STC: Spatio-Temporal Contrastive Learning for Video Instance Segmentation

Zhengkai Jiang[1⋆], Zhangxuan Gu[2*], Jinlong Peng[1], Hang Zhou[3], Liang Liu[1], Yabiao Wang[1], Ying Tai[1], Chengjie Wang[1], and Liqing Zhang[2]

[1] Tencent Youtu Lab
[2] Shanghai Jiao Tong University
[3] The Chinese University of Hong Kong
{zhengkjiang, caseywang}@tencent.com

Abstract. Video Instance Segmentation (VIS) is a task that simultaneously requires classification, segmentation, and instance association in a video. Recent VIS approaches rely on sophisticated pipelines to achieve this goal, including RoI-related operations or 3D convolutions. In contrast, we present a simple and efficient single-stage VIS framework based on the instance segmentation method CondInst by adding an extra tracking head. To improve instance association accuracy, a novel bi-directional spatio-temporal contrastive learning strategy for tracking embedding across frames is proposed. Moreover, an instance-wise temporal consistency scheme is utilized to produce temporally coherent results. Experiments conducted on the YouTube-VIS-2019, YouTube-VIS-2021, and OVIS-2021 datasets validate the effectiveness and efficiency of the proposed method. We hope the proposed framework can serve as a simple and strong baseline for other instance-level video association tasks.

Keywords: Video Instance Segmentation, Spatio-Temporal Contrastive Learning, Temporal Consistency

1 Introduction

While significant progress has been made in instance segmentation [14, 4, 43, 28, 47, 41, 42, 37, 24, 30] with the development of deep neural networks, less attention has been paid to its challenging variant in the video domain. The video instance segmentation (VIS) [50, 52, 44, 17] task requires not only classifying and segmenting instances but also capturing the instance associations across frames. Such technology can benefit a great variety of scenarios, *e.g.,* video editing, video surveillance, autonomous driving, and augmented reality. As a result, it is in great need of accurate, robust, and fast video instance segmentation approach in practice.

Previous researchers have developed sophisticated pipelines for tackling this problem [51, 39, 50, 5, 2, 44, 1, 12]. Generally speaking, previous studies can be divided into the categories of two-stage [51, 39, 50, 2, 12], feature-aggregation [27,

⋆ Equal contributions. This work was done while Zhangxuan Gu was interning at Tencent Youtu Lab.

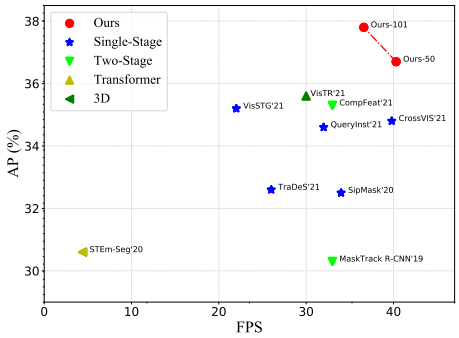

Fig. 1. Speed-Accuracy trade-off curve on the YouTube-VIS-2019 validation set. The baseline results are compared with the same ResNet-50 backbone for fair comparison. We achieve best tradeoff between speed and accuracy. In particular, STC exceeds recent CrossVIS [11] 1.9% mAP with similar running speed.

2] inspired from video object detection domain [55, 20, 19], 3D convolution-based [1], transformer-based [44, 17], and single-stage [5, 52] methods. Two-stage methods, *e.g.*, MaskTrack R-CNN [50] and CompFeat [12], usually rely on the RoIAlign operation to crop the feature and obtain the representation of an instance for further binary mask prediction. Such the RoIAlign operation would lead to great computational inefficiency. 3D convolution-based STEm-Seg [1] holds huge complexity and could not achieve good performance. Transformer-based VisTR [44, 17] could not handle long videos due to largely increasing memory usage and needs a much longer training time for convergence. Feature-aggregation methods [23, 40] enhance features through pixel-wise or instance-wise aggregation from adjacent frames similarly to other video tasks, like video object detection [20, 19, 45]. Although some attempts [5, 46, 52] have been made to tackle VIS in a simple single-stage manner, their performances are still not satisfying.

The key difference between video and image instance segmentation lies in the need of capturing robust and accurate instance association across frames. However, most previous works such as MaskTrack R-CNN [50], and CMaskTrack R-CNN [33] formulate instance association as a multi-label classification problem, focusing only on the intrinsic relationship within instances while ignoring the extrinsic constraint between different ones. Thus different instances with similar distributions may be wrongly associated by using previous tracking embeddings only through such multi-label classification loss constraint.

Alternatively, we propose an *efficient* single-stage fully convolutional network for video instance segmentation task, considering that single-stage instance segmentation is simpler and faster. Based on the recent instance segmentation method CondInst [37], an extra tracking head is added to simultaneously learn instance-wise tracking embeddings for instance association besides original classification head, box head, and mask head by dynamic filter. To improve in-

stance association accuracy between adjacent frames, a spatio-temporal contrastive learning strategy is utilized to exploit relations between different instances. Specifically, for a tracking embedding query, we densely sample hundreds of negative and positive embeddings from reference frames based on the label assignment results, acting as a contrastive manner to jointly pull closer to the same instances and push away from different instances. Different from previous metric learning based instance association methods *i.e., Triplet Loss*, the proposed contrastive strategy enables efficient many-to-many relations learning across frames. We believe this contrast mechanism enhances the instance similarity learning, which provides more substantial supervision than using only the labels. Moreover, this contrastive learning scheme is applied in a bi-directional way to better leverage the temporal information from both forward and backward views. At last, we further propose a temporal consistency scheme for instance encoding, which contributes to both the accuracy and smoothness of the video instance segmentation task.

In summary, our main contributions are:

- We propose a single-stage fully convolutional network for video instance segmentation task with an extra tracking head to simultaneously generate instance-specific tracking embeddings for instance association.
- To achieve accurate and robust instance association, we propose a bi-directional spatio-temporal contrastive learning strategy that aims to obtain representative and discriminative tracking embeddings. In addition, we present a novel temporal consistency scheme for instances encoding to achieve temporally coherent results.
- Comprehensive experiments are conducted on the YouTube-VIS-2019, YouTube-VIS-2021, and OVIS-2021 benchmark. Without bells and whistles, we achieve 36.7% AP and 35.5% AP with ResNet-50 backbone on YouTube-VIS-2019 and YouTube-VIS-2021 datasets, which is the best performance among all listed single-model methods with high efficiency. We also achieve best performance on recent proposed OVIS-2021 dataset. In particular, compared to the first VIS method named MaskTrack R-CNN [50], our proposed method (STC) achieves 36.7% AP on YouTube-VIS-2019, outperforming it by **6.4**% AP with the advantage of being much simpler and faster. Compared with recent method CrossVIS [52], STC outperforms it by 1.9% AP with a slightly faster speed.

2 Related Works

2.1 Instance Segmentation

Instance segmentation aims to represent objects at a pixel level, which is a finer-grained representation compared with object detection. There are mainly two kinds of instance segmentation methods, *i.e.,* two-stage [14, 28, 16], and single-stage [6, 4, 41, 42, 37]. Two-stage methods first detect objects, then crop their region features to further classify each pixel into the foreground or background,

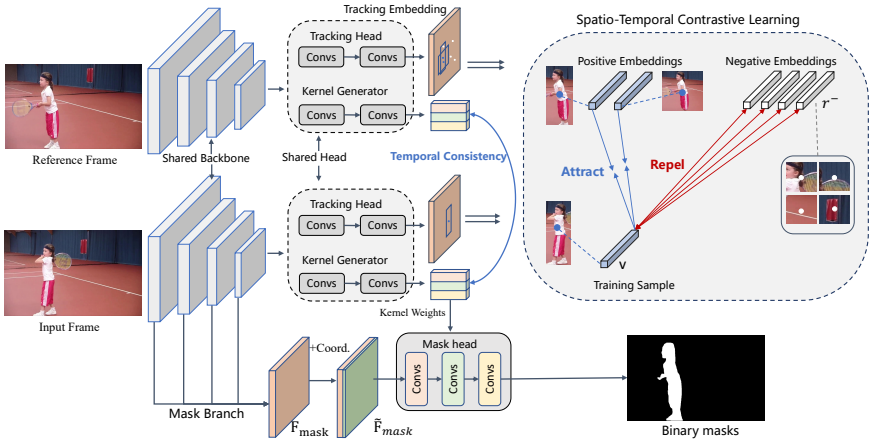

Fig. 2. The overview of our proposed framework. The framework contains the following components: a shared CNN backbone for encoding frames to feature maps, kernel generators with mask heads for instance segmentation, a mask branch to combine multi-scale FPN features, and a shared tracking head with a bi-directional spatio-temporal contrastive learning strategy (the bi-directional learning scheme is omitted here for simplicity) for instance association. A temporal consistency constraint is applied to the kernel weights, as the blue line shows. Best viewed in color.

while the framework of single-stage instance segmentation is much simpler. For example, YOLACT [4] is proposed to generate a set of prototype masks and predict per-instance mask coefficients. The instance masks are then produced by linearly combining the prototypes with the mask coefficients. SOLO [41, 42] reformulates the instance segmentation as two simultaneous category-aware prediction problems, *i.e.*, location prediction, and mask prediction, respectively. Inspired by dynamic filter network [18], CondInst [37] proposes to dynamically predict instance-aware filters for mask generation. SOLOv2 [42] further incorporates dynamic filter scheme to dynamically segments each instance in the image with a novel matrix non-maximum suppression (NMS) technique. Compared to the image instance segmentation, video instance segmentation aims not only to segment object instances in individual frames but also to associate the predicted instances across frames.

2.2 Video Instance Segmentation

Video instance segmentation [50] aims to simultaneously classify, segment, and track instances of the videos. Various complicated pipelines are designed by state-of-the-art methods to solve it. To better introduce the related methods, we separate them into the following groups. (1) The two-stage method MaskTrack R-CNN [50], as the pioneering work for VIS, extends image instance segmentation method Mask R-CNN [14] to video domain by introducing an extra

tracking branch for instance association. Another method in the two-stage group is MaskProp [2], which first uses Hybrid Task Cascade (HTC) [7] to generate the predicted masks and propagates them temporally to the other frames in a video. Recently, CompFeat [12] proposed a feature aggregation approach based on MaskTrack R-CNN, which refines features by aggregating multiple adjacent frames features. (2) Relying on 3D convolutions, STEm-Seg [1] models a video clip as a single 3D spatial-temporal volume and separates object instances by clustering. (3) Based on feature-aggregation, STMask [23] proposes a simple spatial feature calibration to detect and segment object masks frame-by-frame, and further introduces a temporal fusion module to track instances across frames. (4) More recently, a transformer-based method VisTR [44] is proposed to reformulate VIS as a parallel sequence decoding problem. (5) There also exist some single-stage VIS methods, *e.g.,* SipMask [5], and TraDeS [46]. SipMask [5] proposes a spatial preservation module to generate spatial coefficients for the mask predictions while recently proposed TraDeS [46] presents a joint detection and tracking model by propagating the previous instance features with the predicted tracking offset. CrossVIS [52] proposes cross-frame instance-wise consistency loss for video instance segmentation. Although current methods have made good progress, their complicated pipelines or unsatisfying performance prohibit practical application. In contrast, the proposed framework acts in a fully convolutional manner with decent performance and efficiency.

2.3 Contrastive Learning

Contrastive learning has lead to considerable progress in many real-world applications [13, 8, 36, 48, 31, 21, 9]. For example, MOCO [13] builds image-level large dictionaries for unsupervised representation learning using contrastive loss. SimCLR [8] utilizes the elaborate data augmentation strategies and a large batch, which outperforms MOCO by a large margin on self-supervised learning ImageNet [34] classification task. Different from the above methods, which focus on image-level contrastive learning for unsupervised representation learning, we use modified multiple-positives contrastive learning to learn instance-level tracking embeddings accurately for video instance segmentation tasks.

3 Method

In this section, we first briefly review the instance segmentation method CondInst [37] for mask generation of still-image. Then, we introduce the proposed whole framework for the video instance segmentation task. Next, we present a novel spatio-temporal contrastive learning strategy for tracking embeddings to achieve accurate and robust instance association. In addition, we further propose a bi-directional spatio-temporal contrastive learning strategy. At last, the temporal consistency scheme aiming to achieve temporally coherent results is introduced in detail.

3.1 Mask Generation for Still-image

For still-image instance segmentation, we use the dynamic conditional convolutions method CondInst [37, 18]. Specifically, instance mask at location (x, y) can be generated by convolving an instance-agnostic feature map $\tilde{\mathbf{F}}^{x,y}_{mask}$ from mask branch and instance-specific dynamic filter $\boldsymbol{\theta}_{x,y}$, which is calculated as follows:

$$\mathbf{m}_{x,y} = \mathbf{MaskHead}(\tilde{\mathbf{F}}^{x,y}_{mask}; \boldsymbol{\theta}_{x,y}), \tag{1}$$

where $\tilde{\mathbf{F}}^{x,y}_{mask}$ is the combination of multi-scale fused feature map $\mathbf{F}_{mask}$ from FPN features $\{P_3, P_4, P_5\}$ and relative coordinates $\mathbf{O}_{x,y}$. The **MaskHead** consists of three 1×1 conv-layers with dynamic filter $\boldsymbol{\theta}_{x,y}$ at location (x, y) as convolution kernels. $\mathbf{m}_{x,y} \in \mathbb{R}^{H \times W}$ is the predicted binary mask at location (x, y) as shown in Figure 2.

3.2 Proposed Framework for VIS

The overall framework of the proposed method is illustrated in Figure 2. Based on the instance segmentation method CondInst [37], we add a tracking head for instances association. The whole architecture mainly contains following components: (1) A shared CNN backbone (*e.g.* ResNet-50 [15]) is utilized to extract compact visual feature representations with FPN [25]. (2) Multiple heads including a classification head, a box regression head, a centerness head, a kernel generator head, and a mask head as same as CondInst [37]. Since the architectures of the above classification, box regression, and centerness heads are not our main concerns, we omit them here (please refer to [38] for the details). (3) A tracking head where spatio-temporal contrastive learning strategy is proposed to associate instances across frames with comprehensive relational cues in the tracking embeddings. (4) Temporal consistency scheme on instance-wise kernel weights across frames aims to generate temporally coherent results.

3.3 Spatio-Temporal Contrastive Learning

To associate instances from different frames, an extra lightweight tracking head is added to obtain the tracking embeddings [50, 5, 12] in parallel with the original kernel generator head. The tracking head consists of several convolutional layers which take multi-scale FPN features $\{P_3, P_4, P_5\}$ as input. And the outputs are fused to obtain the feature map of tracking embedding. As shown in Figure 2, given an input frame I for training, we randomly select a reference frame I_{ref} from its temporal neighborhood. A location is defined as a positive sample if it falls into any ground-truth box and the class label c of the location is the class label of the ground-truth box. If a location falls into multiple bounding boxes, it is considered as the positive sample of the bounding box with minimal area [38]. Thus, two locations formulate a positive pair if they are associated with the same instance across two frames and a negative pair otherwise.

During training, for a given frame, the model first predicts the object detection results. Then, the tracking embedding of each instance can be extracted

from the tracking feature map by the center of the predicted bounding box. For a training sample with extracted tracking embedding q, we can obtain positive embeddings $\mathbf{k}^+$ and negative embeddings $\mathbf{k}^-$ according to label assignment results at reference frame. Note that traditional unsupervised representation learning [13, 8] with contrastive learning only uses one positive sample and multiple negative samples as follows:

$$\mathcal{L}_q = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k}^+)}{\exp(\mathbf{q} \cdot \mathbf{k}^+) + \sum_{\mathbf{k}^-} \exp(\mathbf{q} \cdot \mathbf{k}^-)}. \tag{2}$$

Since there are many positive embeddings at reference frame for each training sample, instead of randomly selecting one positive embedding at reference frames, we optimize the objective loss with multiple positive embeddings and multiple negative embeddings as:

$$\begin{aligned} \mathcal{L}_{contra} &= -\sum_{\mathbf{k}^+} \log \frac{\exp(\mathbf{q} \cdot \mathbf{k}^+)}{\exp(\mathbf{q} \cdot \mathbf{k}^+) + \sum_{\mathbf{k}^-} \exp(\mathbf{q} \cdot \mathbf{k}^-)} \\ &= \sum_{\mathbf{k}^+} \log[1 + \sum_{\mathbf{k}^-} \exp(\mathbf{q} \cdot \mathbf{k}^- - \mathbf{q} \cdot \mathbf{k}^+)]. \end{aligned} \tag{3}$$

Suppose there are N_{pos} training samples at input frame, the objective track loss with multiple samples is:

$$\mathcal{L}_{track} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \mathcal{L}^i_{contra}. \tag{4}$$

Bi-directional Spatio-Temporal Learning. Many video-related tasks have shown the effectiveness of bi-directional modeling [54, 35]. To fully exploit such temporal context information, we further propose a bi-directional spatio-temporal learning scheme to learn instance-wise tracking embeddings better. Note that we only utilize this scheme in the training stage, and thus it does not affect the inference speed. Similar to Equation 4, the objective function of bi-directional spatio-temporal contrastive learning can be denoted as $\hat{\mathcal{L}}_{track}$ by reversing input frame and reference frame. Thus, the final bi-directional spatio-temporal contrastive loss is:

$$\mathcal{L}_{bi-track} = \frac{1}{2}(\mathcal{L}_{track} + \hat{\mathcal{L}}_{track}). \tag{5}$$

3.4 Temporal Consistency

Compared with image data, the coherent property between frames is also crucial to video-related researches. Thus, we add a temporal consistency constraint on the kernel weights, marked as the blue line in Figure 2, to capture such prior during training so that the predicted masks will be more accurate and robust across frames. Given an instance at location (x, y) appearing at both

input and reference frames, we use (x, y) and $(\hat{x}, \hat{y})$ to denote its positive candidate positions from two frames, respectively. Formally, the temporal consistency constraint during training can be formulated as an L2-loss function:

$$\mathcal{L}_{consistency} = ||\boldsymbol{\theta}_{x,y} - \boldsymbol{\theta}^{ref}_{\hat{x},\hat{y}}||^2 + ||\boldsymbol{m}_{x,y} - \boldsymbol{m}^{ref}_{\hat{x},\hat{y}}||^2, \tag{6}$$

where $\boldsymbol{\theta}^{ref}_{\hat{x},\hat{y}}$ is the dynamic filter at reference frame, $\boldsymbol{m}^{ref}_{\hat{x},\hat{y}}$ is the predicted instance mask by reference dynamic filter. With such a simple constraint, our kernel generator can obtain accurate, robust and coherent mask predictions across frames.

3.5 Training and Inference

Training Scheme. Formally, the overall loss function of our model can be formulated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{condinst} + \lambda_b \mathcal{L}_{bi-track} + \lambda_c \mathcal{L}_{consistency}, \tag{7}$$

where $\mathcal{L}_{condinst}$ denotes the original loss of CondInst [37] for instance segmentation. We refer readers to [37] for the details of $\mathcal{L}_{condinst}$. λ_b and λ_c are the hyper-parameters.

Inference on Frame. For each frame, we forward it through the model to get the outputs, including classification confidence, centerness scores, box predictions, kernel weights, and tracking embeddings. Then we obtain the box detections by selecting the positive positions whose classification confidence is larger than a threshold (set as 0.03), similar to FCOS [38]. After that, following previous work MaskTrack R-CNN [50], the NMS [4] with the threshold being 0.5 is used to remove duplicated detections. In this step, these boxes are also associated with the kernel weights and tracking embeddings. Supposing that there remain T boxes after the NMS, thus we have T groups of the generated kernel weights. Then T groups of kernel weights are used to produce T mask heads. These instance-specific mask heads are applied to the positions encoded mask feature to predict the instance masks following [37]. T is 10 in default following previous work MaskTrack R-CNN.

Inference on Video. Given a testing video, we first construct an empty memory bank for the predicted instance embeddings. Then our model processes each frame sequentially in an online scheme. Our network generates a set of predicted instance embeddings at each frame. The association with identified instances from previous frames relies on the cues of embedding similarity, box overlap, and category label similar to the MaskTrack R-CNN [50]. All predicted instance embeddings of the first frame are directly regarded as identified instances and saved into the memory bank. After processing all frames, our method produces a set of instances sequence. The majority votes are utilized to decide the unique category label of each instance sequence.

Table 1. Comparisons with some state-of-the-art approaches on **YouTube-VIS-2019** `val` set. ✓indicates using extra data augmentation (*e.g.*, random crop, higher resolution input, multi-scale training) [2] or additional data [1, 2, 12, 44]. † indicates the method that reaches higher performance by stacking multiple networks, and we regard it an unfair competitor in general setting. Note that STMask [23] uses deformable convolution network (DCN) [10] as the backbone, which is still inferior to our method at both accuracy and speed, demonstrating the superiority of our proposed framework. †† means transformer on top of ResNet-50 or ResNet-101.

Method	Publication	Augmentations	Backbone	FPS	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN [50]	ICCV'19	✗	ResNet-50	33	30.3	51.1	32.6	31.0	35.5
SipMask [5]	ECCV'20	✗	ResNet-50	34	32.5	53.0	33.3	33.5	38.9
STEm-Seg [1]	ECCV'20	✗	ResNet-50	4.4	30.6	50.7	33.5	31.6	37.1
CompFeat [12]	AAAI'21	✗	ResNet-50	< 33	35.3	56.0	38.6	33.1	40.3
TraDeS [46]	CVPR'21	✗	ResNet-50	26	32.6	52.6	32.8	29.1	36.6
QueryInst [11]	ICCV'21	✗	ResNet-50	32	34.6	55.8	36.5	35.4	42.4
CrossVIS [52]	ICCV'21	✗	ResNet-50	39.8	34.8	54.6	37.9	34.0	39.0
VisSTG [40]	ICCV'21	✗	ResNet-50	22	35.2	55.7	38.0	33.6	38.5
PCAN [22]	NeurIPS'21	✗	ResNet-50	-	36.1	54.9	**39.4**	36.3	41.6
Ours (STC)	-	✗	ResNet-50	**40.3**	**36.7**	**57.2**	38.6	**36.9**	**44.5**
STMask [23]	CVPR'21	DCN backbone [10]	ResNet-50	29	33.5	52.1	36.9	31.1	39.2
SG-Net [27]	CVPR'21	multi-scale training	ResNet-50	23	34.8	56.1	36.8	35.8	40.8
VisTR [44]	CVPR'21	random-crop training	ResNet-50	30	35.6	56.8	37.0	35.2	40.2
QueryInst [11]	ICCV'21	multi-scale training	ResNet-50	32	36.2	56.7	39.7	36.1	42.9
CrossVIS [52]	ICCV'21	multi-scale training	ResNet-50	39.8	36.3	56.8	38.9	35.6	40.7
VisSTG [40]	ICCV'21	multi-scale training	ResNet-50	22	36.5	58.6	39.0	35.5	40.8
Ours (STC)	-	multi-scale training	ResNet-50	**40.3**	**37.6**	**58.9**	**39.7**	**38.2**	**46.2**
MaskTrack R-CNN [50]	ICCV'19	✗	ResNet-101	33	30.3	51.1	32.6	31.0	35.5
SRNet [53]	ACMMM'21	✗	ResNet-101	35	32.3	50.2	34.8	32.3	40.1
STEm-Seg [1]	ECCV'20	✗	ResNet-101	2.1	34.6	55.8	37.9	34.4	41.6
PCAN [22]	NeurIPS'21	✗	ResNet-101	-	37.6	57.2	**41.3**	37.2	43.9
Ours (STC)	-	✗	ResNet-101	**36.6**	**37.8**	**58.5**	40.6	**38.5**	**46.3**
SipMask [5]	ECCV'20	multi-scale training	ResNet-101	24	35.8	56.0	39.0	35.4	42.4
STMask [23]	CVPR'21	DCN backbone [10]	ResNet-101	23	36.8	56.8	38.0	34.8	41.8
SG-Net [27]	CVPR'21	multi-scale training	ResNet-101	20	36.3	57.1	39.6	35.9	43.0
VisTR [44]	CVPR'21	random-crop training	ResNet-101	28	38.6	61.3	42.3	37.6	44.2
Ours (STC)	-	multi-scale training	ResNet-101	**36.6**	**39.2**	**61.5**	**42.4**	**39.7**	**47.3**

4 Experiments

4.1 Dataset

To verify the effectiveness of our approach, we evaluate it on recent three video instance segmentation benchmarks, YouTube-VIS-2019 [50], YouTube-VIS-2021 [49] and OVIS-2021 [33] datasets. Following previous works [50, 2, 52], we evaluate our method on the validation sets of YouTube-VIS-2019, YouTube-VIS-2021 and OVIS-2021.

YouTube-VIS-2019 dataset contains 40 class annotations, including many common objects. The official dataset consists of three subsets: 2238 training videos, 302 validation videos, and 343 test videos.

YouTube-VIS-2021 dataset is an improved version of YouTube-VIS-2019 containing 40 class annotations. It collects more videos and high-quality annota-

tions. This dataset also consists of three subsets: 2985 training videos, 421 validation videos, and 453 test videos.
OVIS-2021 is a new large scale benchmark dataset for video instance segmentation task with 25 common semantic categories. It is designed with object occlusions in videos, which could reveal the complexity of real-world scenes. It consists of 607 training videos, 140 validation videos, and 154 testing videos as the official split.

4.2 Metrics

The evaluation metrics are average precision (AP) and average recall (AR), with the video Intersection over Union (IoU) of the mask sequences as the threshold [50]. Specifically, for a predicted mask $\hat{m}^i$ and a ground-truth mask m^j, we first extend them to the whole video with length T by padding empty mask. Then,

$$\text{IoU}(i,j) = \frac{\sum_{t=1}^{T} \hat{m}_t^i \cap m_t^j}{\sum_{t=1}^{T} \hat{m}_t^i \cup m_t^j}. \tag{8}$$

According to the definition, if the model detects object masks successfully but fails to associate the objects across frames, it still gets a low IoU. Thus, accurate and robust instance association across frames is very crucial for achieving high performance.

4.3 Implementation Details

Model Settings. In our experiments, we choose the ResNet-50 [15] and ResNet-101 with FPN [25] as the backbone in the proposed method. Our model is pre-trained on COCO `train2017` [26] with $1\times$ schedule following previous works [5, 52, 50]. We implement the proposed method with PyTorch [32] and the FPS is measured on an RTX 2080 Ti GPU including the pre- and post-processing steps for fair comparison following previous work [52]. The optimizer of the proposed method is SGD, with a learning rate 5e-3 and a weight decay 1e-4. The models are trained with $1\times$ schedule for 12 epoch, and we decay the lr with the ratio 0.1, in the 8-th and 11-th epoch. The input frames are resized to 640×360 following previous works [50, 52, 12].
Hyper-parameters. There exists some hyper-parameters in our proposed framework, *i.e.*, bi-directional contrastive learning loss λ_b, and temporal consistency loss λ_c. In this paper, we set $\lambda_b = 0.2$ and $\lambda_c = 10$ in default.

4.4 Main Results

Here we compare our method with two-stage [51, 39, 50, 2, 12], single-stage [5, 46, 27], 3D convolution-based [1], feature aggregation-based [23], and transformer-based [44] methods. For some differences in the training settings (*e.g.*, resolution, training epochs) vary from different methods, we strictly follow MaskTrack

Table 2. Comparisons with some recent VIS methods on the **YouTube-VIS-2021** val set. We use ResNet-50 backbone and 1× schedule for all experiments for fair comparison.

Methods	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
SipMask [5]	28.6	48.9	29.6	26.5	33.8
MaskTrack R-CNN [50]	31.7	52.5	34.0	30.8	37.8
STEm-Seg [1]	33.3	53.8	37.0	30.1	37.6
CrossVIS [52]	34.2	54.4	37.9	30.4	38.2
Ours (STC)	**35.5**	**57.4**	**38.0**	**32.8**	**42.2**

Table 3. Comparisons with some recent VIS methods on very challenging **OVIS-2021** val set. We use ResNet-50 backbone and 1× schedule for all experiments for fair comparison.

Methods	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
SipMask [5]	10.3	25.4	7.8	7.9	15.8
MaskTrack R-CNN [50]	10.9	26.0	8.1	8.3	15.2
STEm-Seg [1]	13.8	32.1	11.9	9.1	20.0
CrossVIS [52]	14.9	32.7	12.1	10.3	19.8
Ours (STC)	**15.5**	**33.5**	**13.4**	**11.0**	**20.8**

R-CNN [50], SipMask [5] and CrossVIS [52] with 1× schedule and 640×360 resolution for fair comparison.

YouTube-VIS-2019. Without any bells and whistles, our proposed method achieves the best performance 36.7% AP among the listed single-model methods. More specifically, among the two-stage methods, our model outperforms the original MaskTrack R-CNN [50] by **6.4** % in AP (36.7% vs. 30.3%). As discussed in VisTR [44], we also argue that the performance of MaskProp [2] relies heavily on stacking multiple networks, *e.g.*, Spatio-temporal Sampling Network [3] and Hybrid Task Cascade Network [7], not to mention the larger resolution and more training epochs. Our model also beats the recently proposed CompFeat [12] by 1.4 % in AP with a significant improvement on the performance of speed. Meanwhile, it outperforms STEm-Seg [1] and VisTR [44] with the same backbone on the accuracy, which indicates the superiority of our method. Note that VisTR utilizes multi-scale training and takes a week on 8 NVIDIA Tesla V100 for training. Furthermore, compared with the single-stage methods SipMask [5] and TraDeS [46], our method obtains about 4.2 % and 4.1 % improvement in AP, respectively. Compared with the feature aggregation-based method STMask [23] which uses multi-frames to obtain more robust features, our method surpasses it by 3.2 % in AP for ResNet-50 backbone, and even it uses a stronger ResNet-50-DCN backbone. When compared with recent work CrossVIS [52], our method still shows the superiority of the performance on both performance and speed.

As shown in Table 1, we also compare the FPS (frames per second) with other state-of-the-art methods. Our method achieves 36.7% AP at a 40.3 FPS, which is the best tradeoff for the single model. In addition, our method can run an online mode which is crucial for practical usages.

YouTube-VIS-2021. We evaluate the recently proposed MaskTrack R-CNN [50], SipMask [5] and CrossVIS [52] on YouTube-VIS-2021 using the official implementation for comparison. As shown in Table 2, our method surpasses MaskTrack R-CNN [50] and CrossVIS [52] by 3.8 % and 1.3 % in AP , which verifies the effectiveness of our method.

OVIS-2021. From Table 3 we can observe that all methods meet a large performance degradation due to the complexity and occlusions in OVIS-2021 dataset. Our method achieves the best 15.5% AP, surpassing all methods under the same

Table 4. Ablation studies for each component of the proposed framework on YouTube-VIS-2019 validation set.

Baseline	Consistency	Contrastive	Bi-direction	AP
✓				33.7
✓	✓			34.4
✓	✓	✓		36.3
✓	✓	✓	✓	**36.7**

Table 5. Comparisons among different settings of the track embedding on the YouTube-VIS-2019 validation set.

Contrastive	Bi-direction	Embedding dim	AP
✗	✗	256	34.5
✓	✗	256	36.2
✗	✓	256	35.4
✓	✓	256	**36.7**

Table 6. Comparisons among different settings of the kernel generator head on the YouTube-VIS-2019 validation set.

Consistency	# Conv	AP
✗	1	31.5
✗	2	33.7
✗	3	36.2
✗	4	36.0
✓	3	**36.7**

Table 7. Comparisons among different settings of the mask branch on the YouTube-VIS-2019 validation set.

Coord.	# Channel	AP
✗	1	28.7
✗	4	33.6
✗	8	36.2
✗	16	36.1
✓	8	**36.7**

experimental conditions. We hope that our proposed method can serve as a strong baseline for this challenging benchmark.

4.5 Ablation Studies

We conduct experiments on the YouTube-VIS-2019 validation set with the ResNet-50 backbone and 1× schedule for the ablation studies.

Analysis for Each Component. As shown in Table 4, we first use CondInst [37] to obtain the instance masks instead of utilizing RoIAlign and mask head in MaskTrack R-CNN [50], which achieves 3.4 % in AP improvements (33.7% vs. 30.3%). Besides the performance improvement, this component also changes the two-stage model to a simple single-stage and fully convolutional one with faster speed. Note that our temporal consistency constraint for the kernel generator successfully gains 0.7 % in AP by digging deeper into the temporal information in the video sequence. For the instances association across frames, we conduct experiments to verify the effectiveness of two components ("Contrastive" and "Bi-direction"). Specifically, when only using spatio-temporal contrastive learning module, we could achieve 1.9% in AP improvement. When using the bi-directional contrastive learning strategy, we finally obtain 36.7% in AP, surpassing "Contrastive" baseline by 0.4% in AP, demonstrating the effectiveness of the bi-directional learning strategy.

Kernel Generator. Kernel generator from CondInst [37] plays a critical role in our method. Thus, we conduct ablation studies to show the impact of parameters in kernel generator head. As presented in Table 6, with the number of convolutions in kernel generator head increasing, the performance improves

Table 8. Comparisons among different settings of the negative sampling methods of contrastive learning on the YouTube-VIS-2019 validation set.

Inbox	# Negative	AP
✗	0	34.9
✗	64	36.5
✗	128	**36.7**
✗	256	36.4
✓	128	35.2

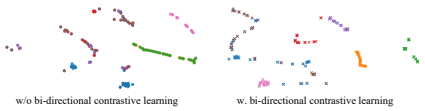

Fig. 3. Visualizations of instance embeddings without or with bi-directional contrastive learning module using t-SNE.

steadily and achieves the peak 36.7% AP with three stacked convolutions. Temporal consistency obtains 0.5% in AP, which demonstrates the effectiveness.

Mask Branch. To enhance the expressiveness of the mask feature, we further explore the channel number and relative coordinate map ("Coord.") used in the mask branch. As illustrated in Table 7, the 8-channel mask feature achieve 36.2% AP without the coordinate map, and extra channels cannot improve the performance. We set the channel number of the mask feature to 8 by default as a result. Relative coordinates are attached to the mask feature for better performance (about 0.5% in AP improvement).

Tracking Embedding. The tracking embedding is crucial for VIS since AP relies heavily on the accuracy of instance association. We compare with different tracking embedding dimensions. As shown in Table 5, AP improves as the embedding dimension increases. However, we can not afford the complexity cost when the embedding dimension is larger than 256 considering the speed. Thus, we set the embedding dimension as 256 by default.

Negative Sampling. The designed contrastive learning strategy aims to obtain representative and discriminative tracking embeddings. Thus, we further explore different numbers of negative embeddings and how they are selected in Table 8. "Inbox" means we randomly select the negative embeddings within boxes from negative locations according to label assignment results. We find that choosing 128 negative embeddings is a good balance of total training time and accuracy. Moreover, randomly selecting negative embeddings from the whole feature map of the reference frame is much better than "Inbox". This observation verifies that the model can learn more discriminative representations from background stuff or objects.

Fig. 4. Visualization of our proposed method and MaskTrack R-CNN on the YouTube-VIS-2019 val set.

4.6 Visualizations

Instance Embedding. To verify the effectiveness of the proposed method qualitatively, we visualize the instance embeddings of the same video sequence using t-SNE [29], which is shown in Figure 3. Comparing with Figure 3(a), the instance embeddings of Figure 3(b) is more separable, which indicates that our proposed STC module helps to distinguish different instances in the embedding space. Thus, compared with the original multi-class classification loss [50], we could obtain more accurate instance association accuracy for video instance segmentation task.

Video Visualization. The visualization of the proposed method on the YouTube-VIS-2019 validation dataset is shown in Figure 4. Compared with baseline method MaskTrack R-CNN [50], as shown in the first row and the second row, STC achieve more accurate segmentation results. From the last two rows, STC could achieve more coherent tracking results compared with MaskTrack R-CNN baseline, which demonstrates the effectiveness of the proposed spatio-temporal contrastive learning strategy. In conclusion, our method can segment and associate instances better with more accurate boundary results in challenging situations while MaskTrack R-CNN suffers from the missing instances or identity mistakes.

5 Conclusion

In this work, we introduced a effective architecture for video instance segmentation. Our model is conceptually simple without requiring RoIAlign operation or 3D convolutions. Moreover, it achieves state-of-the-art single-model results (*i.e.*, ResNet-50 backbone) on the YouTube-VIS-2019, YouTube-VIS-2021, and OVIS-2021 datasets in a fully convolutional fashion. We hope our work could serve an strong baseline, which could inspire designing more efficient framework and rethinking the embeddings loss for challenging video instance segmentation task.

References

1. Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., Leibe, B.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: European Conference on Computer Vision. pp. 158–177 (2020)
2. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9739–9748 (2020)
3. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: European Conference on Computer Vision. pp. 331–346 (2018)
4. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9157–9166 (2019)
5. Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: European Conference on Computer Vision. pp. 1–18 (2020)
6. Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y.: Blendmask: Top-down meets bottom-up for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8573–8581 (2020)
7. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4974–4983 (2019)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607 (2020)
9. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
10. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773 (2017)
11. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6910–6919 (2021)
12. Fu, Y., Yang, L., Liu, D., Huang, T.S., Shi, H.: Compfeat: Comprehensive feature aggregation for video instance segmentation. arXiv preprint arXiv:2012.03400 (2020)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
16. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6409–6418 (2019)

17. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers. Advances in Neural Information Processing Systems **34** (2021)
18. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Advances in Neural Information Processing Systems pp. 667–675 (2016)
19. Jiang, Z., Gao, P., Guo, C., Zhang, Q., Xiang, S., Pan, C.: Video object detection with locally-weighted deformable neighbors. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
20. Jiang, Z., Liu, Y., Yang, C., Liu, J., Gao, P., Zhang, Q., Xiang, S., Pan, C.: Learning where to focus for efficient video object detection. In: European Conference on Computer Vision (2020)
21. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. arXiv preprint arXiv:2010.01028 (2020)
22. Ke, L., Li, X., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F.: Prototypical cross-attention networks for multiple object tracking and segmentation. Advances in Neural Information Processing Systems **34** (2021)
23. Li, M., Li, S., Li, L., Zhang, L.: Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11215–11224 (2021)
24. Lin, H., Wu, R., Liu, S., Lu, J., Jia, J.: Video instance segmentation with a propose-reduce paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1739–1748 (2021)
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
27. Liu, D., Cui, Y., Tan, W., Chen, Y.: Sg-net: Spatial granularity network for one-stage video instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9816–9825 (2021)
28. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8759–8768 (2018)
29. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)
30. Oksuz, K., Cam, B.C., Akbas, E., Kalkan, S.: Rank & sort loss for object detection and instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3009–3018 (2021)
31. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 164–173 (2021)
32. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32**, 8026–8037 (2019)
33. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation. arXiv preprint arXiv:2102.01558 (2021)

34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
35. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)
36. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243 (2020)
37. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: European Conference on Computer Vision. pp. 282–298 (2020)
38. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9627–9636 (2019)
39. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9481–9490 (2019)
40. Wang, T., Xu, N., Chen, K., Lin, W.: End-to-end video instance segmentation via spatial-temporal graph neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10797–10806 (2021)
41. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: Solo: Segmenting objects by locations. In: European Conference on Computer Vision. pp. 649–665 (2020)
42. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. arXiv preprint arXiv:2003.10152 (2020)
43. Wang, Y., Xu, Z., Shen, H., Cheng, B., Yang, L.: Centermask: single shot instance segmentation with point representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9313–9321 (2020)
44. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8741–8750 (2021)
45. Wu, H., Chen, Y., Wang, N., Zhang, Z.: Sequence level semantics aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9217–9225 (2019)
46. Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12352–12361 (2021)
47. Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polarmask: Single shot instance segmentation with polar representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12193–12202 (2020)
48. Xiong, Y., Ren, M., Urtasun, R.: Loco: Local contrastive representation learning. arXiv preprint arXiv:2008.01342 (2020)
49. Xu, N., Yang, L., Yang, J., Yue, D., Fan, Y., Liang, Y., Huang, T.S.: Youtube-vis dataset 2021 version. https://youtube-vos.org/dataset/vis (2021)
50. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5188–5197 (2019)
51. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6499–6507 (2018)
52. Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Crossover learning for fast online video instance segmentation. arXiv preprint arXiv:2104.05970 (2021)

53. Ying, X., Li, X., Chuah, M.C.: Srnet: Spatial relation network for efficient single-stage instance segmentation in videos. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 347–356 (2021)
54. Zhu, L., Xu, Z., Yang, Y.: Bidirectional multirate reconstruction for temporal modeling in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2653–2662 (2017)
55. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision (2017)