CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

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







Xiangtai Li* Wenwei Zhang* Jiangmiao Pang* Kai Chen Guangliang Cheng Yunhai Tong Chen Change Loy

PEKING UNIVERSITY 1898

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

商汤 sensetime

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

Motivation

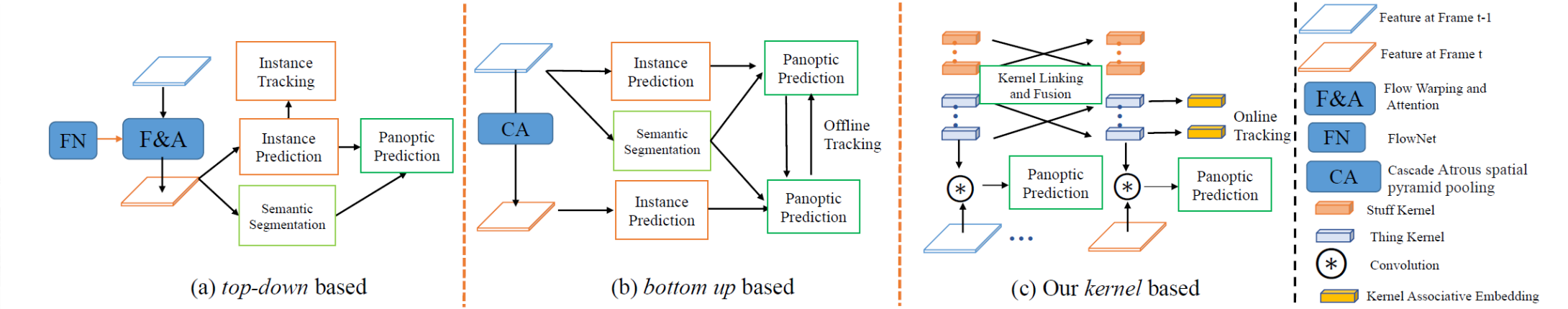


Fig.1 Current Solution For Video Panoptic Segmentation (VPS)

- **Unified Video Segmentation or Video Panoptic Segmentation(VPS):**
  - Complex and hand-crafted Pipeline for VPS.
  - Need the post process or offline tracking.
  - Need Optical flow learning and warping.
  - Tackle the segmentation and tracking with specific task head.

Is there any simpler solution for VPS?

Motivation

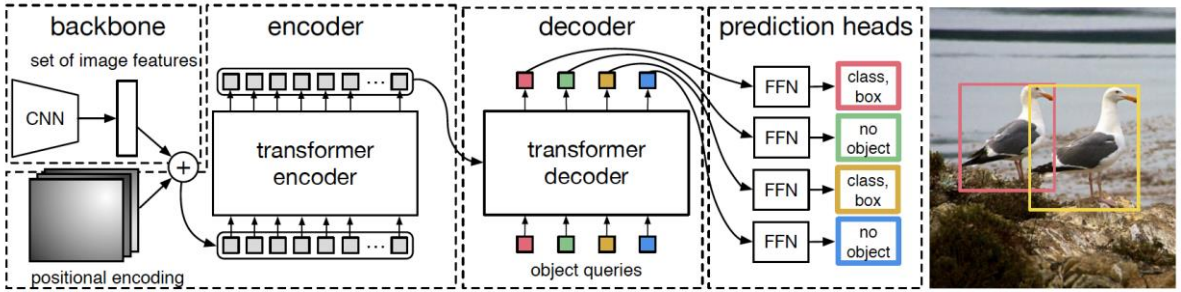
VSS VIS VPS

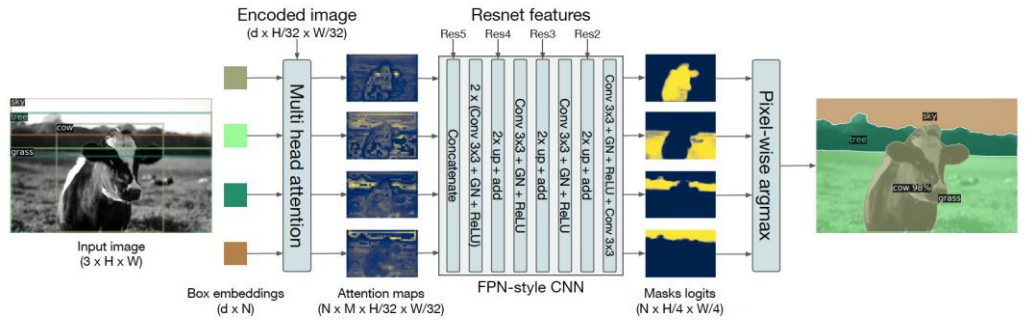
Fig.2 Current Video Segmentation Tasks.

- **Other Video Segmentation Tasks:**
  - Different Tasks have different solutions including specific design. Such as Optical Flow Warping or Clip-Level Transformer.
  - Video Semantic Segmentation (VSS): no instance tracking.
  - Video Instance Segmentation (VIS): no background context.

Is there any unified architecture to solve all video segmentation tasks including VPS, VIS and VSS?

Introduction

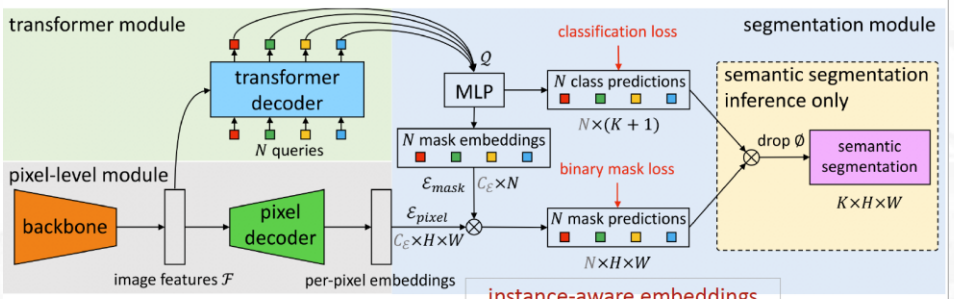




DETR (ECCV-2020-oral)

MaskFormer/K-Net NeurIPS-2021

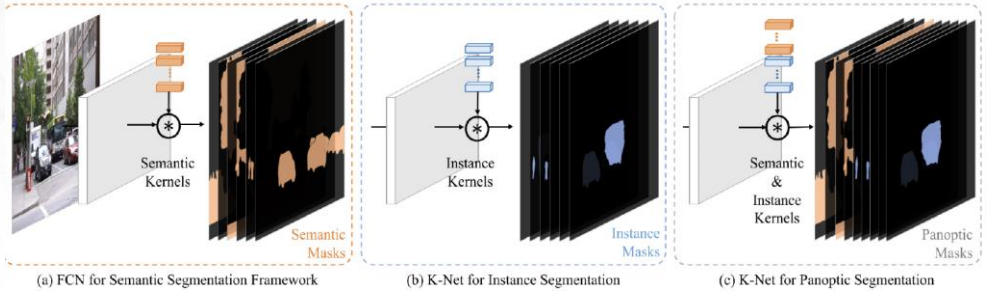
MaskFormer: Per-Pixel Classification is Not All You Need for Semantic Segmentation



Semantic segmentation based on the DETR framework.

instance-aware embeddings
×
instance-agnostic embeddings
(general form of dynamic conv)

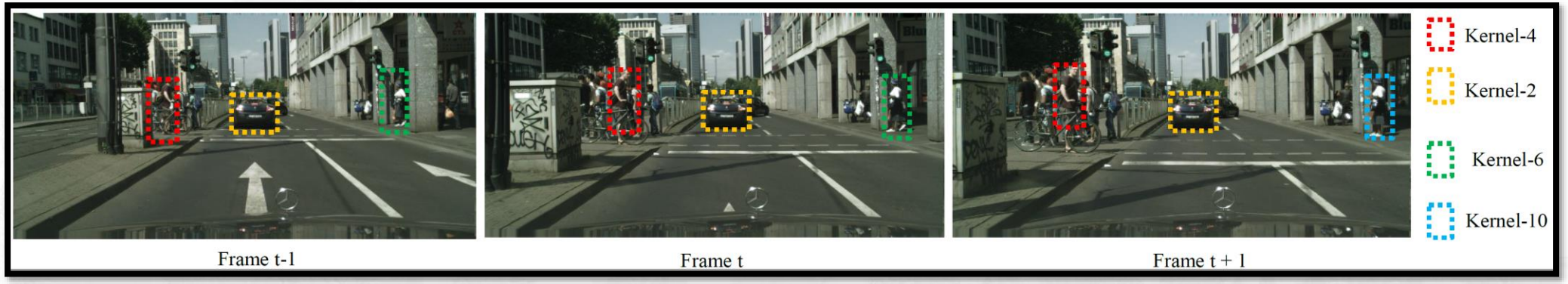
K-Net: Towards Unified Image Segmentation



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

"Instead of generated from dense grids, the kernels in K-Net are a set of learnable parameters **updated** by their corresponding contents in the image."

Introduction



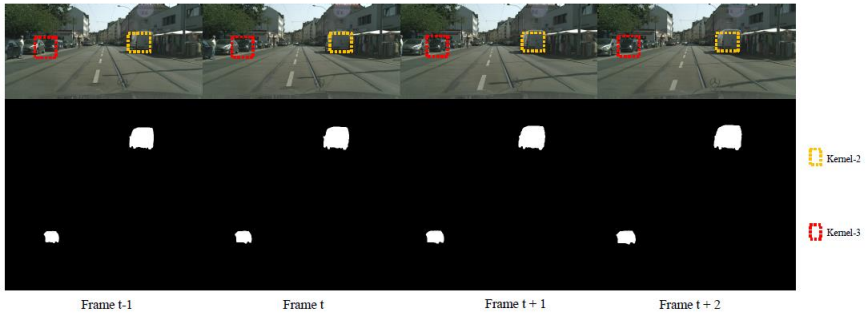


Figure 3. Toy experiment illustration. We use the K-Net directly on Cityscapes video datasets. We find that several instances are originated from **the same kernel** predictions (Red, Yellow boxes, Kernel-2 and Kernel-3). This observation motivates us to use K-Net directly on video. Best view it in color.

Table 1. **T**oy 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.

Method

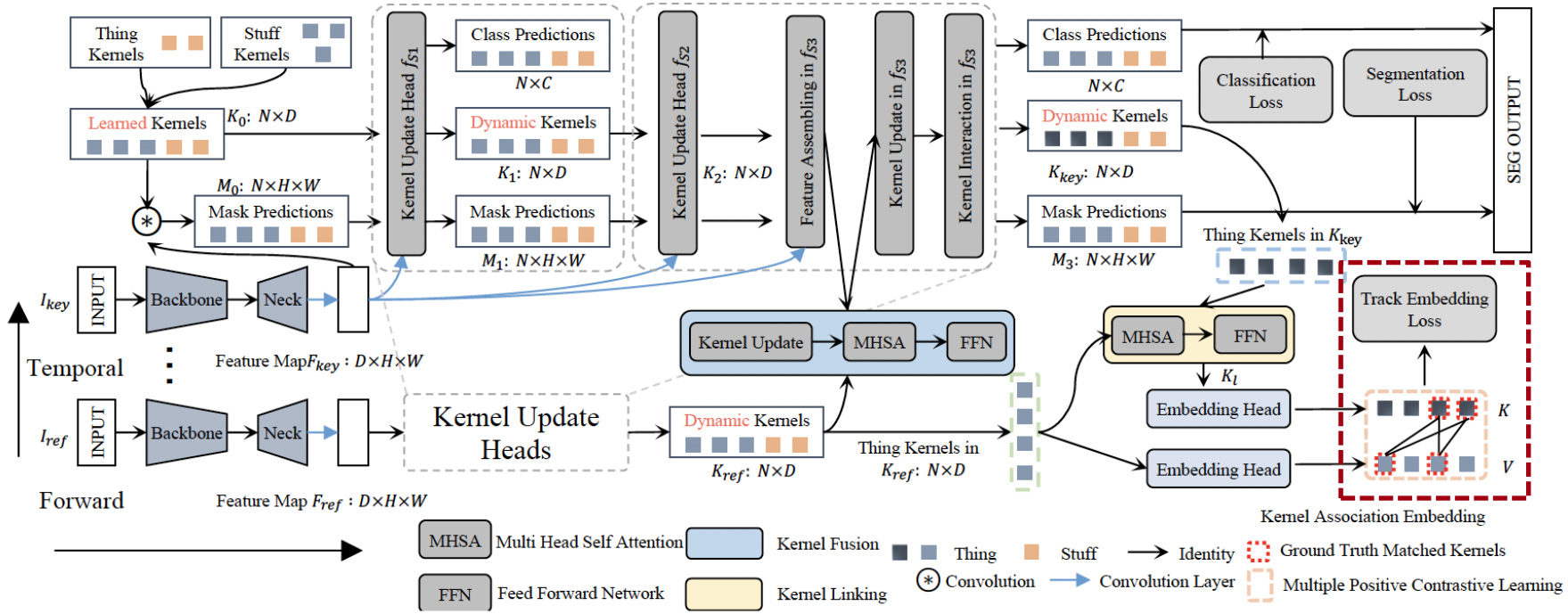


Figure 4. An illustration of our proposed Video K-Net. Our method is based on K-Net [65], which is the top-left part of the figure. Video K-Net adds Kernel Fusion at the start phase of the last stage. The Kernel Linking is performed on the output of dynamic kernels. The Embedding Head is appended at the output of kernel linking and takes kernel outputs from both sampled frames.

➢ **Learning the Kernel Association Embedding**

We propose to use the sparse kernel embedding learning.

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

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

Method

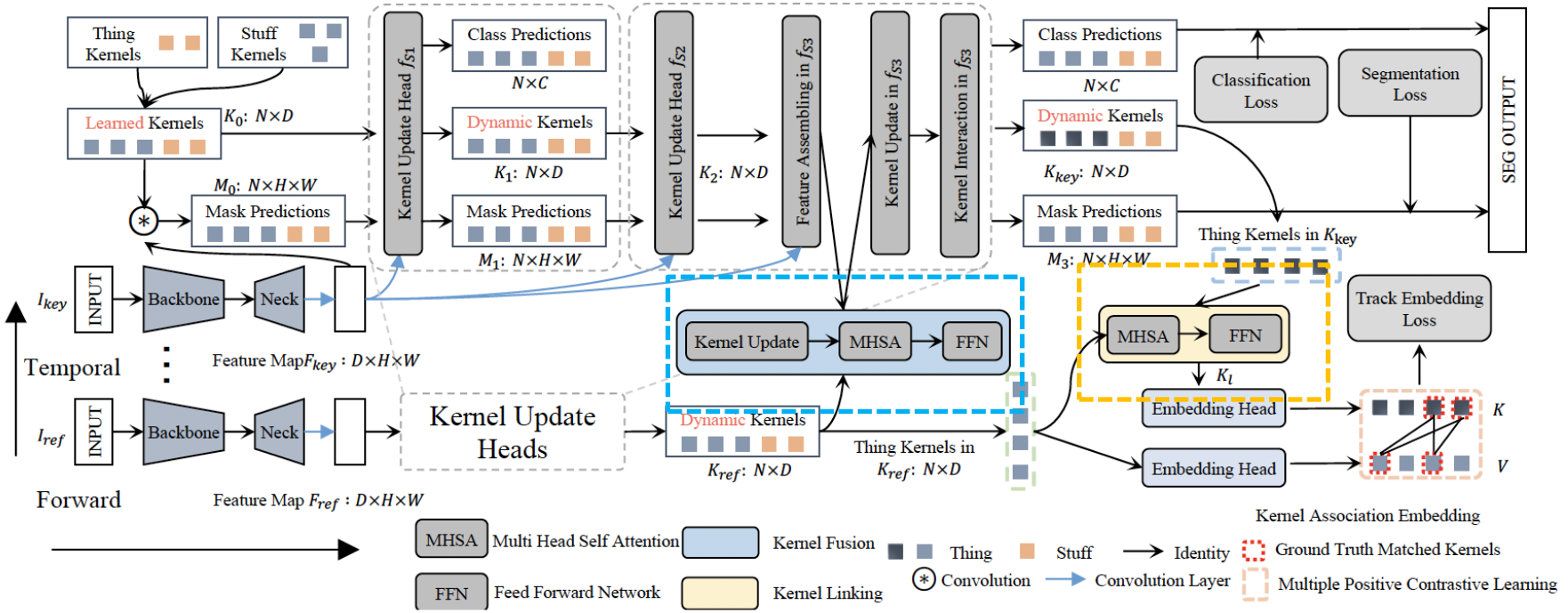


Figure 4. An illustration of our proposed Video K-Net. Our method is based on K-Net [65], which is the top-left part of the figure. Video K-Net adds Kernel Fusion at the start phase of the last stage. The Kernel Linking is performed on the output of dynamic kernels. The Embedding Head is appended at 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.

- **Learning to Fuse Kernels.**

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

- **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 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 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 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	-

Table 5. **Results on Video instance segmentation** AP (%) on the YouTube-VIS-2019 [69] validation dataset. * means using deformable fpn [81]. Axial means using extra Axial Attention [56]. The compared methods are listed by publication date.

Method	backbone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
FEELVOS [54]	ResNet50	26.9	42.0	29.7	29.9	33.4
MaskTrack R-CNN [69]	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 [59]	ResNet50	36.2	59.8	36.9	37.2	42.4
VisTR [59]	ResNet101	40.1	64.0	45.0	38.3	44.9
TubeFormer-DeepLab [23]	ResNet-50 + Axial	38.8	-	-	44.0	51.4
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
Video K-Net	Swin-base*	54.1	79.0	59.5	49.7	59.9

Table 6. **Results on VIPSeg-VPS [37] validation dataset.** We report VPQ and STQ for reference. Following work [37], we report VPQ score at different window sizes (1,2,4,6).

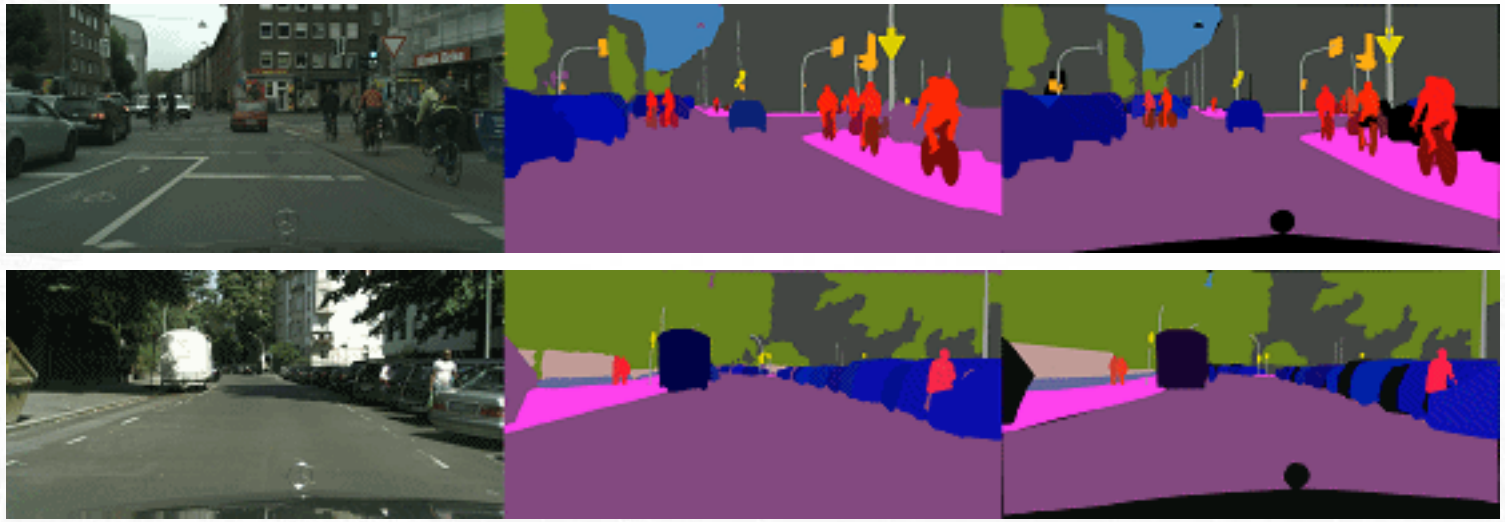
Method	backbone	VPQ^1	VPQ^2	VPQ^4	VPQ^6	VPQ	STQ
VIP-DeepLab [47]	ResNet50	18.4	16.9	14.8	13.7	16.0	22.0
VPSNet [22]	ResNet50	19.9	18.1	15.8	14.5	17.0	20.8
SiamTrack [64]	ResNet50	20.0	18.3	16.0	14.7	17.2	21.1
Clip-PanoFCN [37]	ResNet50	24.3	23.5	22.4	21.6	22.9	31.5
Video K-Net	ResNet50	29.5	26.5	24.5	23.7	26.1	33.1
Video K-Net	Swin-base	43.3	40.5	38.3	37.2	39.8	46.3

New state-of-the art results on All VPS datasets. Even On Recent VIP-Seg dataset(cvpr-2022)

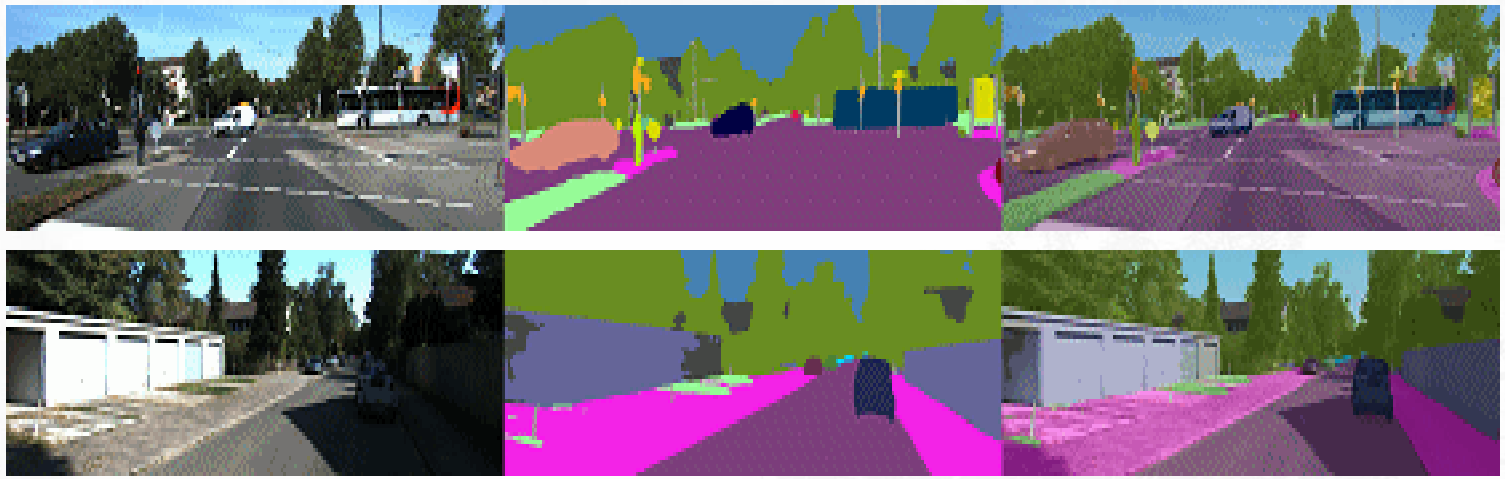
Considerable results on VIS and VSS datasets.

Experiment Results

Short term segmentation and tracking results on Cityscapes VPS dataset.

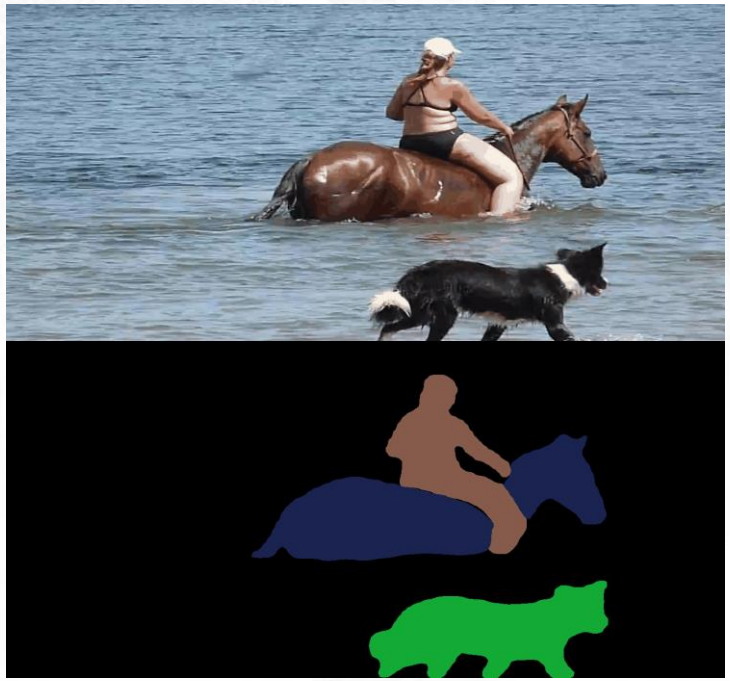


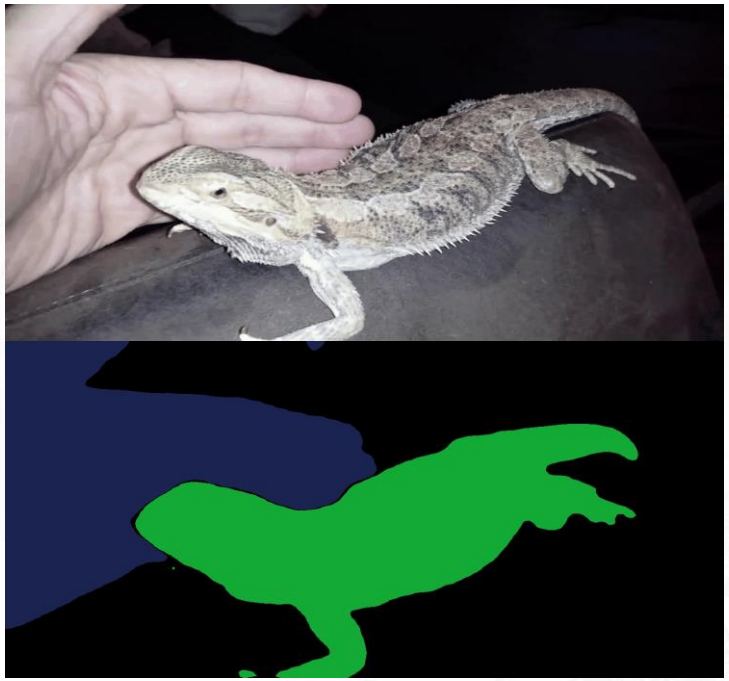
Long term segmentation and tracking results on STEP dataset.



Experiment Results

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





Summary

Thanks For your watching!

Code and Model will be available at:
https://github.com/lxtGH/Video-K-Net