F are sent to the next kernel update head. This process is performed iteratively to progressively refine the kernels and the mask predictions.

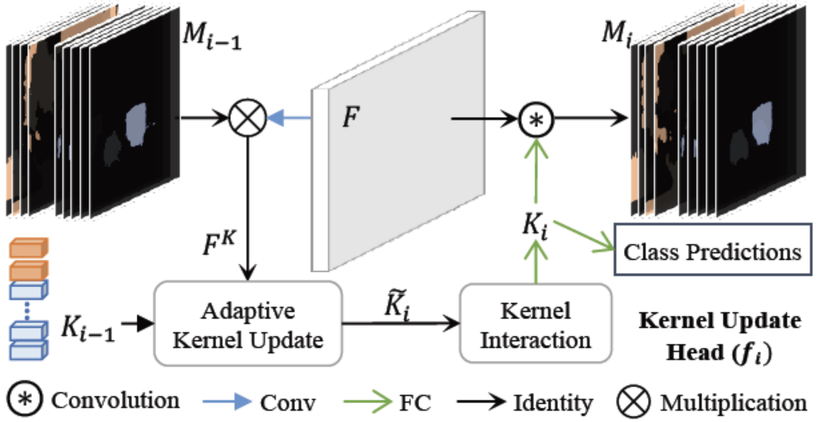


Figure 2: Kernel Update Head.

Kernel Update Head as Cross Attention:

Each kernel (query) is updated via combing corresponding instance/stuff masked features and dynamic convolution.

Motivation

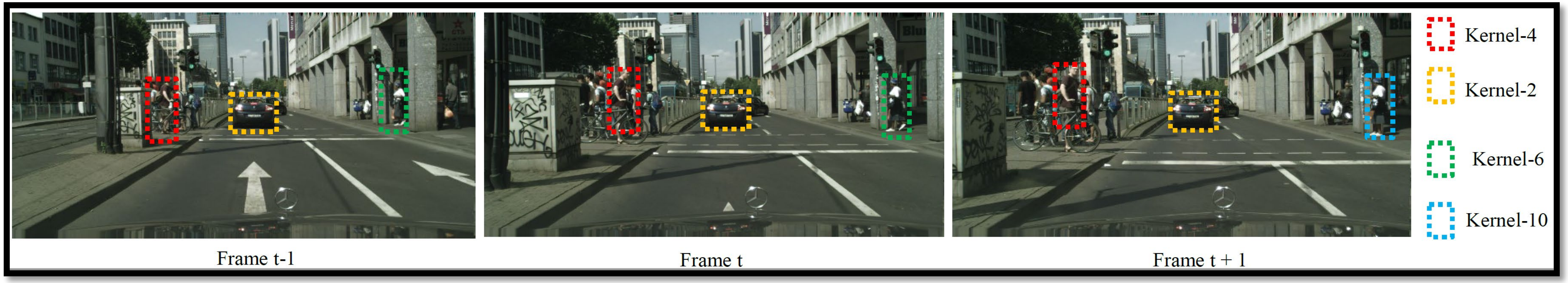


Table 1. Toy Experiment results on KITTI-STEP and Cityscape-VPS set with STQ and VPQ metrics. Unitrack [57] uses ResNet-50 as the appearance model.

KITTI-STEP	Backbone	STQ	AQ	SQ	-
K-Net	ResNet50	67.5	65.5	68.9	-
K-Net + Unitrack [57]	ResNet50	65.1	64.3	68.9	-
Cityscapes-VPS	Backbone	-	-	-	VPQ
K-Net	ResNet50	-	-	-	54.3
K-Net + Unitrack [57]	ResNet50	-	-	-	53.2

We first perform toy experiment where we find the origin K-Net itself can achieve good tracking results and even better than specific Tracker.

Motivation

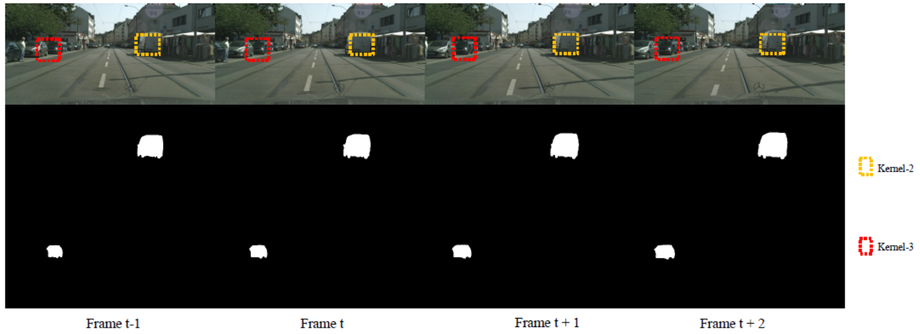


Figure 3. Toy experiment illustration. We use the K-Net directly on Cityscapes video datasets. We find that several instances are originated from **the same kernel** predictions (Red, Yellow boxes, Kernel-2 and Kernel-3). This observation motivates us to use K-Net directly on video. Best view it in color.

Each kernel corresponds to one thing mask.

Conclusion and take away message:

1， Simple Kernel-based Segmenter itself is good mask tracker.

2， How to improve such tracking ability?

1. Temporal feature fusion.
2. More consistent instance association.
3. Directly link the kernel.

Method

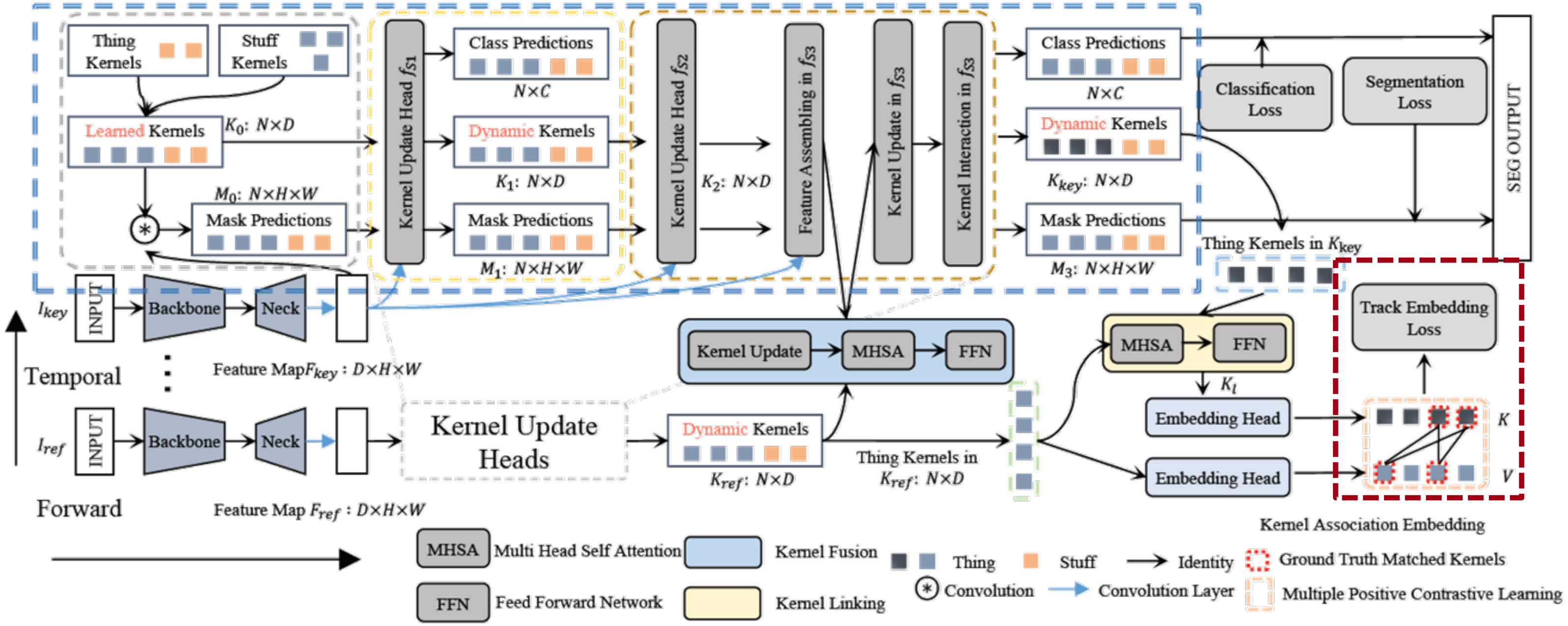


Figure 4. An illustration of our proposed Video K-Net. Our method is based on K-Net [71](in blue dashed box), which is the top-left part of the figure. Video K-Net adds Kernel Fusion at the start phase of the last stage. The Kernel Linking is performed on the output of dynamic kernels. The Embedding Head is appended at the output of kernel linking and takes kernel outputs from both sampled frames.

➢ **Learning the Kernel Association Embedding**

We propose to learn the kernel association embedding on thing kernels.

$\mathbf{v}$ kernels in key frame are matched with $\mathbf{k}$ kernels ($\mathbf{k}^+$ positative, $\mathbf{k}^-$ negative) in reference frames via a temporal contrastive loss

$$\mathcal{L}_{\text{track}} = -\sum_{\mathbf{k}^+} \log \frac{\exp(\mathbf{v}\cdot\mathbf{k}^+)}{\exp(\mathbf{v}\cdot\mathbf{k}^+) + \sum_{\mathbf{k}^-}\exp(\mathbf{v}\cdot\mathbf{k}^-)},$$

Method

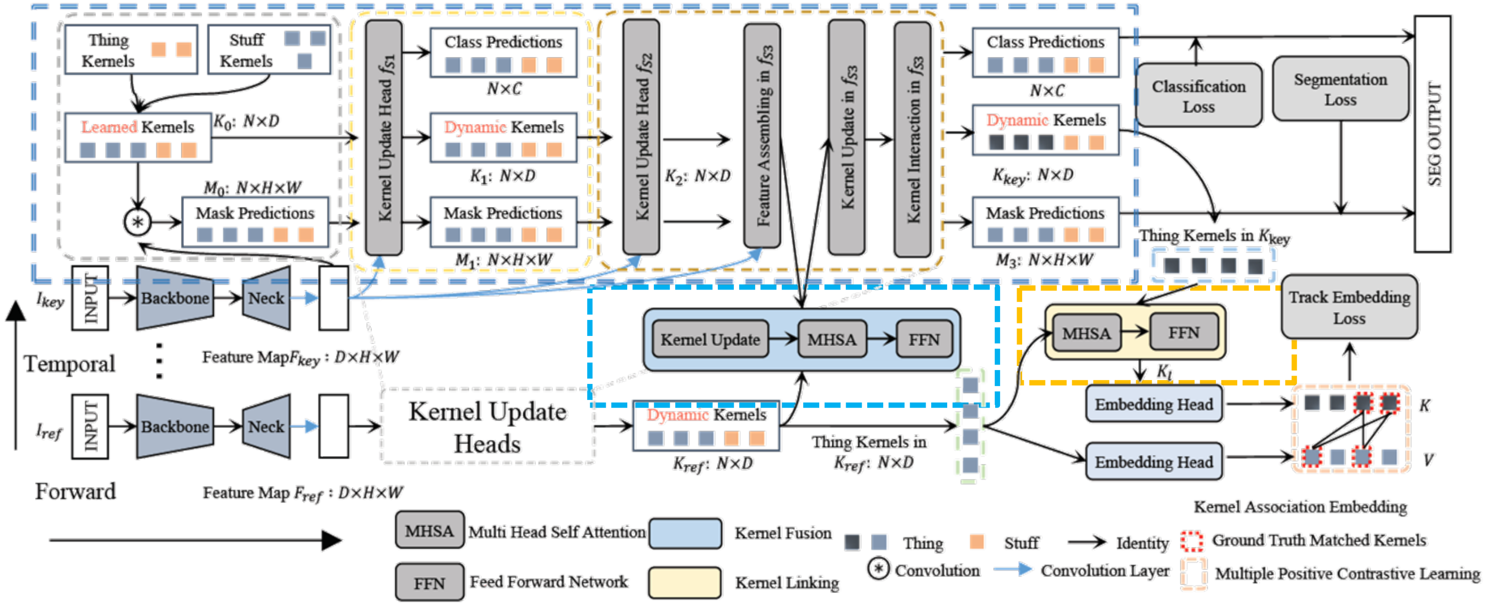


Figure 4. An illustration of our proposed Video K-Net. Our method is based on K-Net [71](in blue dashed box), which is the top-left part of the figure. Video K-Net adds Kernel Fusion at the start phase of the last stage. The Kernel Linking is performed on the output of dynamic kernels. The Embedding Head is appended at the output of kernel linking and takes kernel outputs from both sampled frames.

- ➢ **Learning to Link Kernels.**

We force to link the kernels along tracking heads for thing kernel via MSHA to learn the correspondence within thing kernels.

- ➢ **Learning to Fuse Kernels.**

We propose to fuse the kernel at the last stage of K-Net via kernel update to improve temporal segmentation consistency.

- ➢ **Generation to VSS and VIS**

For VSS: We remove the tracking branch.

For VIS: We remove online tracking and use mean kernels to represent each object in one clip.

Experiment Results

Table 6. Ablation studies and comparison analysis on KITTI-STEP validation set. All the experiments use ResNet-50 as backbone.

(a) Ablation Study on Each Components.

baseline	KAE	KL	KF	STQ	AQ	SQ
K-Net				67.5	65.5	68.9
	✓			69.3	69.0	69.8
	✓	✓		70.2	71.2	69.7
	✓	✓	✓	70.9	70.8	71.2

(b) Needs of Appearance Embeddings

Method	AQ	STQ
RoI-Align [36]	68.8	69.1
Mask-Emb [59]	67.3	68.1
Ours	70.8	70.9
Ours + Mask-Emb [59]	70.3	70.8

(c) Effect of sampling in association.

Method	STQ	AQ	SQ
K-Net	67.5	65.5	68.9
GT-based (ours)	69.3	69.0	69.8
sampling in [36]	63.1	62.1	64.3

(d) Ablation Study on Linking and Fusing Stage.

Stage	STQ	AQ	SQ
3	70.9	70.8	71.2
2	68.5	68.2	69.3
1	66.9	63.4	67.3

(e) Ablation Study on Training Settings

Settings	STQ	AQ	SQ
joint training	70.9	70.8	71.2
only train the key frame	70.1	70.1	69.8

(f) Ablation Study on Kernel Fusing

Settings	STQ	AQ	SQ
K-Net	67.5	65.5	68.9
w Update	70.9	70.8	71.2
w/o Update	67.1	66.2	68.3

We perform ablation studies on KITTI-STEP validation set.

Experiment Results

Table 3. **Results on Cityscapes-VPS validation set**. k is temporal window size in [22]. All the methods use the single scale inference without other augmentations in the test stage. In each cell, we report VPQ, VPQ_{thing} and VPQ_{stuff} in order. There is about 0.5% noise on this dataset where we report the average results (three times).

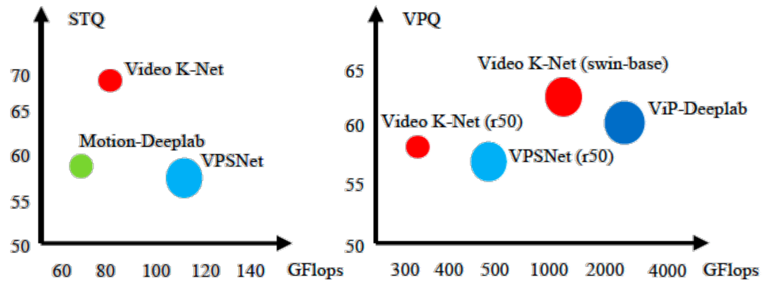
Method	Backbone	k = 0	k = 5	k = 10	k = 15	Average
VPSNet [22]	ResNet50	65.0 | 59.0 | 69.4	57.6| 45.1 | 66.7	54.4| 39.2| 65.6	52.8 | 35.8 | 65.3	57.5 | 44.8 | 66.7
SiamTrack [59]	ResNet50	64.6 | 58.3 | 69.1	57.6 | 45.6 | 66.6	54.2 | 39.2 | 65.2	52.7 | 36.7 | 64.6	57.3 | 44.7 | 55.0
ViP-Deeplab [42]	WideResNet41 [67]	68.2 | N/A | N/A	61.3 | N/A | N/A	58.2 | N/A | N/A	56.2 | N/A | N/A	60.9 | N/A | N/A
ViP-Deeplab [42]	WideResNet41 [67]+RFP [41] + AutoAug [13]	69.2 | N/A | N/A	62.3 | N/A | N/A	59.2 | N/A | N/A	57.0 | N/A | N/A	61.9 | N/A | N/A
Video K-Net	ResNet50	65.6 | 57.4 | 71.5	57.7 | 43.4 | 68.2	54.2 | 36.5 | 67.1	52.3 | 33.1 | 66.3	57.8 | 45.0 | 66.9
Video K-Net	Swin-base [30]	69.2 | 63.6 | 73.3	62.0 | 51.1 | 70.0	58.4 | 44.7 | 68.3	55.8 | 39.8 | 67.5	61.2 | 49.6 | 69.5
Video K-Net	Swin-base + RFP [41]	70.8 | 63.2 | 76.3	63.1 | 49.3 | 73.2	59.5 | 43.4 | 72.0	56.8 | 37.0 | 71.1	62.2 | 49.8 | 71.8

Table 2. **Experiment results on KITTI set with both STQ and VPQ metric.** OF refers to an optical flow network [47]. The results on validation set are shown in the several top rows, and results on test set are in the bottom rows. P means Panoptic Deeplab [10]. Following [57], we keep two decimal numbers. VPQ is obtained via average results of window size k where $k = 1, 2, 3, 4$ [57]. Top: validation set. Bottom: test set. We find 0.5% noise on this dataset where we report the average results(three times).

KITTI-STEP	Backbone	OF	STQ	AQ	SQ	VPQ
P + IoU Assoc.	ResNet50		0.58	0.47	0.71	0.44
P + SORT	ResNet50		0.59	0.50	0.71	0.42
P + Mask Propagation	ResNet50	✓	0.67	0.63	0.71	0.44
Motion-Deeplab [57]	ResNet50		0.58	0.51	0.67	0.40
VPSNet [22]	ResNet50	✓	0.56	0.52	0.61	0.43
Video K-Net	ResNet50		0.71	0.70	0.71	0.46
Video K-Net	Swin-base		0.73	0.72	0.73	0.53
Video K-Net	Swin-large		0.74	0.73	0.75	
Motion-Deeplab [57]	ResNet50		0.52	0.46	0.60	-
Video K-Net	ResNet50		0.59	0.50	0.62	-
Video K-Net	Swin-base		0.63	0.60	0.65	-

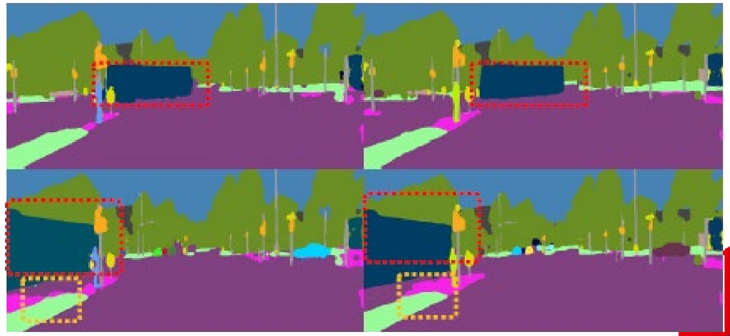
New **state-of-the art results** on VPS datasets. Table.2 and Table.3

Best performance and GFlops Trade-off on KITTI-STEP (a) and Cityscapes VPS (b).



(a) results on KITTT-STEP (b) results on Cityscapes-VPS

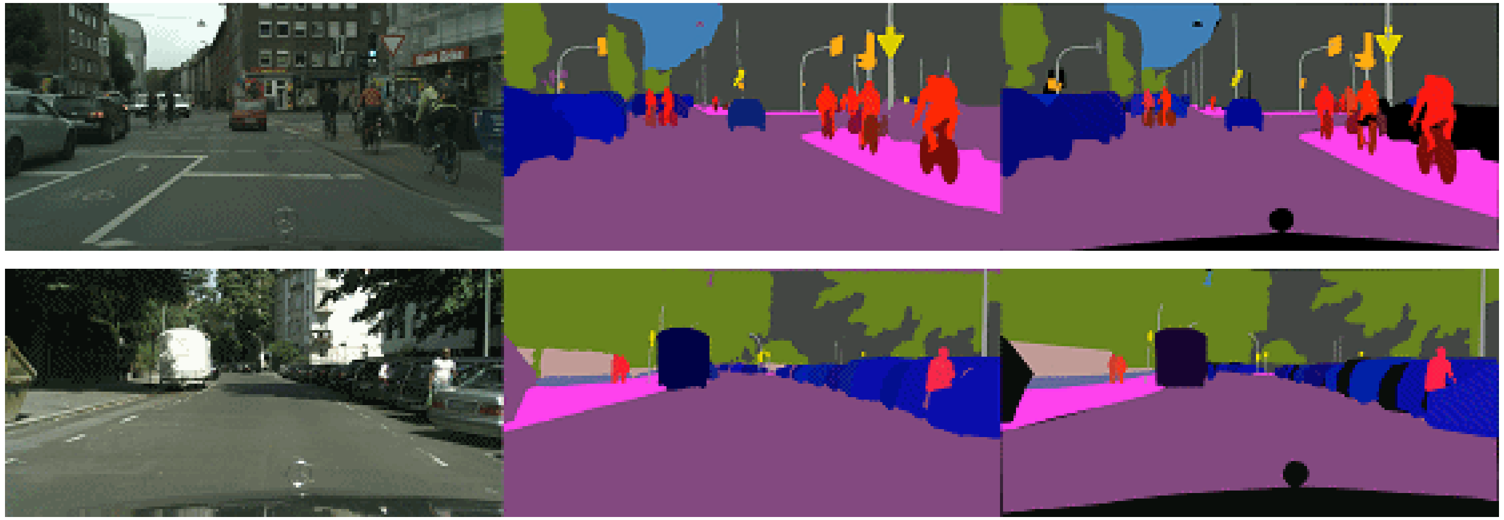
Visual Improvements over K-Net baseline.



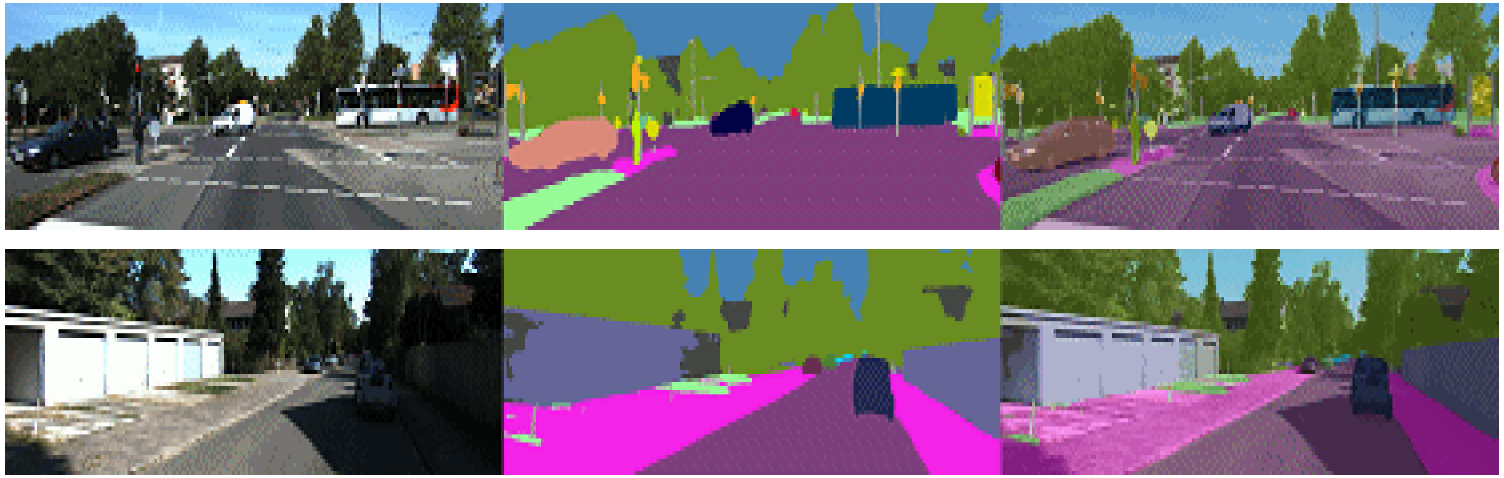
Baseline Our Video K-Net

Video Results

Short term segmentation and tracking results on Cityscapes VPS dataset.



Long term segmentation and tracking results on STEP dataset.



Video Results

Considerable results on VSS and VIS datasets. (Table-4 on VSPW and Table-5 on YouTube-VIS)

Table 4. **Results on VSPW validation set.** mVC_c means that a clip with c frames is used. All methods use the same setting for fair comparison.

VPSW	Backbone	mIoU	mVC_8	mVC_{16}
DeepLabv3+ [8]	ResNet101	35.7	83.5	78.4
PSPNet+ [71]	ResNet101	36.5	84.4	79.8
TCB(PSPNet) [33]	ResNet101	37.5	86.9	82.1
Video K-Net (Deeplabv3+)	ResNet101	37.9	87.0	82.1
Video K-Net (PSPNet)	ResNet101	38.0	87.2	82.3

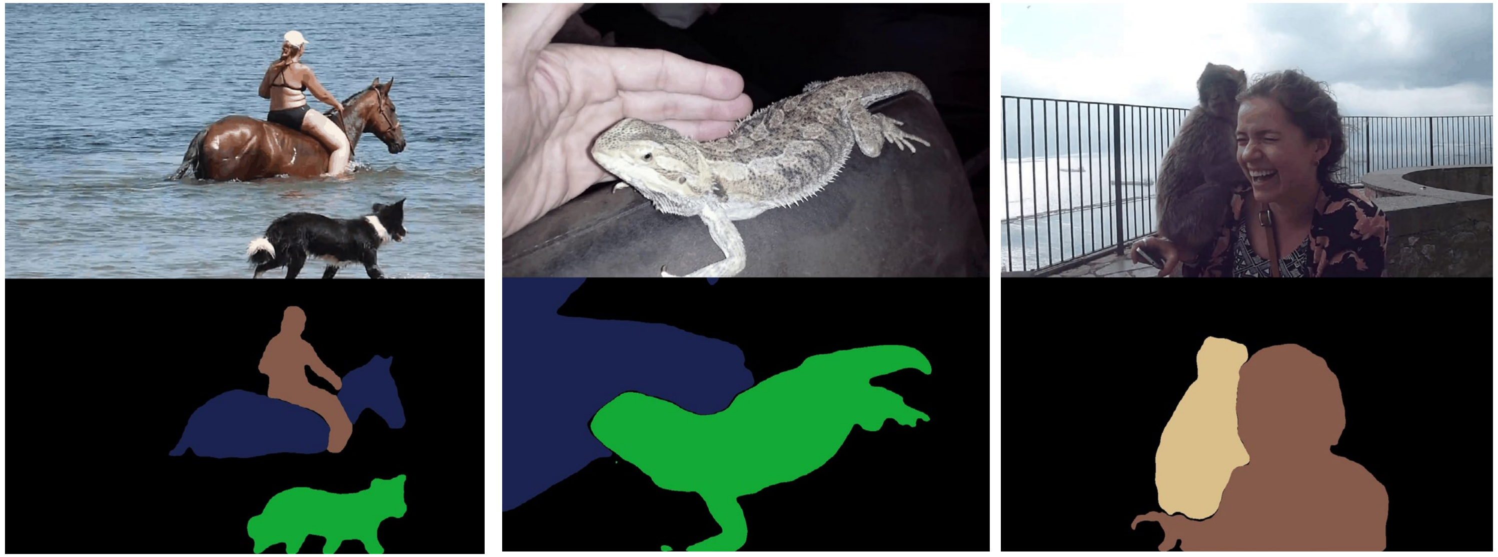
Table 5. **Video instance segmentation** AP (%) on the YouTube-VIS-2019 [66] validation dataset. The compared methods are listed by publication date.

Method	backbone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
FEELVOS [51]	ResNet50	26.9	42.0	29.7	29.9	33.4
MaskTrack R-CNN [66]	ResNet50	30.3	51.1	32.6	31.0	35.5
MaskProp [3]	ResNet-50	40.0	-	42.9	-	-
MaskProp [3]	ResNet101	42.5	-	45.6	-	-
STEm-Seg [1]	ResNet50	30.6	50.7	33.5	31.6	37.1
STEm-Seg [1]	ResNet101	34.6	55.8	37.9	34.4	41.6
CompFeat [15]	ResNet50	35.3	56.0	38.6	33.1	40.3
VisTR [56]	ResNet50	34.4	55.7	36.5	33.5	38.9
VisTR [56]	ResNet101	35.3	57.0	36.2	34.3	40.4
Video K-Net	ResNet50	40.5	63.5	44.5	40.7	49.9
Video K-Net	Swin-base	51.4	77.2	56.1	49.0	58.4

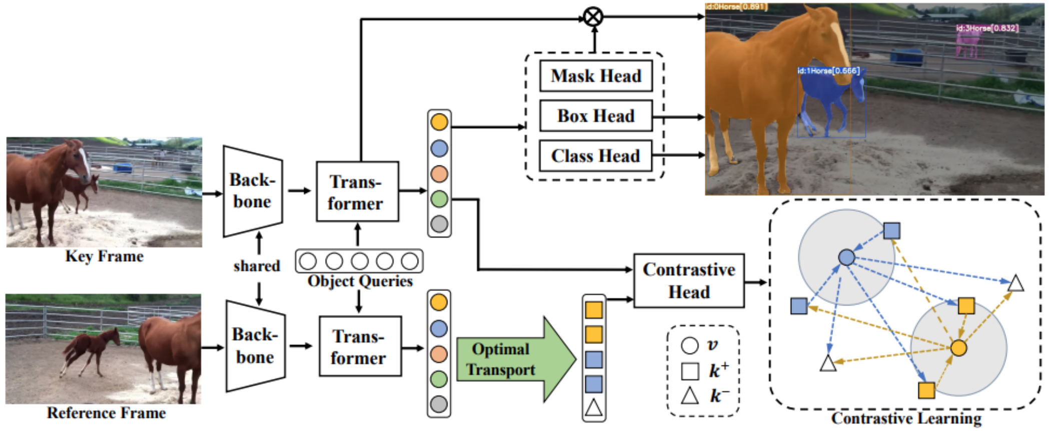
Visual Results of Video K-Net on Youtube-VIS-2019 validation set.

Top: input images. Bottom: Predicted Mask.

The same color represents the same instance



Related Works



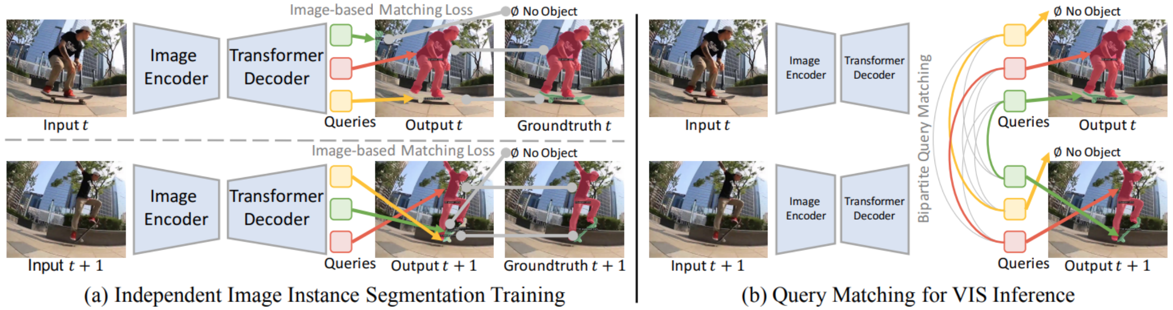


Figure 1: (a) MinVIS trains a query-based image instance segmentation model (Image Encoder + Transformer Decoder) using each frame independently. (b) During inference, the trained image instance segmentation model is used for video instance segmentation by bipartite matching of query embeddings across frames. MinVIS does not require further manually designed heuristics for tracking.

IDOL: In Defense of Online Models for Video Instance Segmentation, ECCV-2022

MinVIS: A Minimal Video Instance Segmentation Framework without Video-based Training, NeurIPS-2022

What are the Nexts?

1, Missing multiple frame information.

2, Segmentation quality is still not good enough.

Motivated by recent near-online approach, we propose Tube-Link.

Tube-Link: A Flexible Cross Tube Baseline for Universal Video Segmentation

Xiangtai Li[1] Haobo Yuan[2] Wenwei Zhang[1,4]
Guangliang Cheng[3] Jiangmiao Pang[4] Chen Change Loy[1✉]
[1] S-Lab, Nanyang Technological University [2] WHU [3] SenseTime Research [4] Shanghai AI Lab
{xiangtai.li, ccloy}@ntu.edu.sg


Xiangtai Li


Haobo Yuan


Wenwei Zhang


Guangliang Cheng


Jiangmiao Pang


Chen Change Loy

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

商汤 sensetime

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Background

Tube-Link aims For Universal Video Segmentation

Video Semantic Segmentation | Video Instance Segmentation | Video Panoptic Segmentation

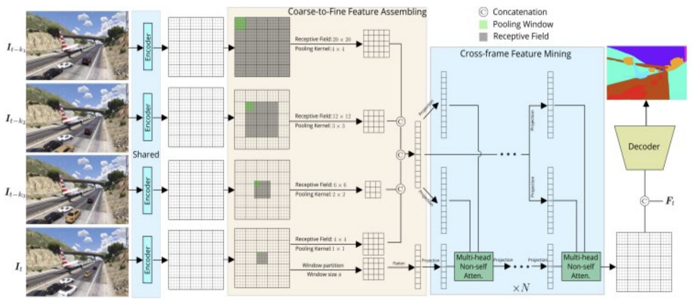
(VSS-VSPW) | (VIS-Youbue-VIS) | (VPS-KITTI-STEP)

-Video Segmentation: Video Semantic Segmentation (VSS), Video Instance Segmentation (VIS), Video Panoptic Segmentation (VPS)

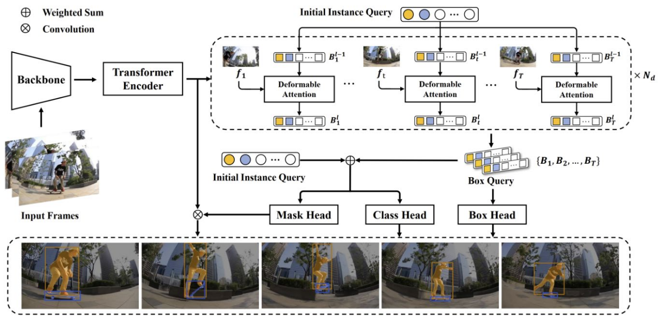
-However, most approaches solve these tasks using specific architectures.

Background

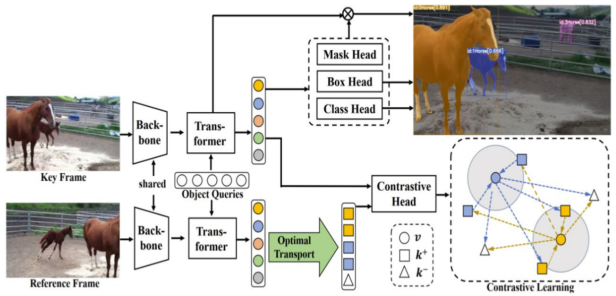
Method	VSS	VIS	VPS	Online	Nearly Online	Joint Mulitple Frames	Frame Matching	Tube Matching	Mask Matching	No Association (or Average Queries)
CFFM [20]	✓				✓	✓				✓
MRCFA [21]	✓				✓	✓				✓
Cross-VIS [29]		✓		✓			✓			
IDOL [26]		✓		✓			✓			
SeqFormer [25]		✓			✓	✓				✓
EfficientVIS [27]		✓			✓	✓				✓
VITA [10]		✓			✓	✓				✓
Min-VIS [11]		✓		✓			✓			
Gen-VIS [9]		✓		✓	✓	✓		✓		
SLOT-VPS [32]			✓		✓	✓				✓
TubeFormer [13]	✓	✓	✓		✓	✓			✓	
Video K-Net [14]	✓	✓	✓	✓			✓			



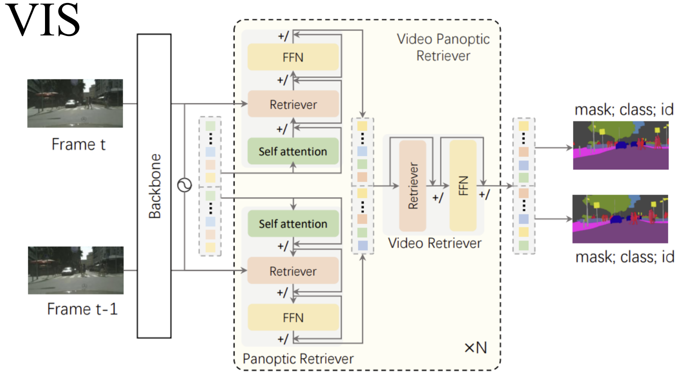
CFFM For VSS



Seqformer For VIS



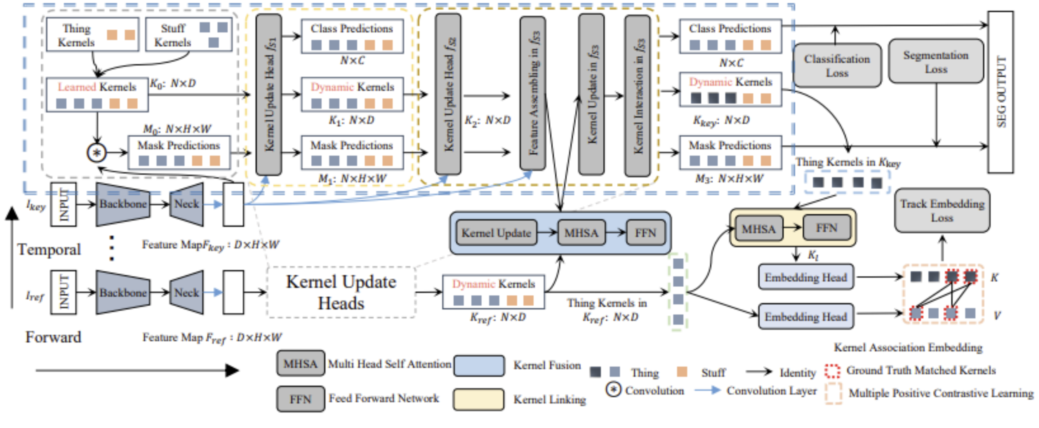
IDOL For VIS



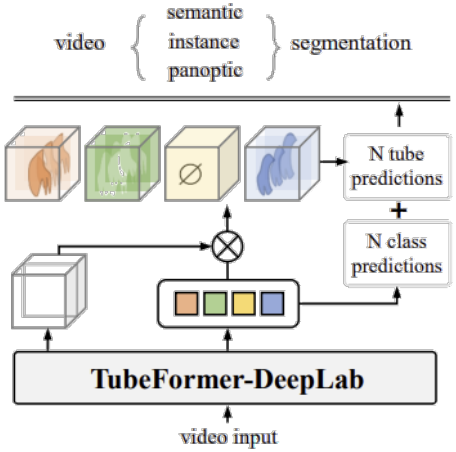
SLOT-VPS For VPS

Motivation

Method	VSS	VIS	VPS	Online	Nearly Online	Joint Mulitple Frames	Frame Matching	Tube Matching	Mask Matching	No Association (or Average Queries)
CFFM [20]	✓				✓	✓				✓
MRCFA [21]	✓				✓	✓				✓
Cross-VIS [29]		✓		✓			✓			
IDOL [26]		✓		✓			✓			
SeqFormer [25]		✓			✓	✓				✓
EfficientVIS [27]		✓			✓	✓				✓
VITA [10]		✓			✓	✓				✓
Min-VIS [11]		✓		✓			✓			
Gen-VIS [9]		✓		✓	✓	✓		✓		
SLOT-VPS [32]			✓		✓	✓				✓
TubeFormer [13]	✓	✓	✓		✓	✓			✓	
Video K-Net [14]	✓	✓	✓	✓			✓			



Video K-Net-CVPR-2022



TubeFormer-CVPR-2022

Motivation

Questions:

- Recently, A more challenging VPS dataset VIP-Seg is introduced, which brings more challenges.

- The performance issue of Universal Video Segmentation
  - Eg: Both Video K-Net and TubeFormer **cannot** achieve better results on VIS datasets.
  - VIS methods cannot generalize to VPS and VSS.

- There should be a trade-off on online and nearly online approaches to support more diverse video inputs.

Is there any architecture or meta-architecture to solve these problems?

Motivation

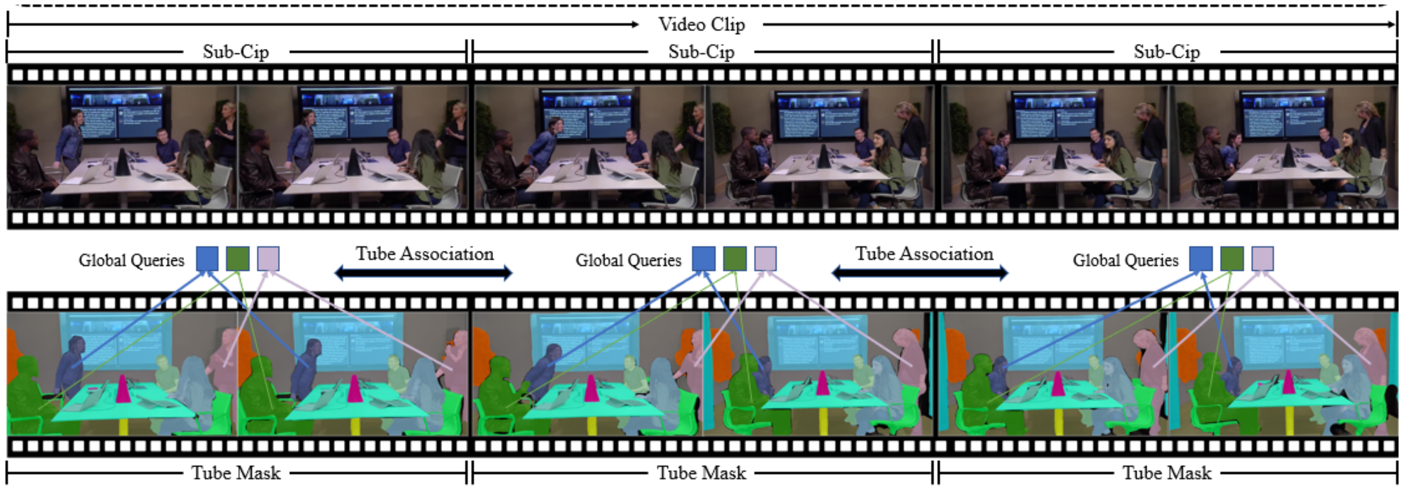
Our Contributions

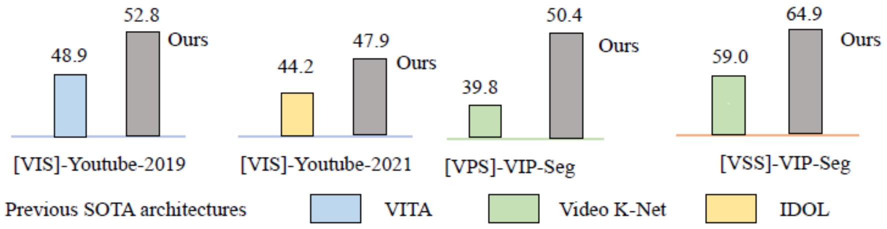
- In this work, we introduce a nearly-online approach, named Tube-Link.

- The key insight to explore the cross tube association rather than cross frame.

- Based on Mask2Former-VIS, our framework is flexible and support three different video segmentation tasks

Method





Our Contributions

- In this work, we introduce a nearly-online approaches named Tube-Link.

- The key insight to explore the cross tube association rather than cross frame.

- Based on Mask2Former-VIS, our framework is flexible and support three different video segmentation tasks

Method

Universal Video Segmentation Formation:

From *definition* of **VPS** task,

We formulate universal video segmentation as **linking** **short tracked tube**.

Table 1: **Exploration experiment on tube-wise matching.** Youtube-VIS: mAP. VIP-Seg:VPQ. We directly use pre-trained models by changing the input to two consecutive frames.

Method	Youtube-VIS-2019	Youtub-VIS-2021	VIP-Seg
MiniVIS [15]	47.4	44.2	-
MiniVIS + tube matching	48.8 (+1.4)	45.5 (+1.3)	-
Video K-Net [21]	-	-	26.1
Video K-Net + tube matching	-	-	27.6 (+1.5)

What we done:

-Use *Tube-wised matching* to replace the *Frame-wised matching*.

What we find:

-We find that using cross tube matching achieves better results even **without** re-training.

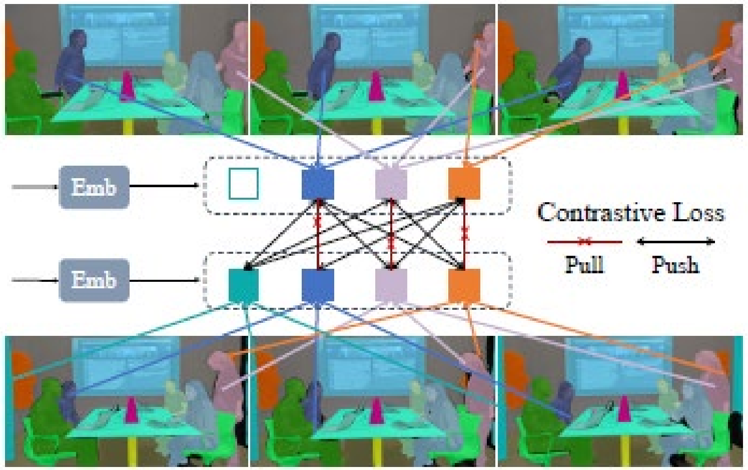
Motivation:

-The findings motivate us to explore cross tube relation in temporal.

Method

1, We adopt extended Mask2Fomer-VIS as strong baseline

In particular, we add stuff queries to adapt such architecture for VSS and VPS.

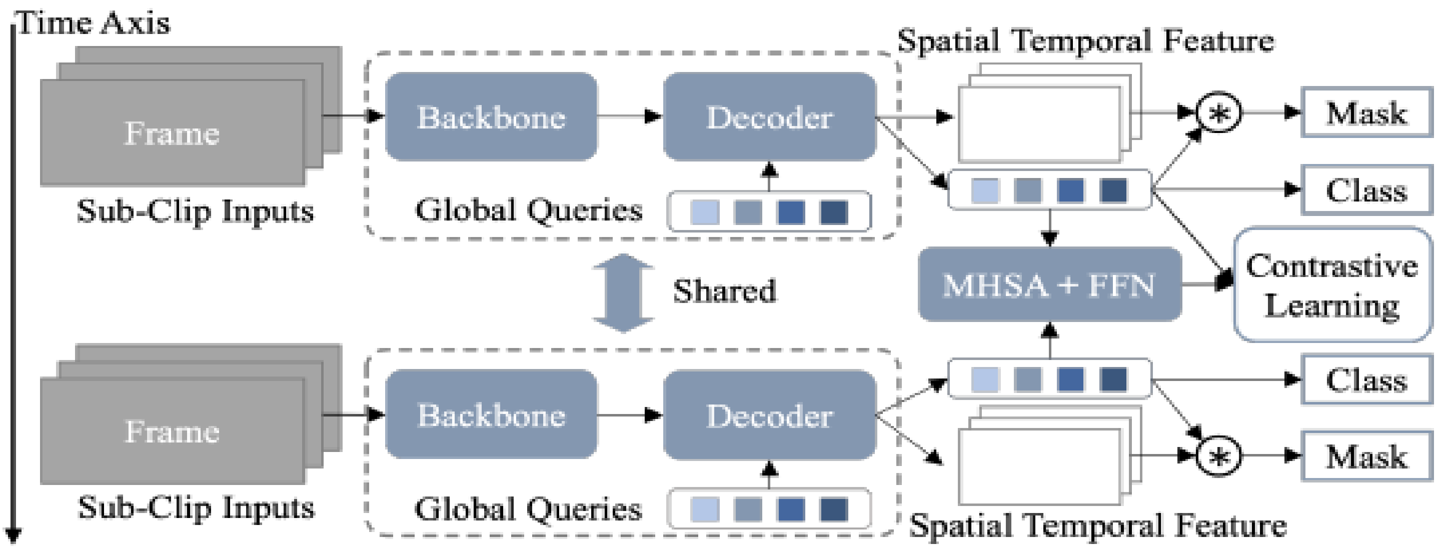


2, Learning Cross Tube Relations:

-2.1 Cross-Tube Linking-> Directly Learn the Relation Of Global Queries.

-2.2 Cross-Tube Temporal Contrastive Learning-> Learn the contrastive query association via the tube-masks.

Method



Training:

We perform tube-wised training for segmentation loss and cross-tube training for tracking loss.

Inference:

We perform **tube-wised matching** for VPS/VIS tasks.

Experiment Results

Table 8: **Ablation studies and comparative analysis on VIPSeg validation set with the ResNet50 backbone.**

(a) Ablation Study on Each Component.

baseline	TCL	CTL	VPQ_{th}	VPQ
Mask2Former-VIS+ (F)	-	-	29.4	32.4
Mask2Former-VIS+ (T)	-	-	31.0	34.5
	✓	-	34.6	36.8
	✓	✓	35.1	37.5

(b) Design Choices of TCL.

Method	VPQ	STQ
Dense Query [32]	30.2	30.1
Sparse Query [25]	34.5	35.1
Global Query(Ours)	37.5	36.5

(c) Association Target Assign.

Method	VPQ	STQ
All-Masks [32]	30.1	29.2
GT-Mask [25]	35.6	35.9
Tube-Mask	37.5	36.5

(d) Input Sub-clip Size with Tube Window Size of 2 as Input.

Clip Size	STQ	VPQ	VPQ_{th}
T=1	34.5	35.6	30.2
T=2	36.5	37.5	35.1
T=2(ovl)	35.9	37.3	35.0
T=3	36.4	37.0	35.3

(e) Tube-Window for Inference with Input Sub-clip Size 2 for Training.

Window Size	STQ	VPQ	VPQ_{th}
W=2	36.5	37.5	35.1
W=4	39.2	39.0	38.2
W=6	39.5	39.2	38.9
W=8	38.3	38.5	37.3

(f) Tracking Choices with the Default Setting of Tab.(d).

Settings	STQ	VPQ	VPQ_{th}
Extra Tracker [51, 53]	33.9	36.6	34.1
RoI Features [32]	34.5	35.9	34.5
Query Embedding [25]	33.1	36.0	33.0
Our Tube embedding	36.5	37.5	35.1

1, Add Temporal Contrastive Learning and Cross-Tube Linking (CTL) improve the performance.

2, The global query work better than sparse sampled query from each frame.

3, Tube-mask as association target.

Experiment Results

Table 3: **Results on VIPSeg-VPS [25] validation dataset.** We report VPQ and STQ for reference. Following Miao *et al.* [25], we report VPQ scores at different window sizes (1, 2, 4, 6). We report the results obtained from either an efficient or strong backbone for comparison.

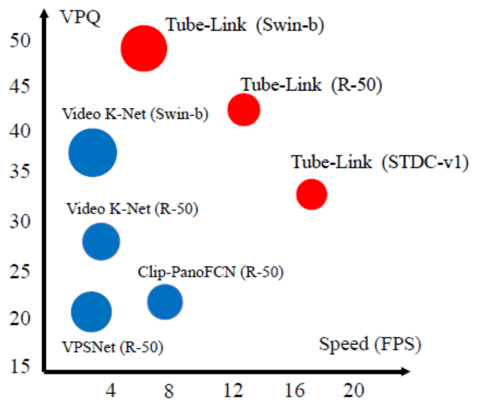
Method	backbone	VPQ^1	VPQ^2	VPQ^4	VPQ^6	VPQ	STQ
VIP-DeepLab [32]	ResNet50	18.4	16.9	14.8	13.7	16.0	22.0
VPSNet [18]	ResNet50	19.9	18.1	15.8	14.5	17.0	20.8
SiamTrack [49]	ResNet50	20.0	18.3	16.0	14.7	17.2	21.1
Clip-PanoFCN [25]	ResNet50	24.3	23.5	22.4	21.6	22.9	31.5
Video K-Net [21]	ResNet50	29.5	26.5	24.5	23.7	26.1	33.1
Video K-Net+ [8, 21]	ResNet50	32.1	30.5	28.5	26.7	29.1	36.6
Video K-Net [21]	Swin-base	43.3	40.5	38.3	37.2	39.8	46.3
Tube-Link	STDCv1	32.1	31.3	30.1	29.1	30.6	32.0
Tube-Link	STDCv2	33.2	31.8	30.6	29.6	31.4	32.8
Tube-Link	ResNet50	41.2	39.5	38.0	37.0	39.2	39.5
Tube-Link	Swin-base	54.5	51.4	48.6	47.1	50.4	49.4

Table 5: **Results on the Youtube-VIS datasets.** We report the mAP metric. † adopt COCO video pseudo labels. Axial means using the extra Axial Attention [43]. Our method does not apply these techniques for simplicity.

Method	Backbone	YTVIS-2019	YTVIS-2021
VISTR [45]	ResNet50	36.2	-
TubeFormer [19]	ResNet50 + Aixal	47.5	41.2
IFC [17]	ResNet50	42.8	36.6
Seqformer [50]	ResNet50	47.4	40.5
Mask2Former-VIS [7]	ResNet50	46.4	40.6
IDOL [51]	ResNet50	46.4	43.9
IDOL [51] †	ResNet50	49.5	-
VITA [14] †	ResNet50	49.8	45.7
Min-VIS [15]	ResNet50	47.4	44.2
Tube-Link	ResNet50	52.8	47.9
SeqFormer [50]	Swin-large	59.3	51.8
Mask2Former-VIS [7]	Swin-large	60.4	52.6
IDOL [51]	Swin-large	61.5	56.1
IDOL [51]	Swin-large †	64.3	-
VITA [14] †	Swin-large	63.0	57.5
Min-VIS [15]	Swin-large	61.6	55.3
Tube-Link	Swin-large	64.6	58.4

Table 6: **Results on the KITTI val set.** OF refers to an optical flow network [49].

KITTI-STEP	Backbone	OF	STQ	AQ	SQ	VPQ
P + Mask Propagation	ResNet50	✓	0.67	0.63	0.71	0.44
Motion-Deeplab [56]	ResNet50		0.58	0.51	0.67	0.40
VPSNet [24]	ResNet50	✓	0.56	0.52	0.61	0.43
TubeFormer-DeepLab [25]	ResNet-50 + Axial		0.70	0.64	0.76	0.51
Video K-Net [29]	ResNet50		0.71	0.70	0.71	0.46
Video K-Net [29]	Swin-base		0.73	0.72	0.73	0.53
Tube-Link	ResNet50		0.68	0.67	0.69	0.51
Tube-Link	Swin-base		0.72	0.69	0.74	0.56



New **SOTA** results on VSS/VIS/VPS using **one** architecture/solution.

Visual Comparison

Mask2Former-VIS | Tube-Link | Difference Maps

Video K-Net+ | Tube-Link

Video K-Net+ | Tube-Link

Summary

1, TransVOD: solve the video object detection problems.

2, Video K-Net: unify VPS subtasks via kernel linking and association.

3, Tube-Link: flexible and universal video segmentation framework.

Summary

1, TransVOD: solve the video object detection problems.

2, Video K-Net: unify VPS subtasks via kernel linking and association.

3, Tube-Link: flexible and universal video segmentation framework.

What are the Nexts?

Beyond pixel wised recognition:

+ Geometry (Depth Estimation)

+ Reasoning (Video Scene Graph)

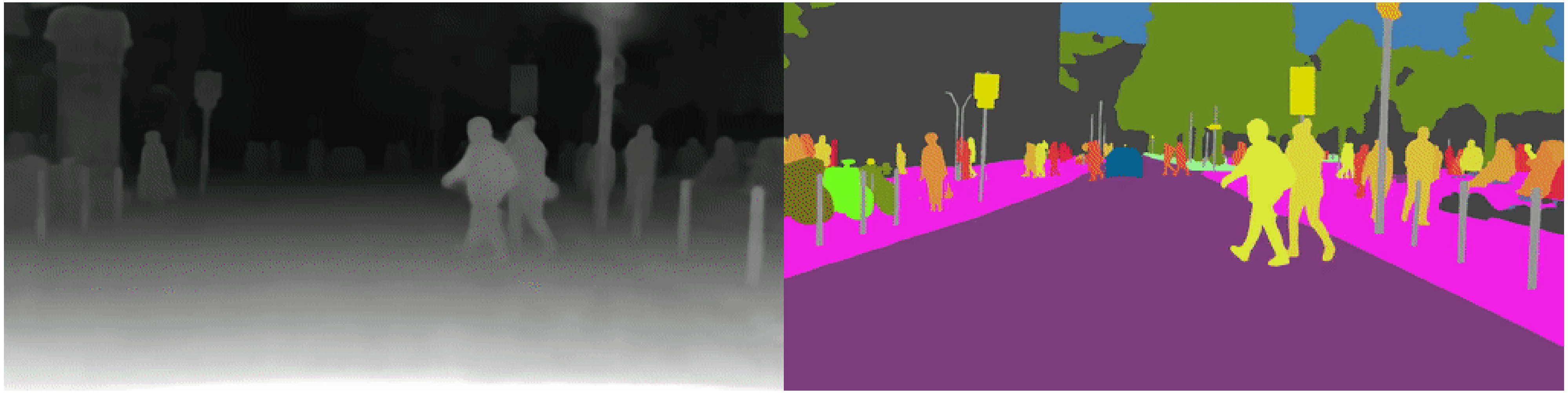
What are the Nexts?

Beyond pixel-wised recognition:

+ Geometry (Depth Estimation)

+ Reasoning (Video Scene Graph)

Task Introduction



Depth-aware Video Panoptic Segmentation (DVPS) Task:

1). Taking raw videos as input.

2). Predicting instance-level temporal-consistent segmentation results.

3). Predicting depth results for every pixel.

4). A complex and holistic scene understanding task.

Learning Visual Perception with Depth-aware Video Panoptic Segmentation CVPR-2021

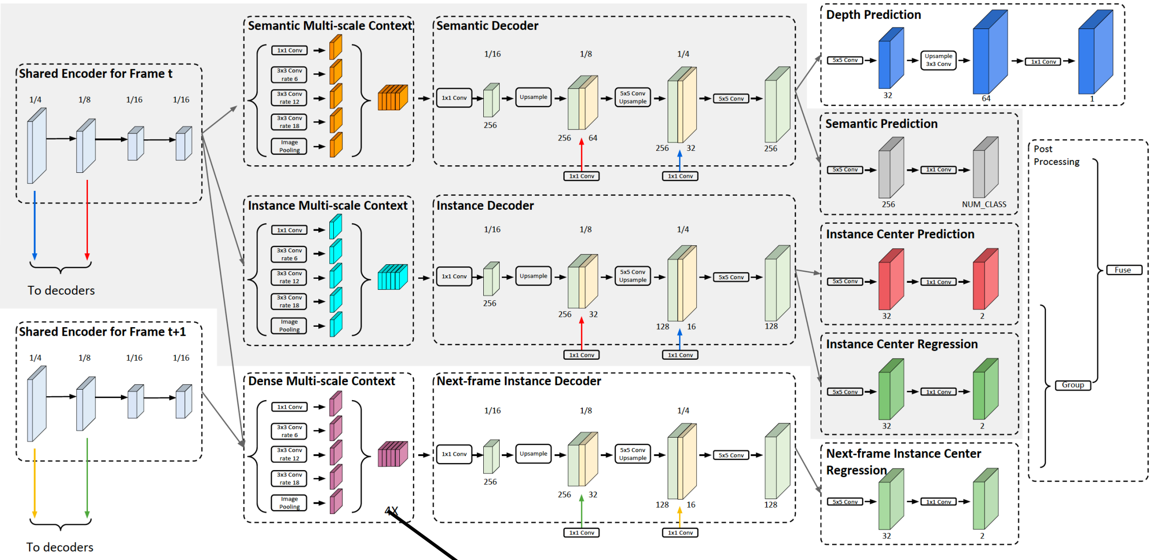
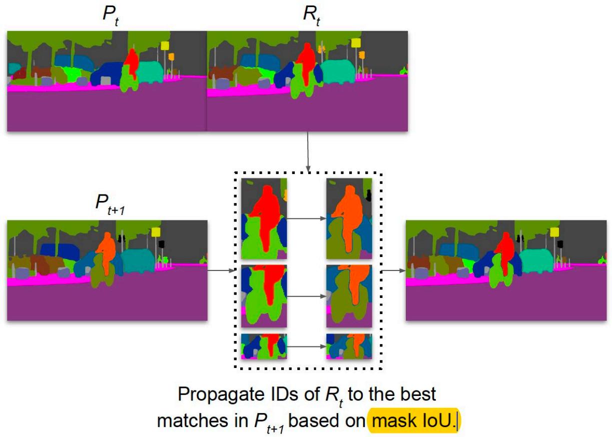


Figure 3: ViP-DeepLab expends Panoptic-DeepLab [17] (the gray part) by adding a depth prediction head to perform monocular depth estimation and a next-frame instance branch which regresses to the object centers in frame t for frame $t+1$.

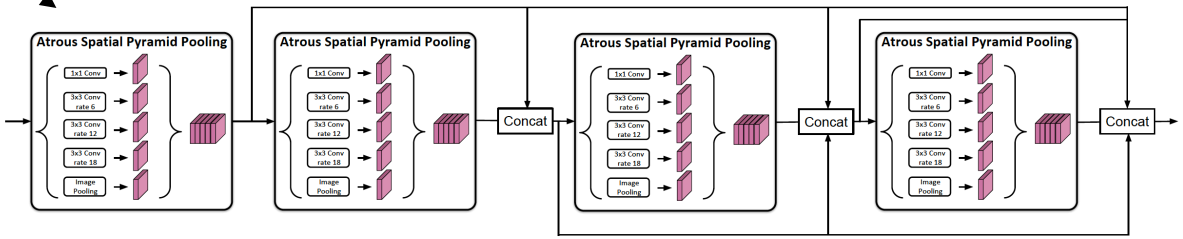


Offline Tracking via Mask IoU matching

(1) monocular depth estimation

(2) video panoptic segmentation

STOA results on VPS datasets



ECCV TEL AVIV 2022

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE | VALSE

♫ PolyphonicFormer : Unified Query Learning for Depth-aware Video Panoptic Segmentation







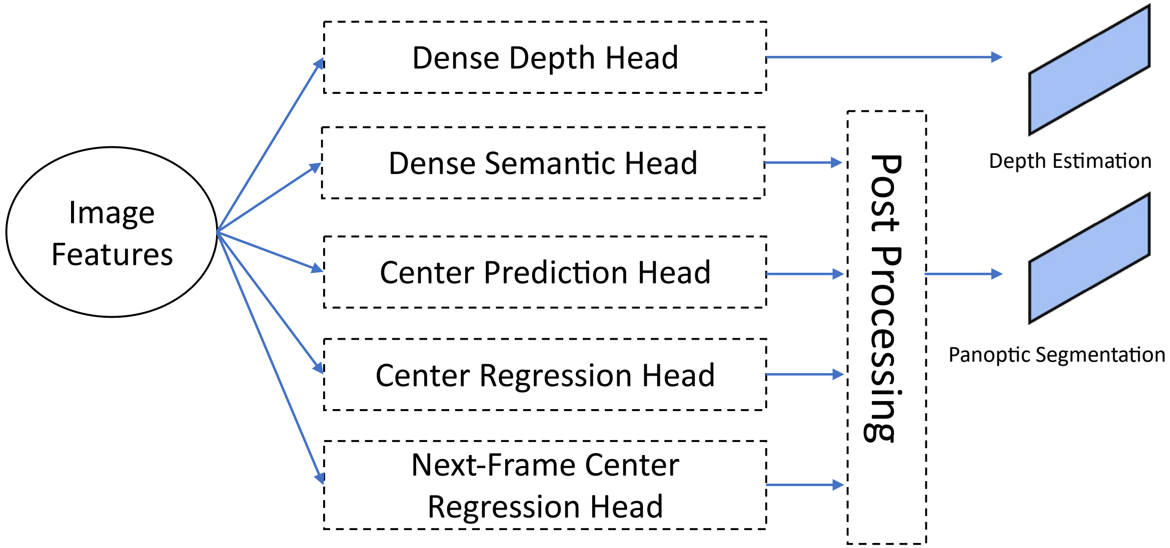

Haobo Yuan[1*] Xiangtai Li[2*] Yibo Yang[3] Guangliang Cheng[4] Jing Zhang[5] Yunhai Tong[2] Lefei Zhang[1] Dacheng Tao[3]

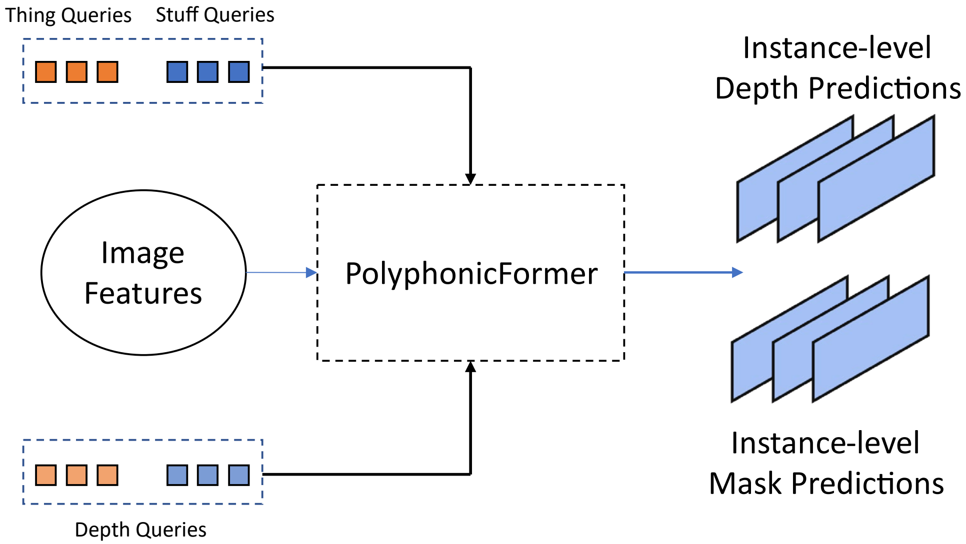
[1]Wuhan University, [2]Peking University, [3]JD Explore Academy,
[4]SenseTime Research, [5]The University of Sydney.

武漢大學 WUHAN UNIVERSITY | PEKING UNIVERSITY 1898 | JD.COM | 商汤 sensetime | THE UNIVERSITY OF SYDNEY

Motivation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE | VALSE





Previous Work on DVPS:

1). Complex.

2). Computationally heavy.

3). Ignore relationship between geometry and semantics.

4). Task competition.

PolyphonicFormer (Ours)

1). **Simple Pipeline.** ☑

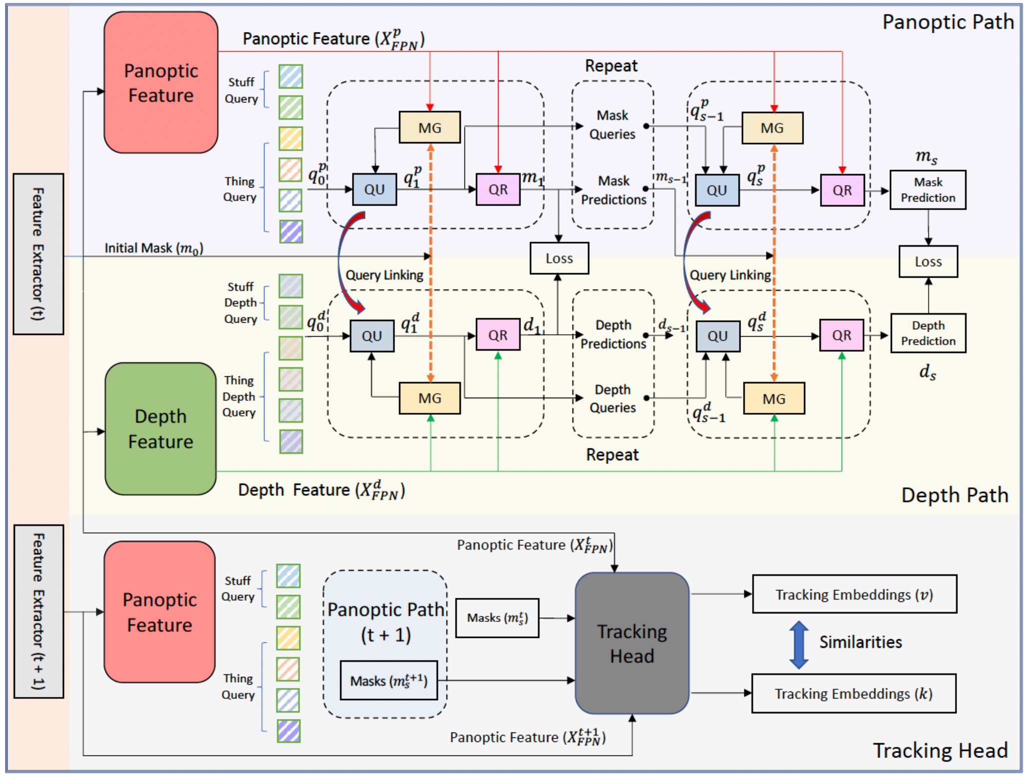
2). **Unified and Efficient (relatively).** ☑

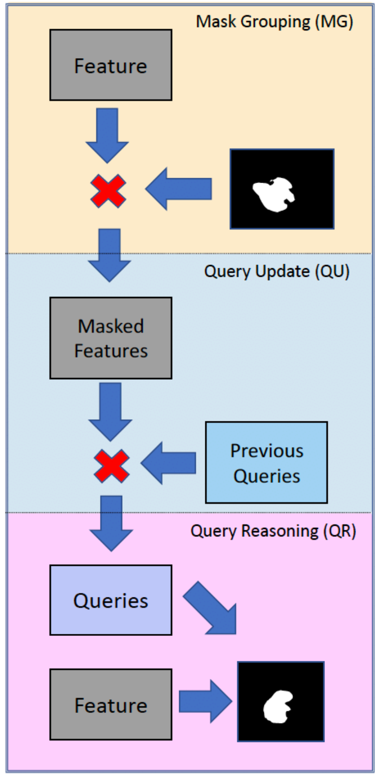
3). **Jointly predicting geometry (depth) and semantics (panoptic segmentation).** ☑

4). **Mutual benefit.** ☑

Our target is to build a UNIFIED framework, and JOINTLY predict semantics and geometry information for DVPS.

Method





Unified Query Modeling for both Depth and Panoptic (Things and Stuff)

1, Backbone Feature Extractor

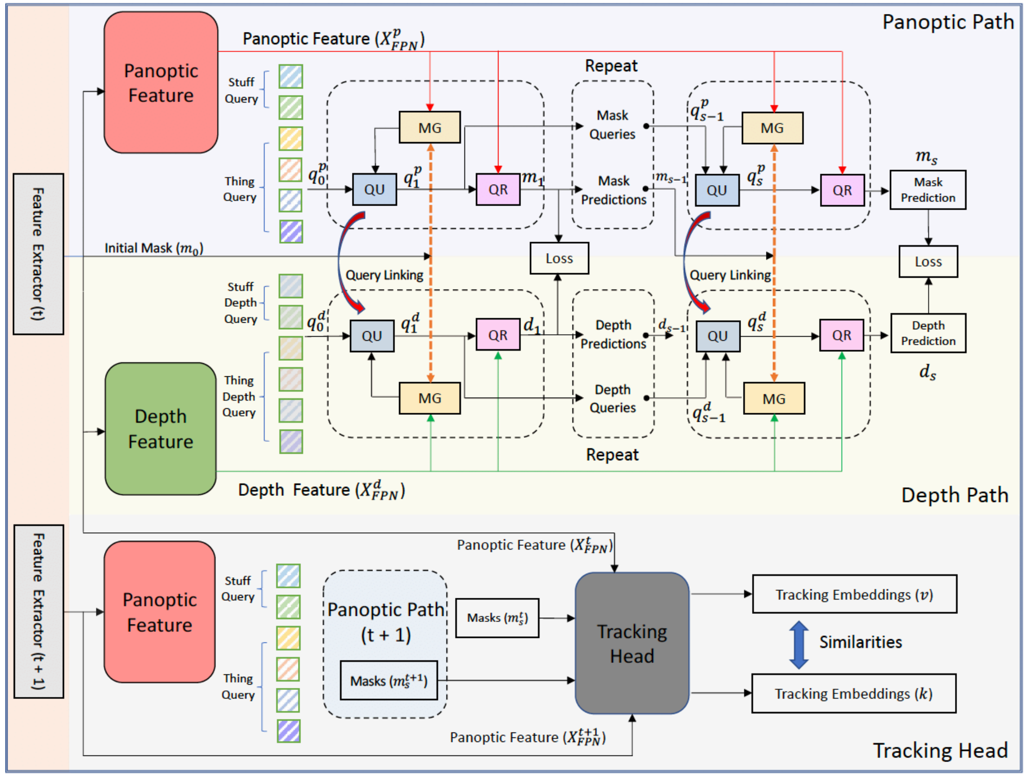
2, Panoptic Path

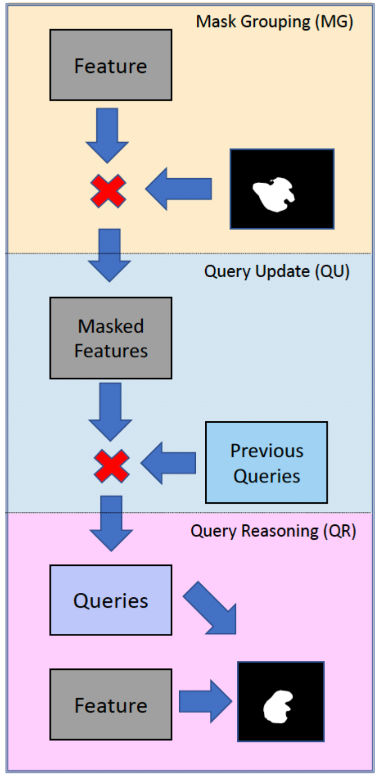
3, Depth Path

4, Panoptic Path in next frame for tracking.

Experiments show the joint modeling Depth Prediction and Panoptic Segmentation via Query Leads to better results for each other.
We term our method **PolyphonicFormer**: Means different queries come from different sources (depth or panoptic) but both can benefit each other which is just like polyphony used in music field.

Method





3, Joint Depth and Panoptic Query Modeling

Each thing and stuff query corresponds to on each depth query. (Doing Broadcasting)

4, Motivated by K-Net and Sparse-RCNN, we proposed to refined and update the above queries and query features via Dynamic Convolution. This process is refined iteratively.

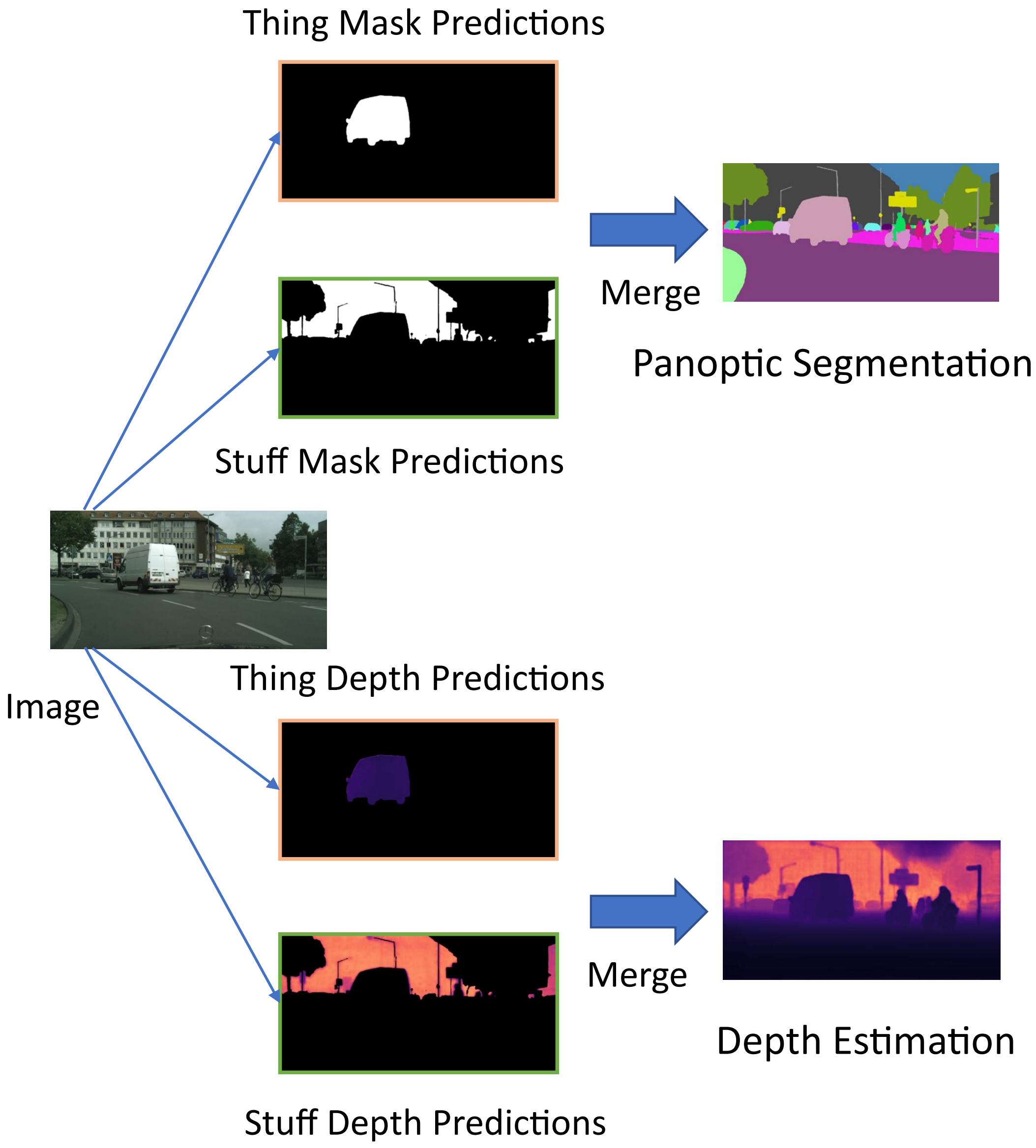
MG: Mask Grouping -> Grouping Query Features
QU: Query Update -> Update Learned Query via MLP
QR: Query Reasoning -> Self Attention along the Query

1, Two path results in different features for further process.

2, Initial Depth Query Weight is obtained from the dense depth prediction

This results in less computation cost and avoids noises from two different modalities (Depth and Semantics)

Method



Panoptic Segmentation: (with query learning)

One Query -> One Thing / Stuff

Bipartite Matching for calculating training loss.

Depth Estimation: (with query learning)

One Query -> Depth Map of Corresponding Thing / Stuff

Cal Training Loss using Panoptic **Bipartite Matching** Results

Summary : Instance-level Depth Estimation Paradigm with Query Learning for implicitly leveraging semantic information.

Experiment Results

DVPQ^k_λ on Cityscapes-DVPS	k = 1			k = 2			k = 3			k = 4			Average			FLOPs
PolyphonicFormer $\lambda = 0.50$	70.6	63.0	76.0	62.9	49.2	72.9	59.3	42.3	71.7	56.5	36.9	70.8	62.3	47.9	72.9	-
PolyphonicFormer $\lambda = 0.25$	67.8	61.0	72.8	60.4	47.6	69.8	56.9	40.8	68.6	54.3	35.8	67.8	59.9	46.3	69.8	-
PolyphonicFormer $\lambda = 0.10$	50.2	43.4	55.2	44.4	33.4	52.4	41.5	28.6	51.0	39.5	24.7	50.4	43.9	32.5	52.3	-
Average: PolyphonicFormer	62.9	55.8	68.0	55.9	43.4	65.0	52.6	37.2	63.8	50.1	32.5	63.0	**55.4**	42.2	65.0	1,675G
Average: ViP-Deeplab [43]	61.9	55.9	66.3	55.6	44.3	63.8	52.4	38.4	62.6	50.4	34.6	61.9	55.1	43.3	63.6	9,451G

DVPQ^k_λ on SemKITTI-DVPS	k = 1			k = 5			k = 10			k = 20			Average			FLOPs
PolyphonicFormer $\lambda = 0.50$	58.5	55.1	61.0	52.0	42.3	59.1	50.6	39.9	58.5	49.9	38.6	58.0	52.8	44.0	59.2	-
PolyphonicFormer $\lambda = 0.25$	56.3	54.0	57.9	49.7	41.1	56.0	48.4	38.7	55.5	47.7	37.6	55.0	50.5	42.9	56.1	-
PolyphonicFormer $\lambda = 0.10$	41.8	41.1	42.4	35.1	28.2	40.1	33.7	26.0	39.3	33.0	25.1	38.7	35.9	30.1	40.1	-
Average: PolyphonicFormer	52.2	50.1	53.8	45.6	37.2	51.7	44.2	34.9	51.1	43.4	33.8	50.6	**46.4**	39.0	51.8	402G
Average: ViP-Deeplab [43]	48.9	42.0	53.9	45.8	36.9	52.3	44.4	34.6	51.6	43.4	33.0	51.1	45.6	36.6	52.2	2,267G

Results on Cityscapes-DVPS and SemKITTI-DVPS (DVPQ).

Our method achieves better results with about ¼ computational cost.

Method	k = 1	k = 2	k = 3	k = 4	VPQ
VPSNet [21]	65.0	57.6	54.4	52.8	57.5
SiamTrack [63]	64.6	57.6	54.2	52.7	57.3
ViP-Deeplab [43]	69.2	62.3	59.2	57.0	61.9
Ours (ResNet50)	65.4	58.6	55.4	53.3	58.2
Ours (Swin-b)	70.8	63.1	59.5	56.8	62.3

Results on Cityscapes-VPS. (VPQ)

Our method also outperforms some other works on VPS (subtask of DVPS).

Experiment Results

Method	Depth	Panoptic	Ins	PQ ↑	abs rel ↓
ViP-Deeplab [43]	✓	✓	-	60.6	0.112
Depth	✓	-	-	N/A	0.084
Panoptic	-	✓	-	63.7	N/A
Hybrid (ours)	✓	✓	-	65.1	0.089
PolyphonicFormer (ours)	✓	✓	✓	65.2	0.080

L_{depth}	PQ ↑	abs rel ↓
0.1	65.4	0.101
1.0	65.3	0.089
5.0	65.2	0.080
10	65.4	0.079

1). Unified framework is good for mutual benefit and robust to loss weight choices between sub-tasks rather than mutual competition.

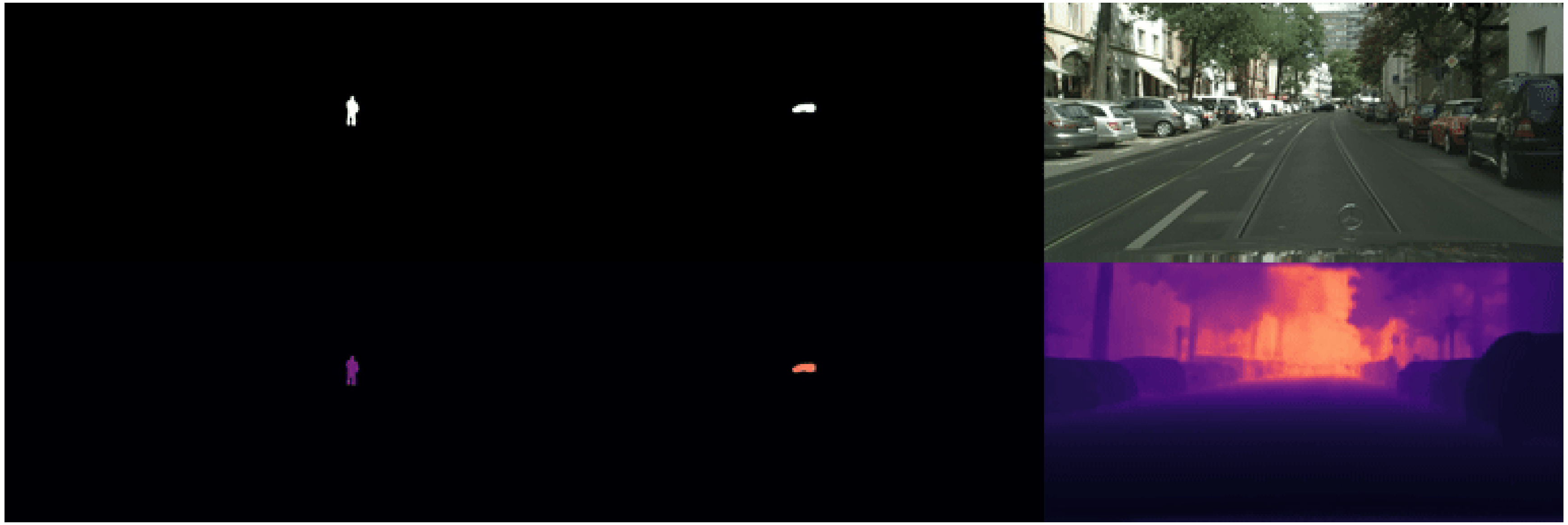
Stages	PQ ↑	abs rel ↓
1	64.1	0.081
2	64.6	0.081
3	65.2	0.080

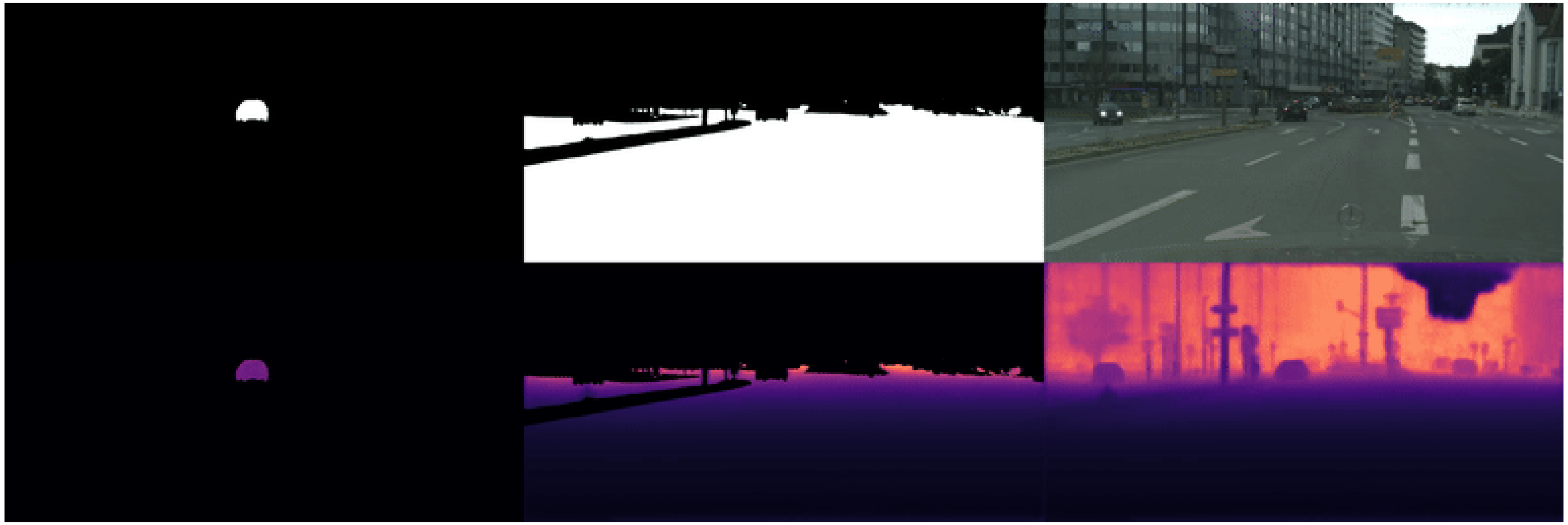
2). Iteratively query modeling for updating instance-level information.

Method	DSTQ ↑	AQ ↑
PolyphonicFormer + DeepSort [62]	51.8	25.9
PolyphonicFormer + Unitrack [59]	49.3	22.5
PolyphonicFormer + QuasiDense [38]	63.6	46.2

3). PolyphonicFormer is capable of tracking with different appearance-based tracking heads.

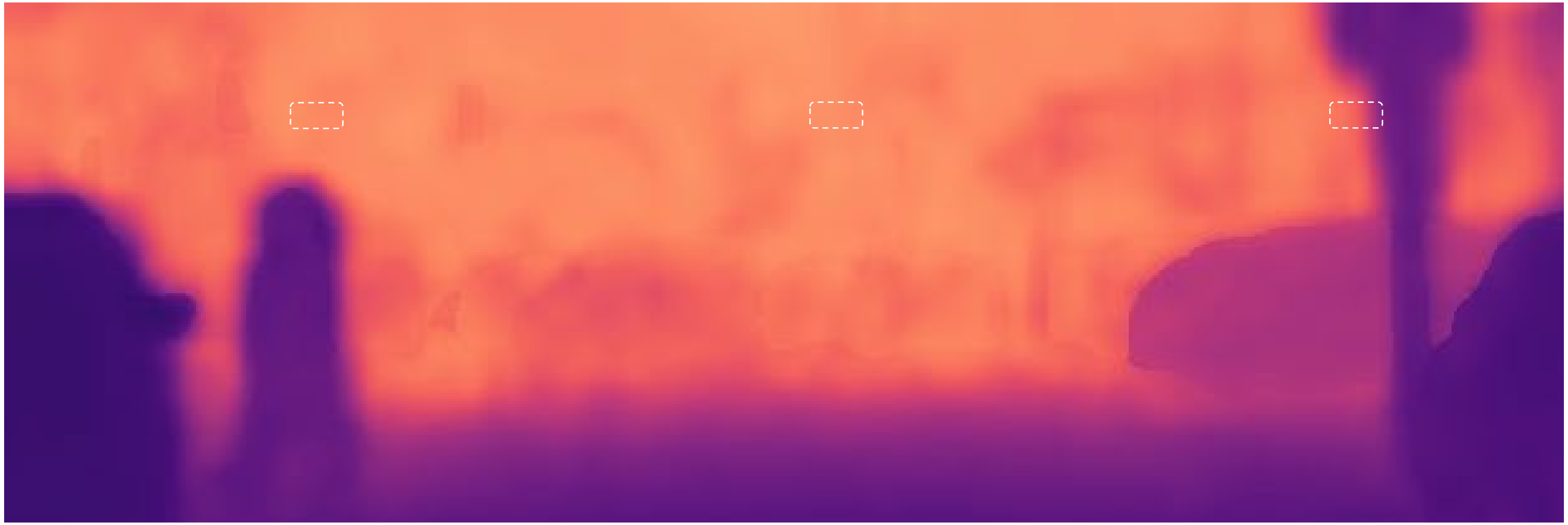
Experiment Results





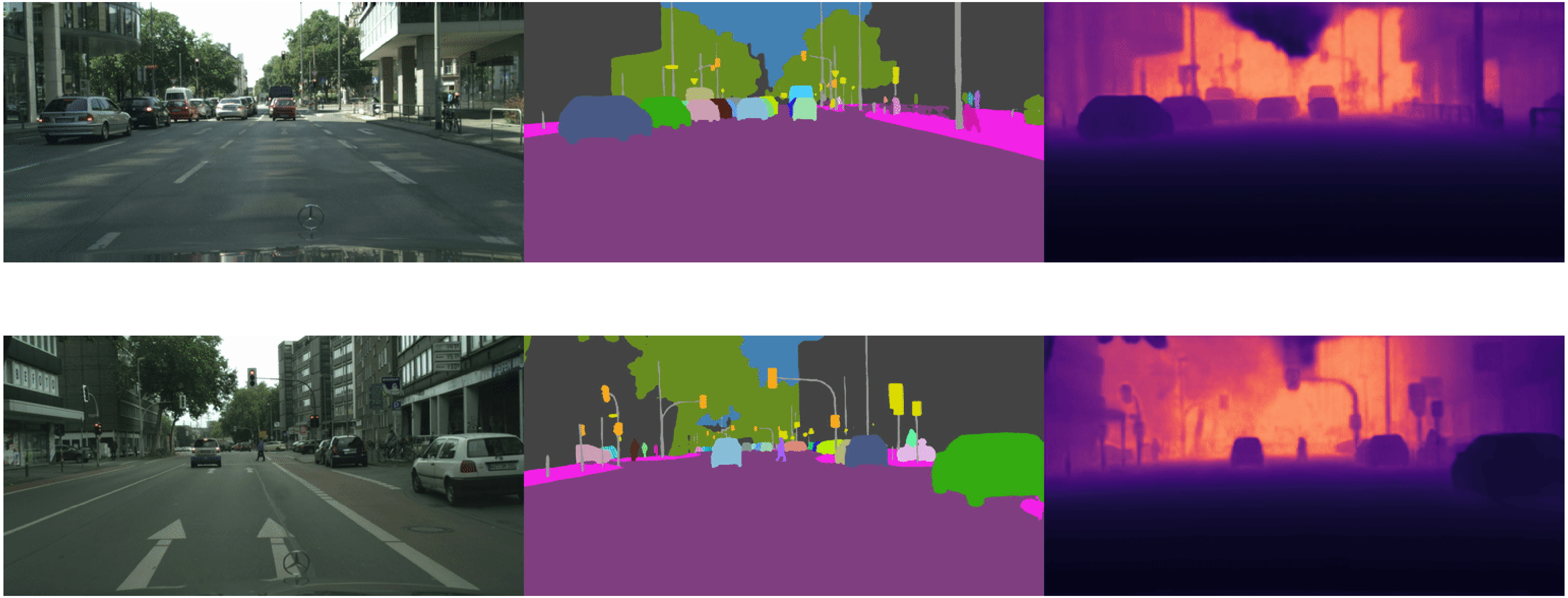
Experiment Results

Instance-level Depth Estimation Paradigm for implicitly leveraging semantic information.



The neural network can better predict the depth map with the semantic information with query learning.

Experiment Results



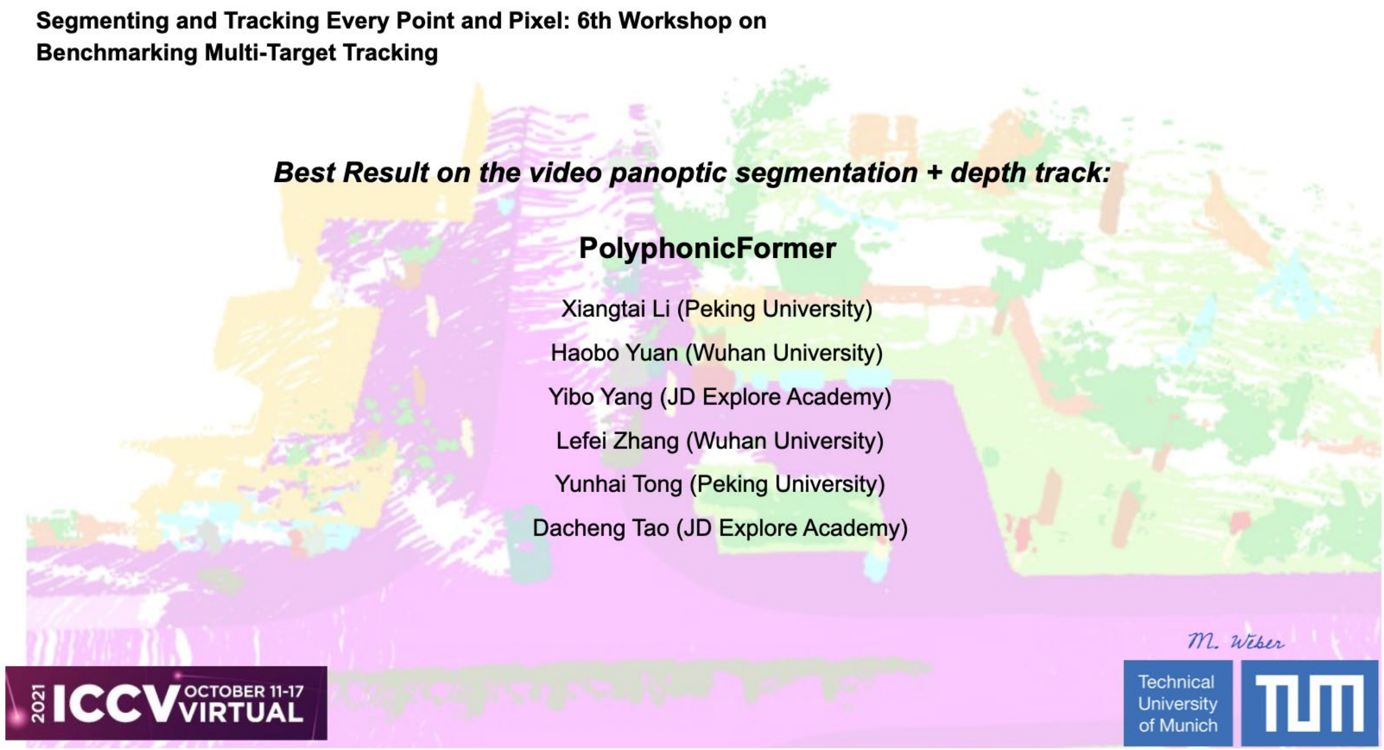
The output of the Depth-aware Video Panoptic Segmentation with PolyphonicFormer.

Experiment Results

Winner of the ICCV-2021 SemanticKITTI DVPS Challenge

Results

#	User	Entries	Date of Last Entry	DSTQ ▲
1	HarborY	10	10/08/21	63.63 (1)
2	ViP-DeepLab	4	07/15/21	63.36 (2)
3	ywang26	8	10/09/21	55.59 (3)
4	rl_lab	3	10/08/21	54.77 (4)



Segmenting and Tracking Every Point and Pixel: 6th Workshop on Benchmarking Multi-Target Tracking

Best Result on the video panoptic segmentation + depth track:

PolyphonicFormer

Xiangtai Li (Peking University)
Haobo Yuan (Wuhan University)
Yibo Yang (JD Explore Academy)
Lefei Zhang (Wuhan University)
Yunhai Tong (Peking University)
Dacheng Tao (JD Explore Academy)

M. Weber

2021 ICCV OCTOBER 11-17 VIRTUAL

Technical University of Munich

What are the Nexts?

Beyond pixel-wised recognition:

+ Geometry (Depth Estimation)

+ Reasoning (Video Scene Graph)

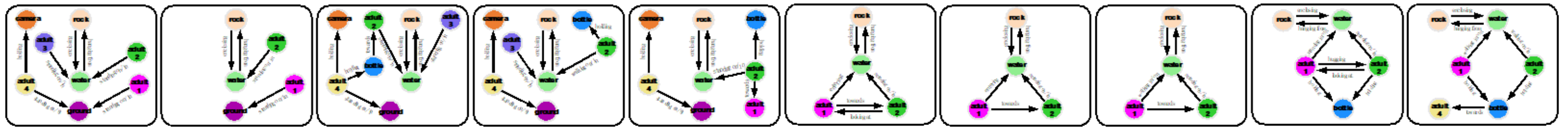
PVSG-dataset

Input: a video sequence

(a) An example video from the PVSG dataset. The PVSG dataset carefully selects 400 first/third-person long videos (avg. 1 min) with clear storyline.

Output: a frame-level panoptic segmentation & video-level scene graph

(b) The mask annotation of the example video. The PVSG dataset has dense (5 fps) and accurate video panoptic segmentation annotation.



(c) The frame-level scene graph representation of the example video. Nodes represent object category and its status. Edges represent relations.

Status (Open Vocabulary)

- adult-1 roll up pants | adult-1 smiling
- adult-2 smiling | adult-2 smiling

Relations (80 Predicate Classes)

- adult-1 standing on/in ground | bottle towards adult-2 | adult-1 towards adult-2 | adult-1 kissing adult-2
- adult-2 standing on/in water | adult-2 walking on/in water | adult-2 standing on/in water
- adult-4 holding camera

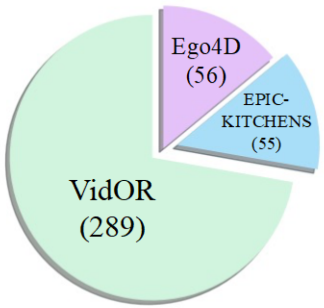
……

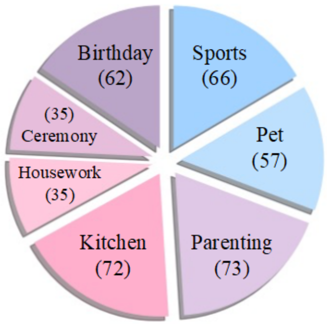
- (Static Relations) water hanging from rock; rock enclosing water; rock in front of tree

(d) The video-level status and relation annotation, which contains interchangeable information of frame-level scene graph in (c).

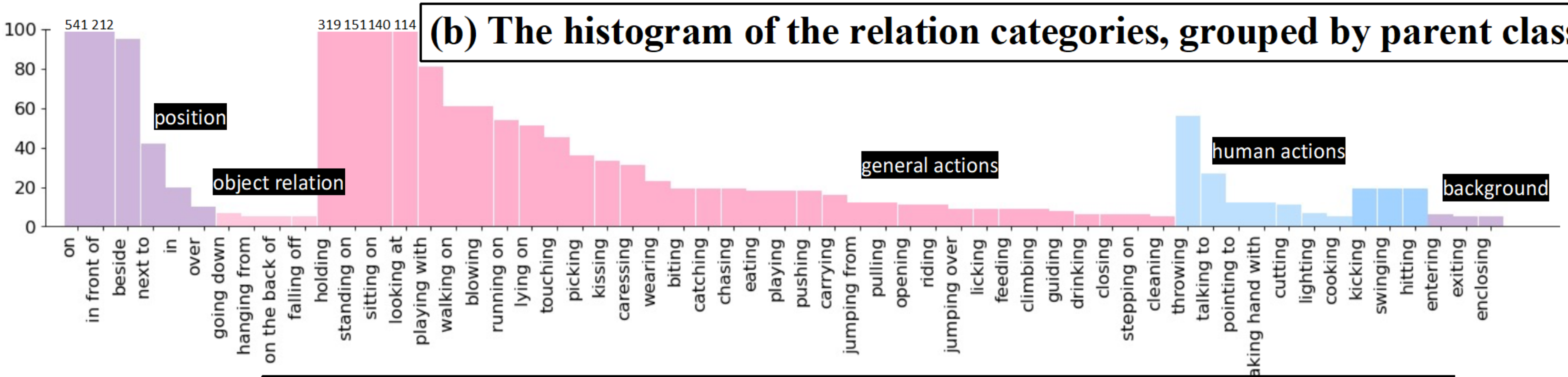
The PVSG task aims to abstract/parse all the information in a video, into a representation of dynamic scene graph, each node is grounded by temporal mask tube.

PVSG-dataset

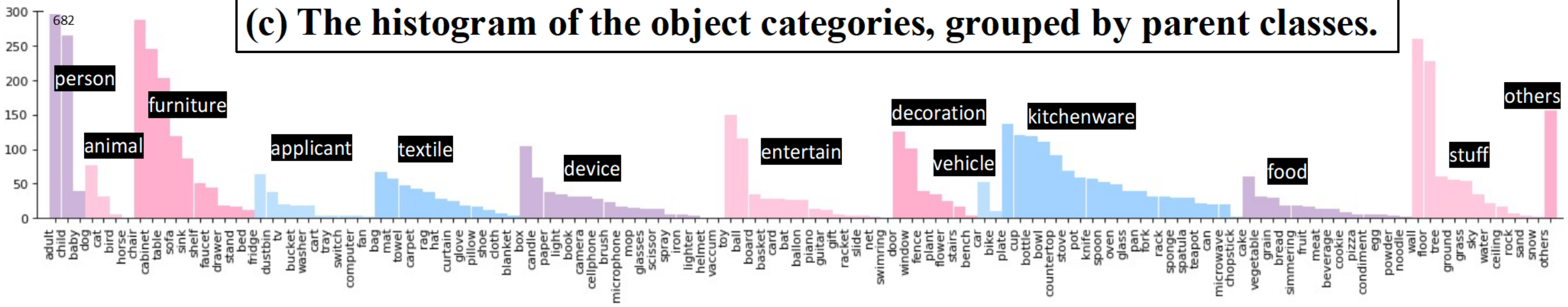




(a)Video Source & Types



(b) The histogram of the relation categories, grouped by parent classes.



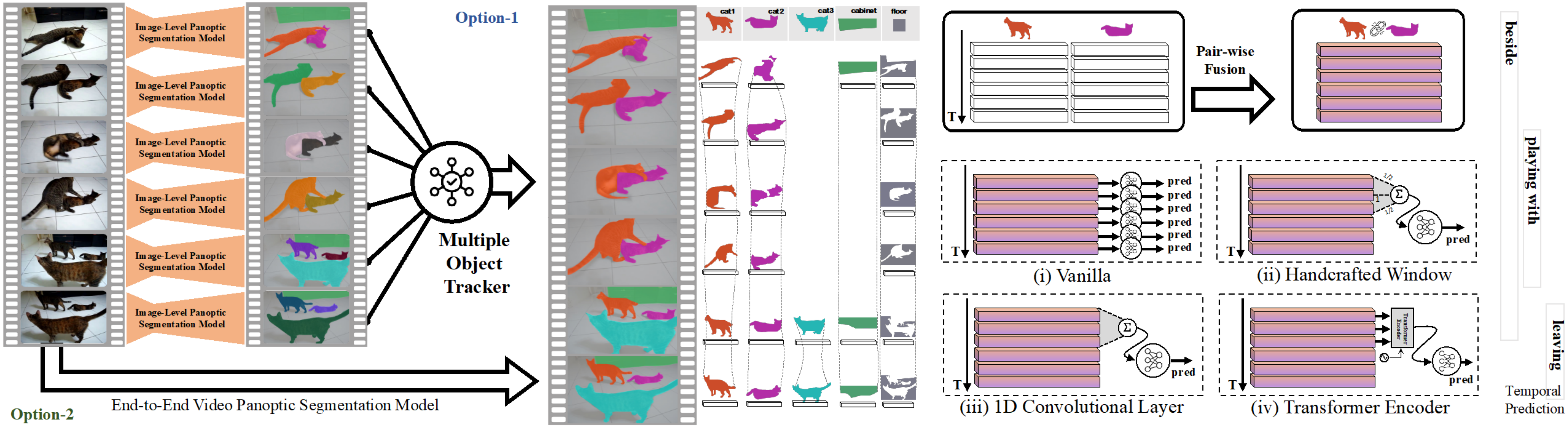
(c) The histogram of the object categories, grouped by parent classes.

Large-Scale (150K frames) **Long Videos** (avg. 77s)

Complex & Dynamic Scene / Multiple Viewpoints

Dense Annotation: Scene Graph, Caption, Conversational Instruction

Panoptic Video Scene Graph Generation-CVPR-2023



(a) Stage-1: For Feature Tube and Mask Tube Output

(b) Stage 2: Relation Prediction

Conclusion

1, TransVOD: End-to-End Video Object Detection with Spatial-Temporal Transformers.
(Video Object Detection)

2, Video K-Net: A Simple, Strong, and Unified Baseline for Video Segmentation
(Video Panoptic Segmentation, online)

3, Tube-Link: A Flexible Cross Tube Baseline for Universal Video Segmentation
(Universal Video Segmentation, semi-online)

Video Perception

4, PolyphonicFormer: Unified Query Learning for Depth-aware Video Panoptic Segmentation
(Unified Transformer For Depth + Panoptic Segmentation)

5, Panoptic Video Scene Graph Generation
(A challenging video pixel-level segmentation and relation detection benchmark)

Video Perception and Beyond

Open Sourced Codebases

We release all codebases of our video research works!!

Video K-Net

Tube-Link

TransVOD

PolyphonicFormer

Future Work

1, Joint Learning with Multi-Modality

2, Generative Segmentation

3, Extremely Long Video Segmentation in Dynamic Scenes

4, Life-Long Learning for Segmentation

5, Video Segmentation in Open Vocabulary Setting

Thanks For Your Watching

Q&A