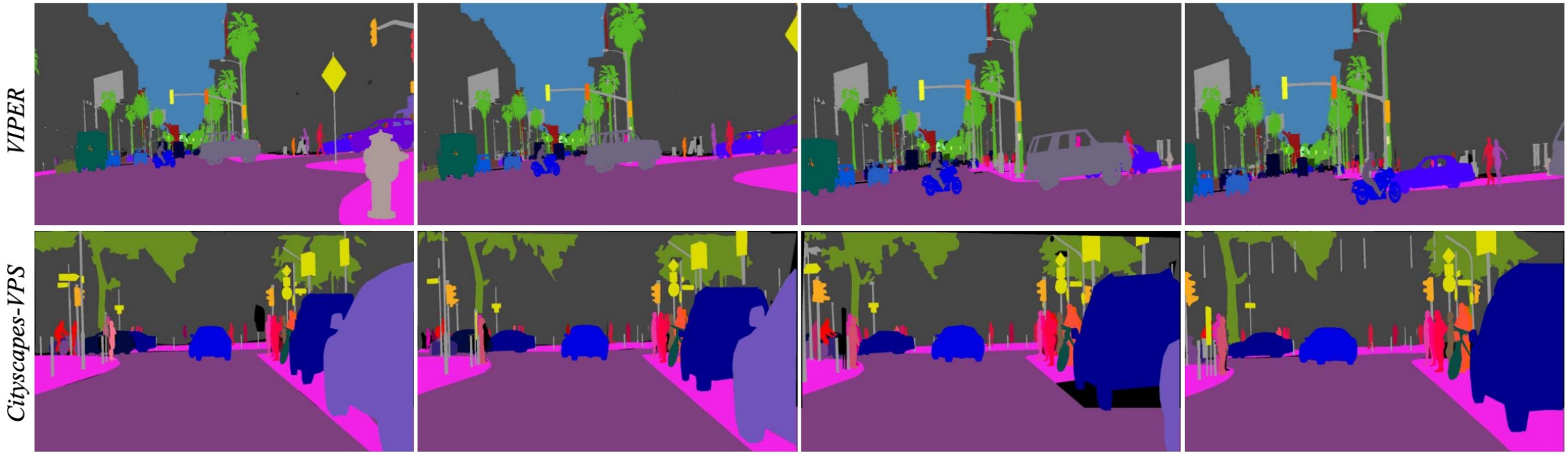
Video K-Net: A Simple, Strong, and Unified Baseline for Video Segmentation

Xiangtai Li[1*] Wenwei Zhang[2*] Jiangmiao Pang[3,5*] Kai Chen [4,5]
Guangliang Cheng[4✉] Yunhai Tong[1✉] Chen Change Loy[2]

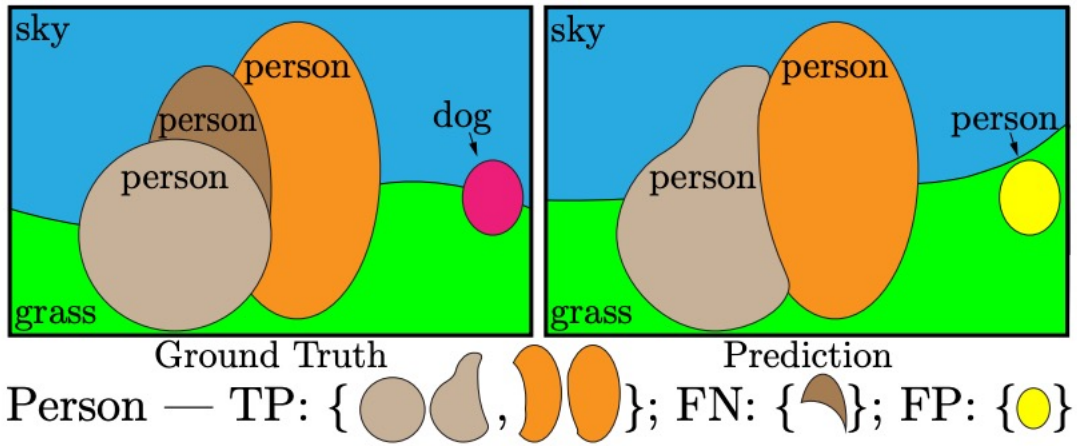
[1] School of Artificial Intelligence, Key Laboratory of Machine Perception (MOE), Peking University
[2] S-Lab, Nanyang Technological University [3] CUHK-SenseTime Joint Lab, the Chinese University of Hong Kong
[4] SenseTime Research [5] Shanghai AI Lab

{lxtpku, yhtong}@pku.edu.cn {wenwei001, ccloy}@ntu.edu.sg
pangjiangmiao@gmail.com {chenkai, chengguangliang}@sensetime.com

First proposed： CVPR 2020 VPSnet



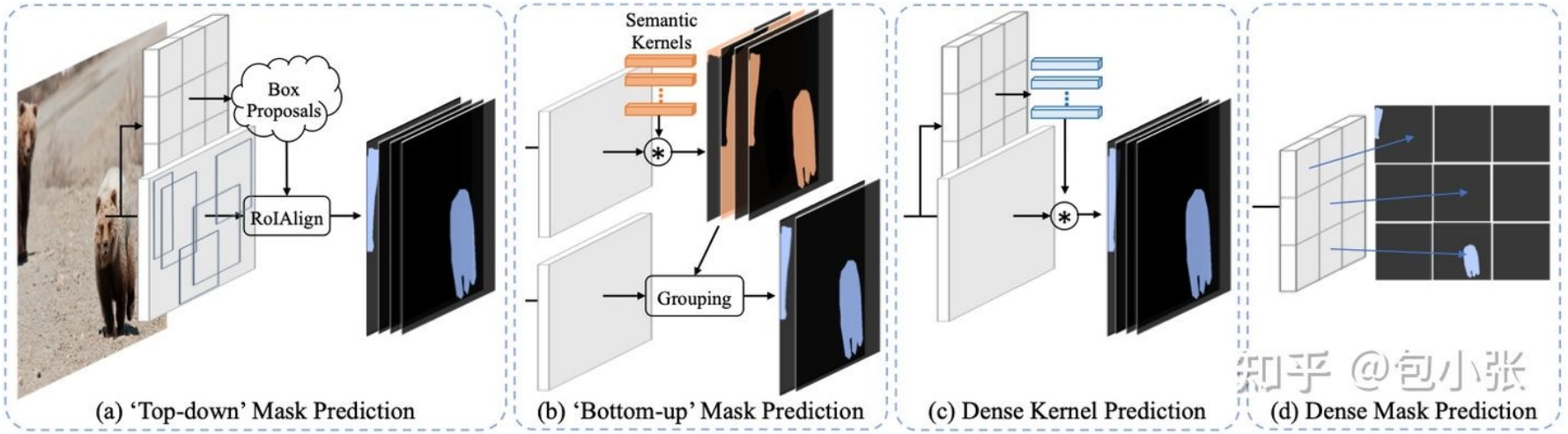
Metric： Video Panoptic Quality (VPQ)、Segmentation and Tracking Quality (STQ)



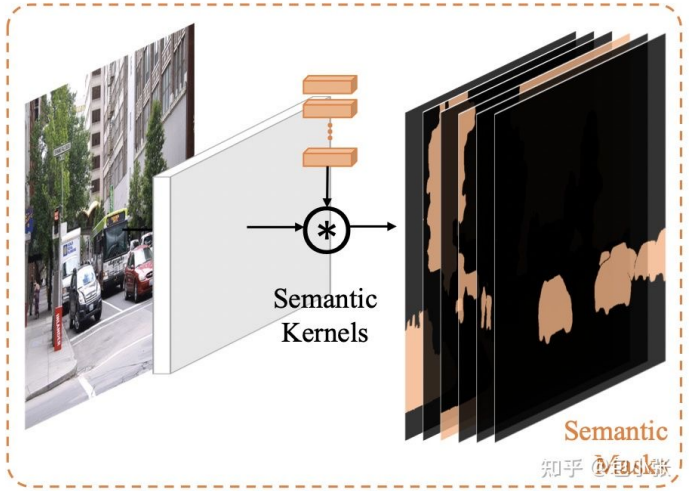
$$\mathrm{PQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}.$$

$$VPQ = \frac{1}{K}\sum_{k}\frac{1}{N_{classes}}\sum_{c}\frac{\sum_{(p,g)\in TP_c^k} IoU_{2D/3D}(p,g)}{|TP_c^k| + \frac{1}{2}|FP_c^k| + \frac{1}{2}|FN_c^k|}$$

Different instance segmentation method：



General Semantic segmentation method：



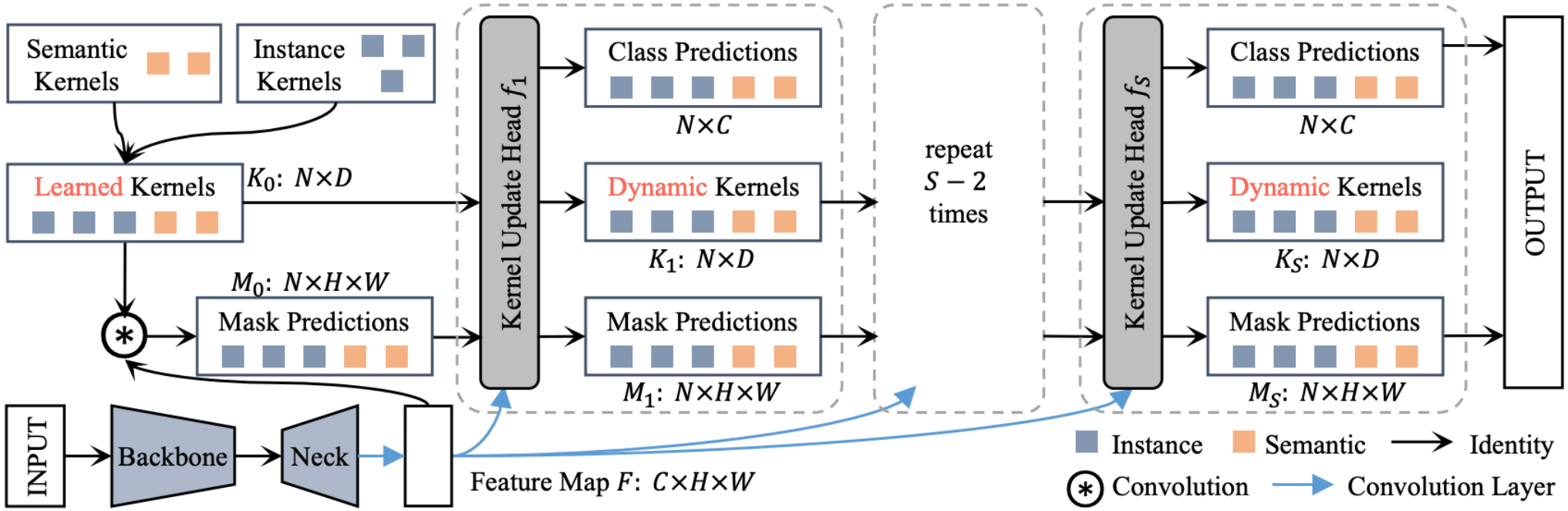


Figure 3: **K-Net for panoptic segmentaion.** A set of learned kernels first performs convolution with the feature map F to predict masks M_0. Then the kernel update head takes the mask predictions M_0, learned kernels K_0, and feature map F as input and produce class predictions, group-aware (dynamic) 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.

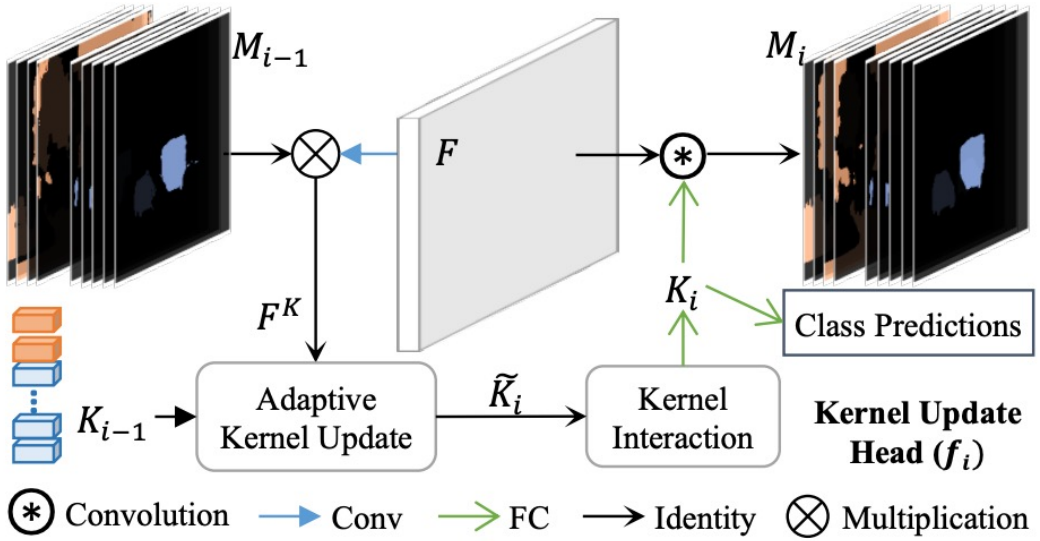


Figure 2: Kernel Update Head.

Feature assembling: $F^K = \sum_u^H \sum_v^W M_{i-1}(u,v) \cdot F(u,v), F^K \in R^{B \times N \times C}$

Kernel update: $F^G = \phi_1(F^K) \otimes \phi_2(K_{i-1}), F^G \in R^{B \times N \times C}$ ϕ_1 and ϕ_2 are linear transformations

$$G^K = \sigma(\psi_1(F^G)), G^F = \sigma(\psi_2(F^G)),$$
$$\tilde{K} = G^F \otimes \psi_3(F^K) + G^K \otimes \psi_4(K_{i-1}),$$

where $\psi_n, n = 1, ..., 4$ are different fully connected (FC) layers followed by LayerNorm (LN) and σ is the Sigmoid function. $\tilde{K}$ is then used in kernel interaction.

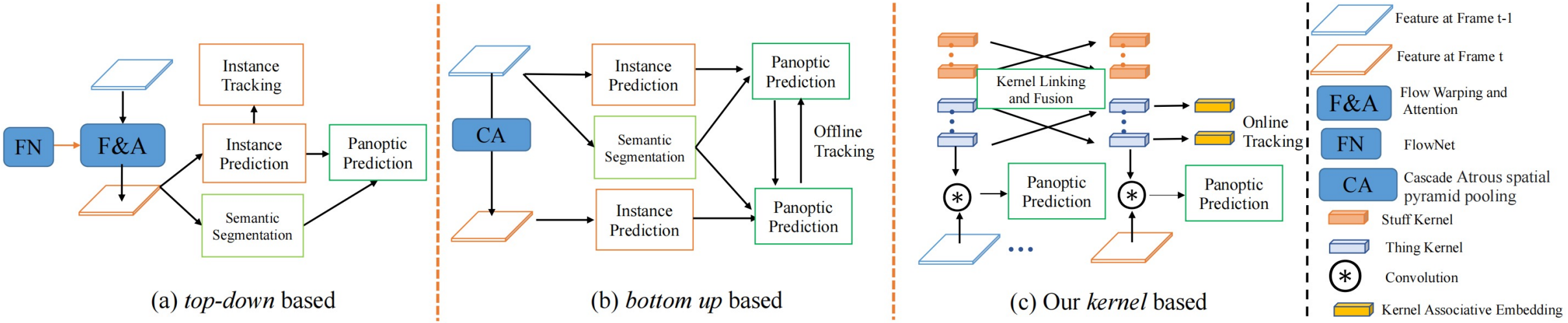


Figure 1. An illustration of previous *top-down* based VPS method (a), *bottom-up* based VPS method (b) and the proposed Video K-Net (c). Unlike previous approaches [22, 43] that perform panoptic segmentation and object tracking with independent modules, our method unifies panoptic segmentation and instance level tracking via kernels in a simpler framework.

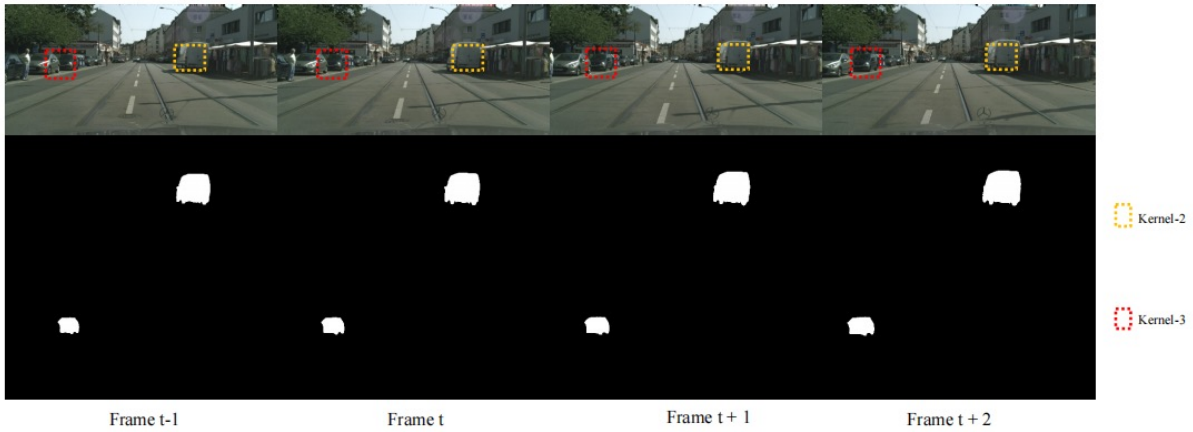


Figure 3. Toy experiment illustration. We use the K-Net directly on Cityscapes video datasets. We find that several instances are originated from **the same kernel** predictions (Red, Yellow boxes, Kernel-2 and Kernel-3). This observation motivates us to use K-Net directly on video. Best view it in color.

Table 1. **T**oy Experiment results on KITTI-STEP and Cityscape-VPS set with STQ and VPQ metrics. Unitrack [56] 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 [56]	ResNet50	65.1	64.3	68.9	-
Cityscapes-VPS	Backbone	-	-	-	VPQ
K-Net	ResNet50	-	-	-	54.3
K-Net + Unitrack [56]	ResNet50	-	-	-	53.2

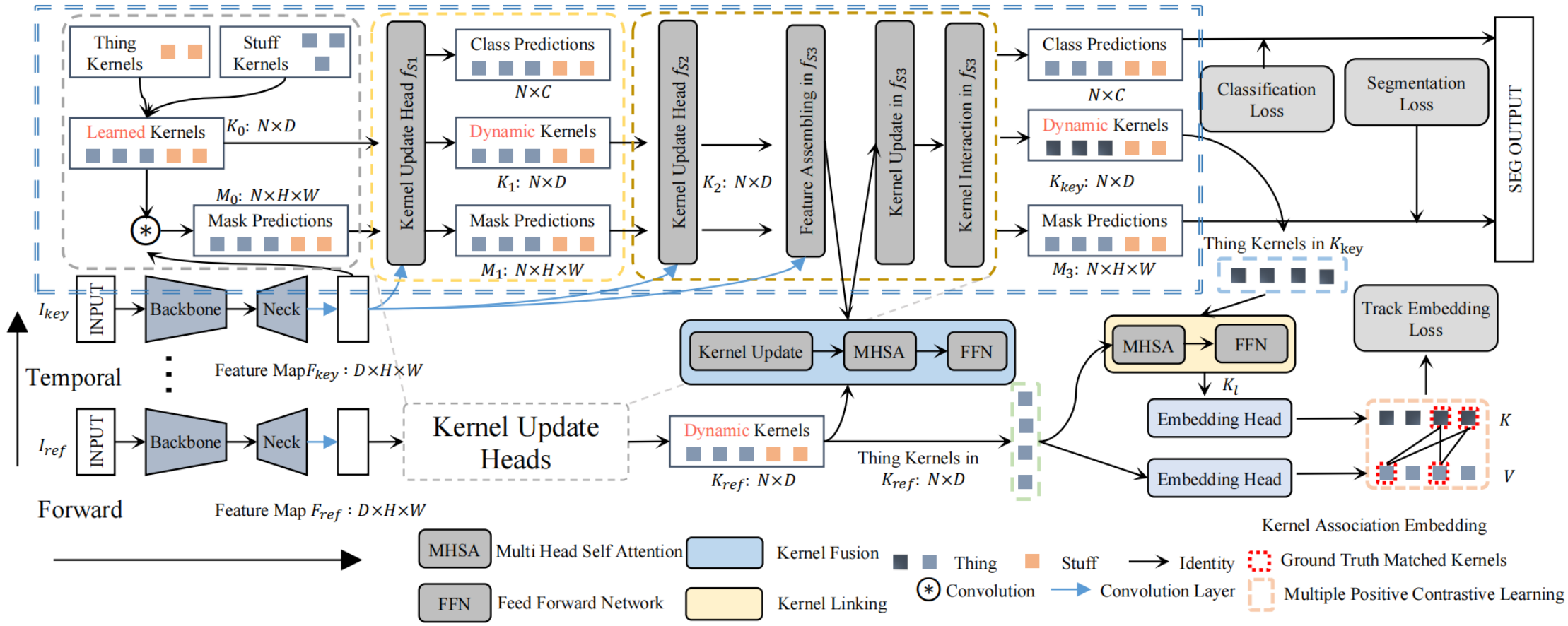
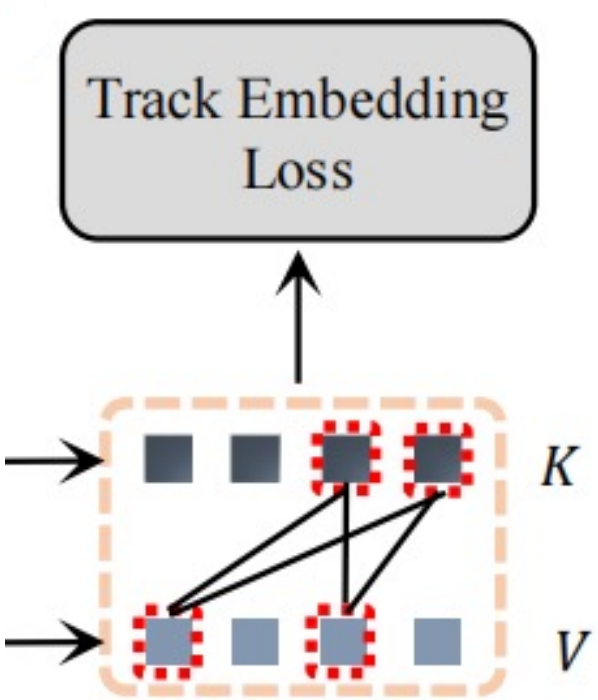


Figure 4. An illustration of our proposed Video K-Net. Our method is based on K-Net [70](in blue dashed box), which is the top-left part of the figure. Video K-Net adds Kernel Fusion at the start phase of the last stage. The Kernel Linking is performed on the output of dynamic kernels. The Embedding Head is appended at the output of kernel linking and takes kernel outputs from both sampled frames.

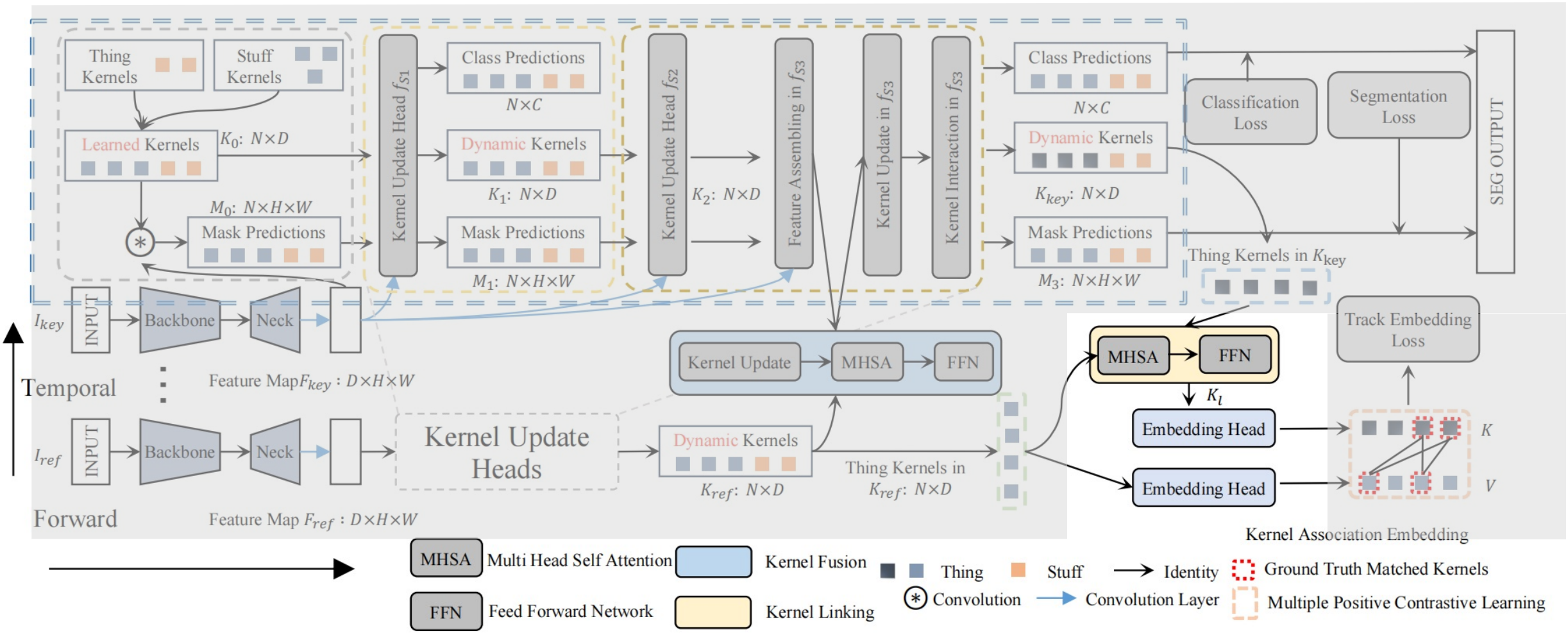
Learning Kernel Association Embeddings



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

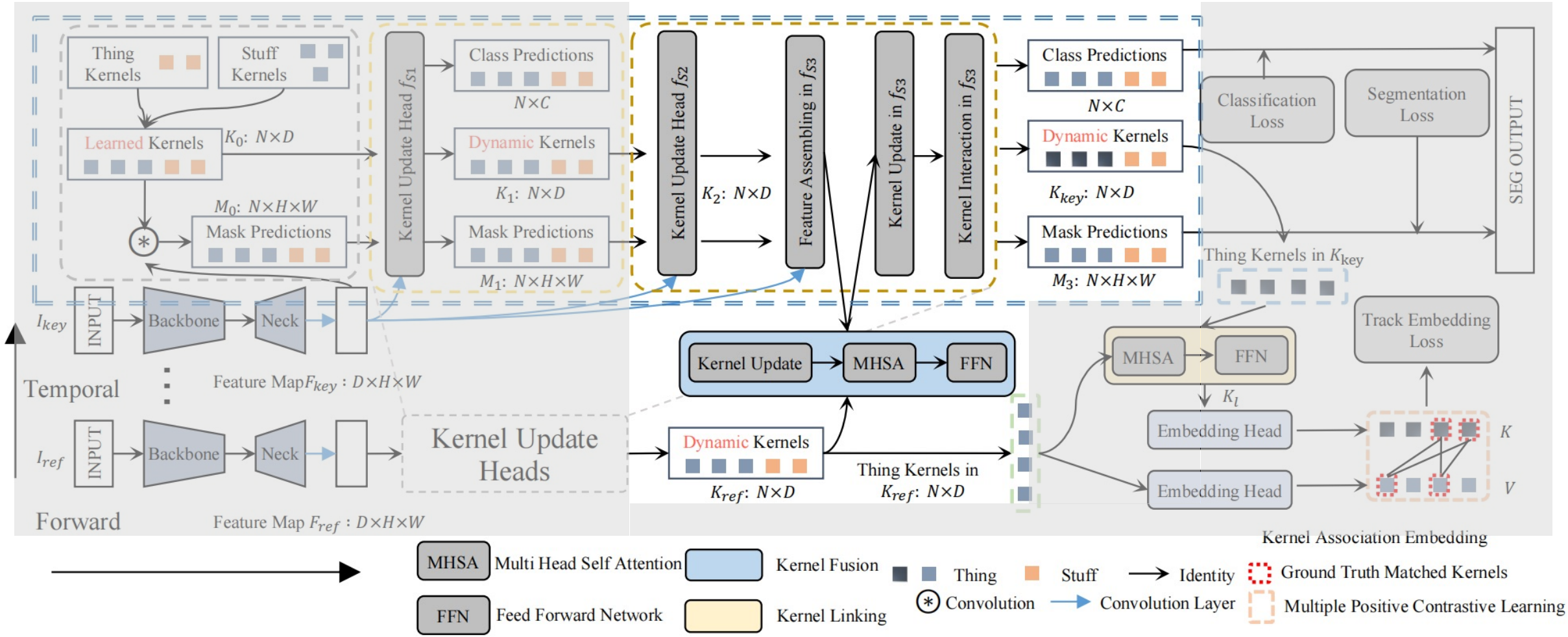
$$\mathcal{L}_{\text{aux}} = \left(\frac{\mathbf{v}\cdot\mathbf{k}}{||\mathbf{v}||\cdot||\mathbf{k}||} - c\right)^2,$$

Learning to Link Kernels



$$K_l = FFN(MHSA(K_{key}, K_{ref}, K_{ref}) + K_{key})$$

Learning to Fuse Kernels



Inference

Bidirectional softmax:

$$\mathbf{f}(i,j) = \left[\frac{\exp\left(\mathbf{n}_i \cdot \mathbf{m}_j\right)}{\sum_{k=0}^{M-1} \exp\left(\mathbf{n}_i \cdot \mathbf{m}_k\right)} + \frac{\exp\left(\mathbf{n}_i \cdot \mathbf{m}_j\right)}{\sum_{k=0}^{N-1} \exp\left(\mathbf{n}_k \cdot \mathbf{m}_j\right)} \right] /2$$

Experiment

KITTI-STEP validation and test set

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 [58]	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		0.74	0.72	0.73	0.53
Motion-Deeplab [58]	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	-

Experiment

Cityscapes-VPS validation set

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 [60]	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 [43]	WideResNet41 [68]	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 [43]	WideResNet41 [68]+RFP [42] + 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 [31]	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 [42]	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

VSPW validation set

VPSW	Backbone	mIoU	mVC_8	mVC_{16}
DeepLabv3+ [8]	ResNet101	35.7	83.5	78.4
PSPNet+ [72]	ResNet101	36.5	84.4	79.8
TCB(PSPNet) [34]	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

Experiment

YTVIS-2019 validation set

Method	backbone	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
FEELVOS [50]	ResNet50	26.9	42.0	29.7	29.9	33.4
MaskTrack R-CNN [65]	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 [55]	ResNet50	34.4	55.7	36.5	33.5	38.9
VisTR [55]	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

Ablation study

(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 [37]	68.8	69.1
Mask-Emb [60]	67.3	68.1
Ours	70.8	70.9
Ours + Mask-Emb [60]	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 [37]	63.1	62.1	64.3

(d) Ablation Study on Linking and Fusing Stage.

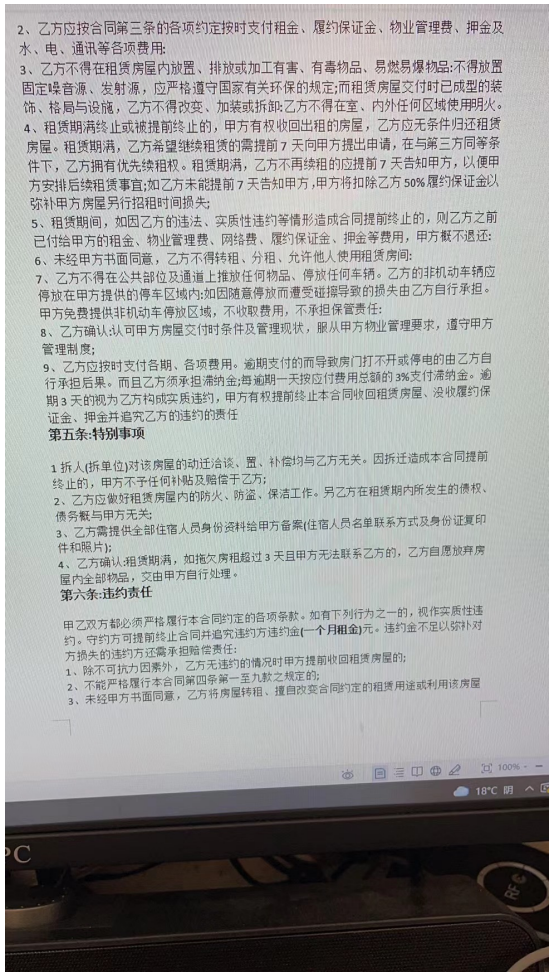
Stage	STQ	AQ	SQ
3	70.9	70.8	71.2
2	68.5	68.2	69.3
1	66.9	63.4	67.3

(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



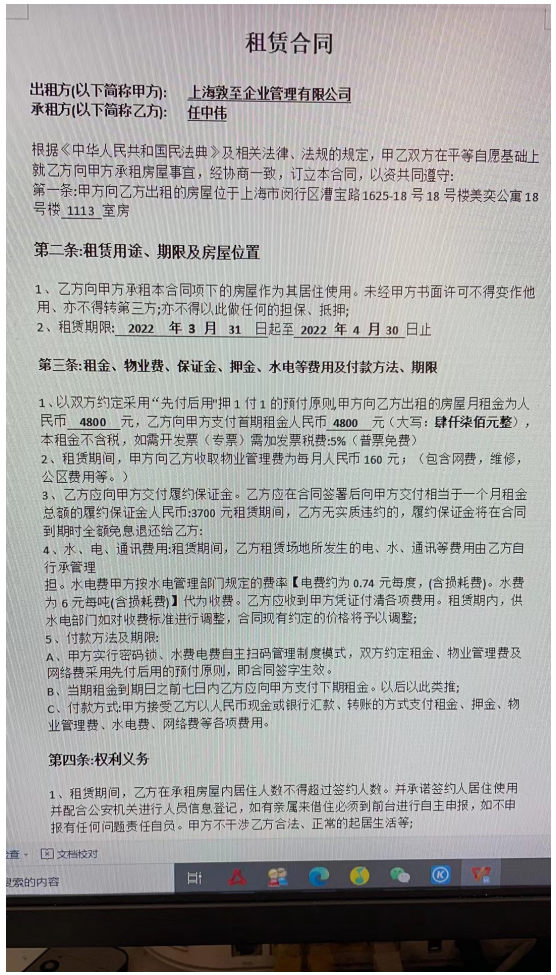
2、乙方应按合同第三条的各项约定按时支付租金、履约保证金、物业管理费、押金及水、电、通讯等各项费用;
3、乙方不得在租赁房屋内放置、排放或加工有害、有毒物品、易燃易爆物品;不得放置固定噪音源、发射源，应严格遵守国家有关环保的规定;而租赁房屋交付时已成型的装饰、格局与设施，乙方不得改变、加装或拆卸;乙方不得在室、内外任何区域使用明火。
4、租赁期满终止或被提前终止的，甲方有权收回出租的房屋，乙方应无条件归还租赁房屋。租赁期满，乙方希望继续租赁的需提前 7 天向甲方提出申请，在与第三方同等条件下，乙方拥有优先续租权。租赁期满，乙方不再续租的应提前 7 天告知甲方，以便甲方安排后续租赁事宜;如乙方未能提前 7 天告知甲方，甲方将扣除乙方 50%履约保证金以弥补甲方房屋另行招租时间损失;
5、租赁期间，如因乙方的违法、实质性违约等情形造成合同提前终止的，则乙方之前已付给甲方的租金、物业管理费、网络费、履约保证金、押金等费用，甲方概不退还;
6、未经甲方书面同意，乙方不得转租、分租、允许他人使用租赁房间;
7、乙方不得在公共部位及通道上摊放任何物品、停放任何车辆。乙方的非机动车辆应停放在甲方提供的停车区域内;如因随意停放而遭受碰撞导致的损失由乙方自行承担。甲方免费提供非机动车停放区域，不收取费用，不承担保管责任;
8、乙方确认:认可甲方房屋交付时条件及管理现状，服从甲方物业管理要求，遵守甲方管理制度;
9、乙方应按时支付各期、各项费用。逾期支付的而导致房门打不开或停电的由乙方自行承担后果。而且乙方须承担滞纳金;每逾期一天按应付费用总额的 3%支付滞纳金。逾期 3 天的视为乙方构成实质违约，甲方有权提前终止本合同收回租赁房屋、没收履约保证金、押金并追究乙方的违约的责任

第五条:特别事项

1 拆人(拆单位)对该房屋的动迁洽谈、置、补偿均与乙方无关。因拆迁造成本合同提前终止的，甲方不予任何补贴及赔偿于乙方;
2、乙方应做好租赁房屋内的防火、防盗、保洁工作。另乙方在租赁期内所发生的债权、债务概与甲方无关;
3、乙方需提供全部住宿人员身份资料给甲方备案(住宿人员名单联系方式及身份证复印件和照片);
4、乙方确认:租赁期满，如拖欠房租超过 3 天且甲方无法联系乙方的，乙方自愿放弃房屋内全部物品，交由甲方自行处理。

第六条:违约责任

甲乙双方都必须严格履行本合同约定的各项条款。如有下列行为之一的，视作实质性违约。守约方可提前终止合同并追究违约方违约金**(一个月租金)**元。违约金不足以弥补对方损失的违约方还需承担赔偿责任:
1、除不可抗力因素外，乙方无违约的情况时甲方提前收回租赁房屋的;
2、不能严格履行本合同第四条第一至九款之规定的;
3、未经甲方书面同意，乙方将房屋转租、擅自改变合同约定的租赁用途或利用该房屋



租赁合同

出租方(以下简称甲方): 上海敦至企业管理有限公司
承租方(以下简称乙方): 任中伟

根据《中华人民共和国民法典》及相关法律、法规的规定，甲乙双方在平等自愿基础上就乙方向甲方承租房屋事宜，经协商一致，订立本合同，以资共同遵守:
第一条:甲方向乙方出租的房屋位于上海市闵行区漕宝路1625-18 号 18 号楼美奕公寓 18 号楼 1113 室房

第二条:租赁用途、期限及房屋位置

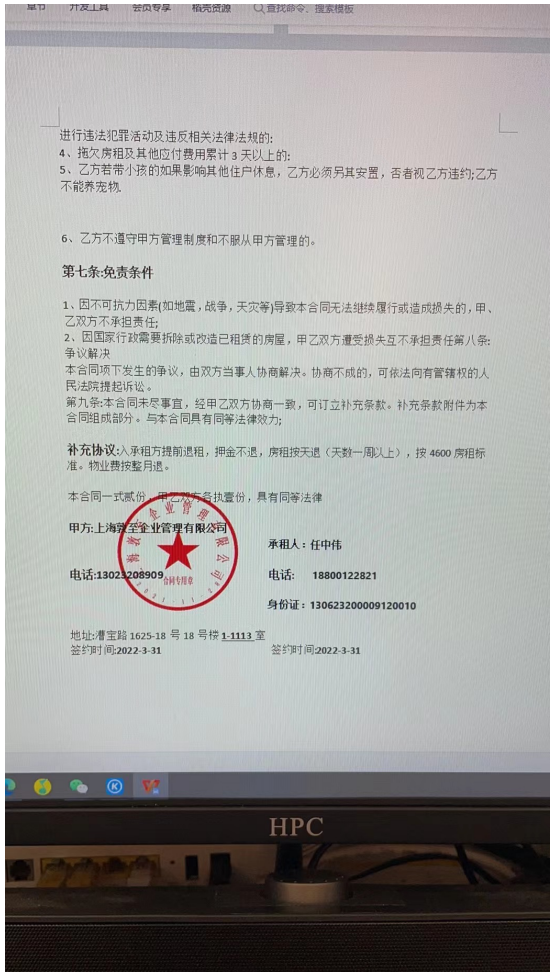
1、乙方向甲方承租本合同项下的房屋作为其居住使用。未经甲方书面许可不得变作他用、亦不得转第三方;亦不得以此做任何的担保、抵押;
2、租赁期限: 2022 年 3 月 31 日起至 2022 年 4 月 30 日止

第三条:租金、物业费、保证金、押金、水电等费用及付款方法、期限

1、以双方约定采用"先付后用"押 1 付 1 的预付原则,甲方向乙方出租的房屋月租金为人民币 4800 元，乙方向甲方支付首期租金人民币 4800 元（大写：**肆仟柒佰元整**），本租金不含税，如需开发票（专票）需加发票税费:5%（普票免费）
2、租赁期间，甲方向乙方收取物业管理费为每月人民币 160 元；（包含网费，维修，公区费用等。）
3、乙方应向甲方交付履约保证金。乙方应在合同签署后向甲方交付相当于一个月租金总额的履约保证金人民币:3700 元租赁期间，乙方无实质违约的，履约保证金将在合同到期时全额免息退还给乙方;
4、水、电、通讯费用:租赁期间，乙方租赁场地所发生的电、水、通讯等费用由乙方自行承管理
担。水电费甲方按水电管理部门规定的费率【电费约为 0.74 元每度，(含损耗费)。水费为 6 元每吨(含损耗费)】代为收费。乙方应收到甲方凭证付清各项费用。租赁期内，供水电部门如对收费标准进行调整，合同现有约定的价格将予以调整;
5、付款方法及期限:
A、甲方实行密码锁、水费电费自主扫码管理制度模式，双方约定租金、物业管理费及网络费采用先付后用的预付原则，即合同签字生效。
B、当期租金到期日之前七日内乙方应向甲方支付下期租金。以后以此类推;
C、付款方式:甲方接受乙方以人民币现金或银行汇款、转账的方式支付租金、押金、物业管理费、水电费、网络费等各项费用。

第四条:权利义务

1、租赁期间，乙方在承租房屋内居住人数不得超过签约人数。并承诺签约人居住使用并配合公安机关进行人员信息登记，如有亲属来借住必须到前台进行自主申报，如不申报有任何问题责任自负。甲方不干涉乙方合法、正常的起居生活等;



进行违法犯罪活动及违反相关法律法规的;
4、拖欠房租及其他应付费用累计 3 天以上的;
5、乙方若带小孩的如果影响其他住户休息，乙方必须另其安置，否者视乙方违约;乙方不能养宠物.

6、乙方不遵守甲方管理制度和不服从甲方管理的。

第七条:免责条件

1、因不可抗力因素(如地震，战争，天灾等)导致本合同无法继续履行或造成损失的，甲、乙双方不承担责任;
2、因国家行政需要拆除或改造已租赁的房屋，甲乙双方遭受损失互不承担责任第八条:争议解决
本合同项下发生的争议，由双方当事人协商解决。协商不成的，可依法向有管辖权的人民法院提起诉讼。
第九条:本合同未尽事宜，经甲乙双方协商一致，可订立补充条款。补充条款附件为本合同组成部分。与本合同具有同等法律效力;

补充协议:入承租方提前退租，押金不退，房租按天退（天数一周以上），按 4600 房租标准。物业费按整月退。

本合同一式贰份，甲乙双方各执壹份，具有同等法律

甲方:上海敦至企业管理有限公司

电话:13023208909

承租人：任中伟

电话: 18800122821

身份证：130623200009120010

地址:漕宝路 1625-18 号 18 号楼 1-1113 室
签约时间:2022-3-31　　签约时间:2022-3-31