

Temporally Efficient Vision Transformer for Video Instance Segmentation

Shusheng Yang^{1,3*}; Xinggang Wang^{1†}; Yu Li^{4*}, Yuxin Fang¹,
Jiemin Fang^{2,1}, Wenyu Liu¹, Xun Zhao³, Ying Shan³

¹School of EIC, Huazhong University of Science & Technology

²Institute of Artificial Intelligence, Huazhong University of Science & Technology

³Applied Research Center (ARC), Tencent PCG ⁴International Digital Economy Academy (IDEA)

Abstract

Recently vision transformer has achieved tremendous success on image-level visual recognition tasks. To effectively and efficiently model the crucial temporal information within a video clip, we propose a **Temporally Efficient Vision Transformer (TeViT)** for video instance segmentation (VIS). Different from previous transformer-based VIS methods, TeViT is nearly convolution-free, which contains a transformer backbone and a query-based video instance segmentation head. In the backbone stage, we propose a nearly parameter-free messenger shift mechanism for early temporal context fusion. In the head stages, we propose a parameter-shared spatiotemporal query interaction mechanism to build the one-to-one correspondence between video instances and queries. Thus, TeViT fully utilizes both frame-level and instance-level temporal context information and obtains strong temporal modeling capacity with negligible extra computational cost. On three widely adopted VIS benchmarks, i.e., YouTube-VIS-2019, YouTube-VIS-2021, and OVIS, TeViT obtains state-of-the-art results and maintains high inference speed, e.g., 46.6 AP with 68.9 FPS on YouTube-VIS-2019. Code is available at <https://github.com/hustvl/TeViT>.

1. Introduction

Video Instance Segmentation (VIS) [62] is a representative and challenging video understanding task that requires detecting, segmenting and tracking video instances across frames simultaneously. Similar to other instance-level video recognition tasks, making full use of temporal context information is critical for building high-performance VIS systems. Vision transformer (ViT) [13], which is based on self-attention [51], has shown strong long-range context modeling ability and obtained great successes on im-

age classification [13, 15, 21, 31, 47, 50, 55], object detection [7, 33, 44, 68], semantic segmentation [11, 43, 59], instance segmentation [12, 17, 66], and video recognition [1, 4, 14, 32, 36, 65, 67].

Recently, how to design ViTs for instance-level video understanding, especially VIS, becomes an emerging problem. Different from the detection transformers [7, 16, 33, 44, 68], semantic segmentation transformers [11, 43, 59], and instance segmentation transformers [12, 17, 66], which focus on 2D contextual information modeling, VIS transformers additionally require to perform temporal context modeling. To this end, VisTR [57] firstly proposes a transformer encoder to fuse patch features from a sequence of frames using a CNN backbone and leverages a query-based decoder to predict video instances, IFC [22] introduces memory tokens to store frame-level features and performs cross-frame feature interaction by computing self-attention among memory tokens, and then decodes instance-level results using a conditional mask head.

In this paper, we focus on the efficiency of modeling temporal information for ViT-based VIS. This is a very important problem since (1) computing self-attention among all video patches has extremely high time and space complexity [57], (2) additional multi-head self attention (MHSA) layers for temporal modeling have extra parameters and are sensitive to pre-training [22], (3) the CNN or transformer backbones in these methods [22, 37, 57, 64] only support single frame feature extraction and fail to capture temporal information in the backbone stage. To remedy the above issues, we present Temporally Efficient ViT (TeViT) to fully utilize temporal contextual information for efficient and effective video instance segmentation.

TeViT contains a transformer backbone and a series of query-based VIS heads. In the backbone stage, we use messenger tokens [15] to extract intra-frame information via self-attention and propose a messenger shift mechanism for frame-level context modeling, in which messenger tokens are divided into several groups to perform temporal shift with various of time steps. Different from previous

*This work was done while Shusheng Yang and Yu Li were at Applied Research Center (ARC), Tencent PCG.

†Corresponding author, E-mail: xgwang@hust.edu.cn.

VIS methods, the messenger shift transformer enables early temporal feature fusion. In the head stages, we convert the QueryInst [17] instance segmentation head into our VIS head by reusing the multi-head self-attention (MHSA) [51] parameters for instance-level temporal information interaction. The instance-level MHSA fuses the features for a single video instance among input frames, thus it realizes the concept of a video instance as a query.

Experiments are conducted on three large-scale VIS datasets, *i.e.*, YouTube-VIS-2019 [62], YouTube-VIS-2021 [61], and OVIS [39]. New state-of-the-art (SoTA) performance has been obtained, *e.g.*, TeViT obtains 46.6 AP with 68.9 FPS on YouTube-VIS-2019. Our main contributions are summarized as follows.

- TeViT is the first video instance segmentation transformer that can efficiently capture temporal contextual information at both frame level and instance level.
- Benefiting from the flexibility of self-attention, the proposed temporal modeling modules, *i.e.*, messenger token shift and spatiotemporal query interaction, both are friendly to the image-level pre-trained models, cost marginal extra computation overhead and parameters.
- TeViT is a nearly convolution-free framework and obtains SoTA VIS results. In TeViT, the concepts of “early temporal feature fusion” and “a video instance as a query” shield lights on how to build effective video transformers for instance-level recognition tasks.

2. Related Work

Video instance segmentation. How to achieve efficiently temporal modeling is always the focus of video tasks, such as video object segmentation (VOS) [5, 5, 38], multi-object tracking and segmentation (MOTS) [52] and VIS [62]. Though VOS and MOTs are very related with VIS, MOTs mainly focuses on the urban scene understanding and VOS aims at tracking specific object by a given mask. Representative VIS works are reviewed as follows. MaskTrack R-CNN [62] extends Faster R-CNN [41] and Mask R-CNN [19] to VIS with a tracking branch and external memory that saves instance features across multiple frames. MaskProp [3] builds on the Hybrid Task Cascade Network [9] and propagates instance region features to adjacent frames to perform temporal modeling. STEM-Seg [2] treats video clip as 3D spatiotemporal volume and captures temporal information by 3D convolutional backbone network. CompFeat [18] refines temporal features at both frame-level and instance-level. CrossVIS [63] introduces a crossover learning scheme upon [45, 46] to make use of contextual information across video frames. SeqMask R-CNN [26] establishes temporal relation across frames by adding an extra

sequence propagation head upon Mask R-CNN. Both VisRGNN [23] and VisSTG [53] model temporal information in VIS by a graph neural network. VisTR [57] proposes the first fully end-to-end VIS method upon DETR [7], temporal contexts are fused by the multi-head attention mechanism in transformer encoder layers. IFC [22] presents inter-frame communication to exchange frame-level information. In this paper, we present a temporally efficient framework to model temporal contexts at both frame-level and instance-level.

Vision Transformer. Transformer [51] is firstly proposed to model long-range sequence data in natural language process (NLP). ViT [13] firstly adopts transformer to image domain. After that various high-performance vision transformers [15, 21, 31, 47, 50, 54, 55] have been proposed as backbones for image understanding. Beyond serving as backbone networks, transformer has motivated lots of novel object detection [7, 16, 33, 44, 68], instance segmentation [12, 17, 66], and semantic segmentation [11, 43, 59] frameworks. Recently, VisTR [57], IFC [22], QueryTrack [64], and TCIS [37] bring transformer to video instance segmentation and achieve excellent performance. In this paper, we investigate how to efficiently model temporal context across video frames and propose TeViT. TeViT is a nearly convolution-free transformer while VisTR and IFC both use ResNet [20] backbone.

Temporal context modeling. Temporal context modeling is the key issue in video understanding. A lot of works [8, 40, 48, 49, 60] model temporal context by 3D convolutional block. TSM [27] proposes an efficient temporal shift module by moving the convolutional feature map along the temporal dimension. Non-local network [56] applies self-attention to capture long-range spatiotemporal dependencies directly. Recently, TimeSFormer [4], ViViT [1], VidTR [67], and MViT [14] extends ViT to capture spatiotemporal context for video classification. Video Swin Transformer [32] extends Swin Transformer [31] to video by conducting shift window MHSA in both space and time. TokShift [65] proposes a temporal shift mechanism on CLASS tokens of ViT. Different from these video transformers focus on video classification, we target at building temporally efficient transformer for instance-level video understanding.

3. Method

3.1. Overall Architecture

The overall architecture of our VIS method TeViT is shown in Fig. 1, which contains a transformer-based backbone network and a query-driven head network. Given a sequence of video frames, the transformer backbone performs feature extraction and generates multi-scale pyramid features. The query-driven head network takes randomly

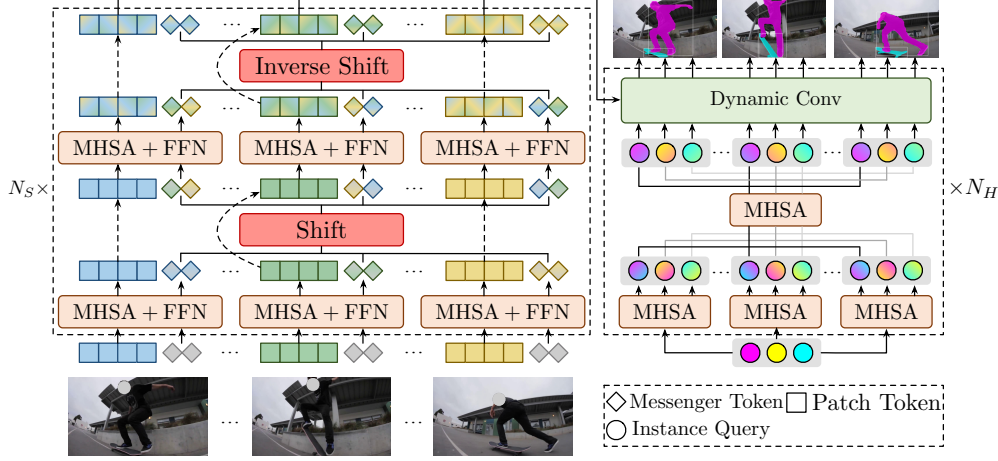


Figure 1. The overall illustration of our TeViT framework. TeViT contains a messenger shift transformer backbone and a series of spatiotemporal query-driven instance heads. The messenger shift mechanism performs efficient frame-level temporal modeling by simply shifting messenger tokens along the temporal axis. Spatiotemporal query interaction conducts two successive and parameter-shared multi-head self attention (MHSA) with feed forward network (FFN) upon video instance queries. The “Dynamic Conv” design follows QueryInst [17]. Best viewed in color.

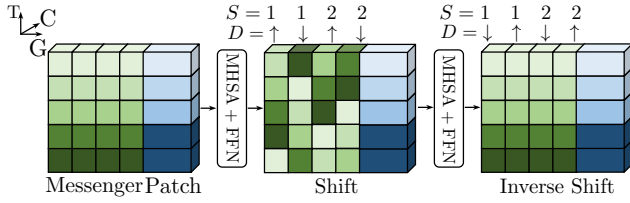


Figure 2. An illustration of the messenger shift mechanism. Messenger tokens are first segmented into several groups (4 in the figure) and then shifted along the temporal axis with different stride (S) and direction (D) to exchange frame-level information. For every two successive messenger shift mechanisms, we apply an inverse shift operation. As shown in figure above, the shift direction (D) of each token group in the second shift mechanism (right) is exactly inverse to the first (medium). Messenger tokens are shown by green cubes, while blue cubes denote the patch tokens.

initialized instance queries with backbone feature maps to predict video instances. Our whole network is end-to-end for both training and inference.

3.2. Messenger Shift Transformer Backbone

In previous VIS methods, the backbone networks only perform feature extraction in per-frame fashion [22, 57] and neglect the rich contextual information inherent in video frames. In contrast, inspired by MSG-Transformer [15], we propose messenger shift transformer (MsgShift) which performs highly efficient temporal context modeling in a bottom-up manner, as shown in Fig. 1 (left). Without loss of generality, we build MsgShift based on the pyramid vision transformer (PVT) [54, 55].

To be specific, given an input video with T frames of

resolution $H \times W$, denoted as $\{x_i\}_{i=1}^T \in \mathbb{R}^{T \times 3 \times H \times W}$, we first divide these frames into $T \times \frac{HW}{P^2}$ patch tokens frame-by-frame, where P denotes the size of each patch. Then we feed the flattened patch tokens to a linear projection and get embedded patches $\{f_i^0\}_{i=1}^T$ with size of $T \times \frac{HW}{P^2} \times C$, C denotes the channel dimension. Meanwhile, a group of randomly initialized learnable embeddings with size of $M \times C$ are introduced as messenger tokens, denoted as m^0 , where M indicates the number of messenger tokens. Then we simply copy and concatenate messenger tokens with patch tokens:

$$\{[f_i^0, m_i^0]\}_{i=1}^T \in \mathbb{R}^{T \times (\frac{HW}{P^2} + M) \times C}, \quad (1)$$

where m_i^0 indicates the copycat of messenger tokens m^0 . The concatenated joint tokens $\{[f_i^0, m_i^0]\}_{i=1}^T$ are taken as inputs for our MsgShift.

Our transformer architecture consists of $N_S = 4$ stages and each stage has the same architecture as in Fig. 1 (left). The multi-head self attention (MHSA) and feed forward network (FFN) act on the concatenated joint tokens in a per-frame manner:

$$\{[f_i^l, m_i^l]\}_{i=1}^T = \left\{ \text{FFN}^l \left(\text{MHSA}^l \left([f_i^{l-1}, m_i^{l-1}] \right) \right) \right\}_{i=1}^T. \quad (2)$$

Next, a messenger shift manipulation performs temporal information exchange across video frames.

In short, the messenger shift mechanism takes temporal messenger tokens $\{m_i^l\}_{i=1}^T$ as inputs and builds temporal context modeling by shifting messenger tokens along the temporal axis. Fig. 2 gives a detailed illustration. First, messenger tokens are divided into $G = 4$ groups and shifted

along the temporal axis with different time steps ($S = 1$ or 2) and direction (forward or backward). With various time steps and directions, messenger tokens are able to achieve temporal context exchange with both past and future frames. Moreover, for every two messenger shift operations, we apply an inverse operation to the second one, which implies the messenger tokens will be shifted back to their original corresponding frames after two contiguous messenger shift manipulations. This design aims to maintain a stable temporal receptive field as the network goes deeper.

After the above process, the messenger tokens and patch tokens go through one of four stages, and the output tokens are reshaped to feature maps $\{F_i^1\}_{i=1}^T$ which is $\frac{1}{4}$ smaller than the original image. In the same way, using the output messenger tokens and patch tokens of prior stage as inputs, we obtain the following pyramid feature maps $\{F_i^2\}_{i=1}^T$, $\{F_i^3\}_{i=1}^T$ and $\{F_i^4\}_{i=1}^T$, whose strides are 8, 16 and 32 pixels with respect to the input image. The pyramid feature maps $\{F_i^1, F_i^2, F_i^3, F_i^4\}_{i=1}^T$ will be used to predict video instances in the head network.

MsgShiT performs early temporal fusion in the backbone network, while the previous transformer-based VIS approaches [22, 37, 57, 64] only perform temporal feature fusion using transformer encoders after image-level feature extraction. It is almost parameter-free, friendly to image-level pre-training models, and brings negligible computation costs. The messenger tokens are randomly initialized and the shift manipulation has no parameter, so this module is insensitive to the pre-training process, which will be further discussed in the experiments in Tab. 6.

3.3. Spatiotemporal Query Interaction Head

MsgShiT achieves frame-level spatiotemporal context modeling. Meanwhile, in the VIS head network, our method still emphasizes temporally efficient spatiotemporal context modeling, but at the instance level. To this end, we propose a spatiotemporal query interaction (STQI) head network based on the recent SoTA query-based image-level instance segmentation method, *i.e.*, QueryInst [17].

As shown in Fig. 1 (right), our head network contains $N_H = 6$ STQI heads and takes a fixed-length instance queries $Q \in \mathbb{R}^{N_q \times C}$ along with pyramid features extracted by MsgShiT $\{F_i^1, F_i^2, F_i^3, F_i^4\}_{i=1}^T$ as inputs, and generates $N_q \times T$ instance predictions. N_q and C denotes the numbers and the channel dimensions of instance query respectively. Instance queries Q are randomly initialized and optimized during training. Additionally, our VIS network also contains a set of proposal boxes $B \in \mathbb{R}^{N_q \times 4}$ as prior proposals, for more details about this, we refer readers to QueryInst [17].

Instance queries are firstly copied by T times to each

frame. Two successive and parameter-shared MHSA modules act on instance queries along spatial and temporal dimensions:

$$\hat{Q}_{1:N_q}^{1:T} = \left\{ \text{MHSA} \left(Q_{1:N_q}^i \right) \right\}_{i=1}^T, \quad (3)$$

$$\tilde{Q}_{1:N_q}^{1:T} = \left\{ \text{MHSA} \left(\hat{Q}_j^{1:T} \right) \right\}_{j=1}^{N_q}. \quad (4)$$

“1:K” denotes ranging from 1 to K . Enhanced instance queries \tilde{Q} are fed into a dynamic convolution module and perform interactions with instance region features. Its output serves as the input queries of the next head. Finally, task specific heads (*i.e.*, classification head, box head and mask head) predict a sequence of video instances:

$$\{\hat{y}_i^t\}_{1:N_q}^{1:T} = \left\{ \left(\hat{p}_i^t(c), \hat{b}_i^t, \hat{m}_i^t \right) \right\}_{1:N_q}^{1:T}, \quad (5)$$

where $\hat{p}(c)$, \hat{b} and \hat{m} denotes predicted confidence scores, bounding boxes and instance foreground masks, respectively.

The advantages of STQI mainly stem from the minimum modifications on the still-image instance prediction head in [17]. STQI achieves highly efficient temporal context modeling at instance-level by a parameter-shared MHSA (Eq. 3 and Eq. 4) while does not involve extra parameters.

3.4. Matching and Loss Function

The loss function is motivated by [7]. We first compute the one-to-one assignment between predicted instances and ground-truth annotations. The ground-truth annotations are denoted as follows:

$$\{y_j^t\}_{1:N_{gt}}^{1:T} = \left\{ (c_j^t, b_j^t, m_j^t) \right\}_{1:N_{gt}}^{1:T}, \quad (6)$$

in which N_{gt} indicates the number of ground-truth video instances, c , b and m indicates the category, bounding box and mask respectively. We then perform sequence-level bipartite matching between predictions and annotations by Hungarian algorithm [25]. The cost matrix with size of $N_q \times N_{gt}$ between each predicted video instance and each annotation is defined as follows.

$$\begin{aligned} \mathcal{L}_{Hung}(\hat{y}_i^{1:T}, y_j^{1:T}) &= \lambda_{cls} \cdot \mathcal{L}_{cls}(\hat{p}_i^{1:T}(c), p_j^{1:T}) \\ &+ \lambda_{L1} \cdot \mathcal{L}_{L1}(\hat{b}_i^{1:T}, b_j^{1:T}) \\ &+ \lambda_{giou} \cdot \mathcal{L}_{giou}(\hat{b}_i^{1:T}, b_j^{1:T}), \end{aligned} \quad (7)$$

where \mathcal{L}_{cls} indicates the focal loss [28] for classification, \mathcal{L}_{L1} and \mathcal{L}_{giou} indicates the L1 loss and GIoU loss [42] respectively. $\lambda_{cls}, \lambda_{L1}, \lambda_{giou} \in \mathbb{R}$ are hyper-parameters which we simply follow [17, 44, 68]. Besides we use dice coefficient [34] to optimize mask predictions. For more details, please refer to [17].

Method	Backbone	MST	FPS	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
MaskTrack R-CNN [62]	ResNet-50		32.8	30.3	51.1	32.6	31.0	35.5
MaskTrack R-CNN [62]	ResNet-101		28.6	31.9	53.7	32.3	32.5	37.7
SipMask [6]	ResNet-50	✓	34.1	33.7	54.1	35.8	35.4	40.1
SG-Net [30]	ResNet-50		–	34.8	56.1	36.8	35.8	40.8
SG-Net [30]	ResNet-101		–	36.3	57.1	39.6	35.9	43.0
CrossVIS [63]	ResNet-50	✓	39.8	36.3	56.8	38.9	35.6	40.7
CrossVIS [63]	ResNet-101	✓	35.6	36.6	57.3	39.7	36.0	42.0
STEm-Seg [2]	ResNet-50	✓	3.0	30.6	50.7	33.5	31.6	37.1
STEm-Seg [2]	ResNet-101	✓	–	34.6	55.8	37.9	34.4	41.6
MaskProp [3]	ResNet-50	✓	–	40.0	–	42.9	–	–
MaskProp [3]	ResNet-101	✓	–	42.5	–	45.6	–	–
MaskProp [3]	STSN-X-101	✓	–	46.6	–	51.2	44.0	52.6
SeqMask R-CNN [26]	ResNet-50	✓	–	40.4	63.0	43.8	41.1	49.7
SeqMask R-CNN [26]	ResNet-101	✓	–	43.8	65.5	47.4	43.0	53.2
VisTR [57]	ResNet-50		51.1	36.2	59.8	36.9	37.2	42.4
VisTR [57]	ResNet-101		43.5	40.1	64.0	45.0	38.3	44.9
EfficientVIS [58]	ResNet-50	✓	36.0	37.9	59.7	43.0	40.3	46.6
EfficientVIS [58]	ResNet-101	✓	32.0	39.8	61.8	44.7	42.1	49.8
IFC [22]	ResNet-50	✓	107.1	41.2	65.1	44.6	42.3	49.6
IFC [22]	ResNet-101	✓	89.4	42.6	66.6	46.3	43.5	51.4
TeViT (ours)	MsgShifT		68.9	45.9	69.1	50.4	44.0	53.4
TeViT (ours)	MsgShifT	✓	68.9	46.6	71.3	51.6	44.9	54.3

Table 1. Comparisons on YouTube-VIS-2019 dataset [62]. “✓” under “MST” indicates using multi-scale training strategy, The FPS is measured with a single TESLA V100 GPU. All methods in the figure are organized into four groups. According to their basic architectures, the first two groups of methods are built upon CNN architecture, while the last two are transformer-based. According to their training inference paradigms, the first group follows the online and track-by-detect fashion, while the rest all follow offline and sequence-in-sequence-out paradigm.

3.5. Online and Offline Inference

Our method is flexible for both offline and online inference. Under the offline scenario, our TeViT takes the *whole* video clips as inputs and then outputs all possible video instances with a single run. No post-tracking process is needed. When it comes to the near online [2] scenario, an entire video is split into several overlapping segments. TeViT takes clips in time order and generates predictions. A rule-based post-tracking procedure is applied to linking instances across different video clips. For instances from two overlapping video clips, we first compute the similarity score between each instance, and then a Hungarian matcher gives the assignment according to the similarity matrix. The similarity score is defined as a combination of box IoU and mask IoU.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate TeViT on three challenging video instance segmentation benchmarks, *i.e.*, YouTube-VIS-2019 [62], YouTube-VIS-2021 [61], and OVIS [39]. **YouTube-VIS-2019** is the first dataset that focuses on the VIS problem. It

contains 40 common object categories, 4, 883 unique video instances and about 131k high-quality instance-level annotations. **YouTube-VIS-2021** dataset is the new version of YouTube-VIS-2019 with $1.5\times$ more video frames and $2\times$ more annotations. **OVIS** dataset aims to explore the VIS problem under high-occlusion scenarios. It consists of 296k high-quality instance masks and 5.80 instances per video from 25 semantic categories. Following previous works, we report the performance on the validation set for all three datasets. We follow the standard VIS evaluation metrics defined in [62].

4.2. Implementation Details

TeViT is built upon the `mmdetection` toolbox [10]. Unless otherwise noted, hyper-parameters follow the settings of QueryInst [17]. We use $Nq = 100$ video instance queries as [17, 22]. Due to the temporal efficient designs in TeViT, we do not need to create pseudo video data, *e.g.* [2, 22], to train the temporal modeling parameters, instead, we first train a transformer-based QueryInst for image-level instance segmentation on the COCO dataset [29] and then initialize TeViT with the COCO pre-trained QueryInst weights. Besides, we provide a MindSpore [35] implementation of TeViT.

Methods	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
MaskTrack R-CNN [†] [62,63]	28.6	48.9	29.6	26.5	33.8
SipMask [†] [6,63]	31.7	52.5	34.0	30.8	37.8
CrossVIS [63]	34.2	54.4	37.9	30.4	38.2
IFC [22]	35.2	57.2	37.5	–	–
TeViT	37.9	61.2	42.1	35.1	44.6

Table 2. Comparisons with previous VIS methods on YouTube-VIS-2021 datasets. Methods with superscript “†” are reported in [63].

When training on the VIS datasets, we use the AdamW [24] optimizer with an initial learning rate of 0.00025, and a weight decay of 0.0001. Especially, the backbone learning rate is slightly lower with a multiplier set to 0.1. We also apply gradient clipping with a maximal gradient norm of 0.1. TeViT is trained with a batch size of 16 and a clip length of $T = 5$. The total training process contains 12 epochs, and the learning rate is decreased by 10 at the 8-th and 11-th epoch respectively. For example, our TeViT can be trained in about 4 hours with 8 Tesla V100 GPUs on YouTube-VIS-2019, which is much faster than previous transformer-based method (*i.e.*, VisTR [57]). The number of instance queries N_q is set to 100 for all experiments. Following [62], all input frames are resized to 360×640 in single-scale experiments. Settings of multi-scale training simply follow [6]. For inference, all frames are resized to 360×640 regardless of the training setups. During inference, we use $T = 36$ for most results and report the near online results in ablation study. For main results, we evaluate our framework on YouTube-VIS-2019, YouTube-VIS-2021, and OVIS datasets, with PVT-B1 based MsgShiT as backbone.

It’s noted that all reported results in main results and ablation studies are average performance from multiple runs (*i.e.*, we choose five different random seeds and run each random seed for three times). The standard deviations (*i.e.*, σ_{AP} in following tables) are calculated in the same way.

4.3. Main Results

Main results on YouTube-VIS-2019 dataset. We compare our TeViT to state-of-the-art methods on YouTube-VIS-2019 dataset in Tab. 1. The longest video in YouTube-VIS-2021 dataset only contains 36 frames, so that our TeViT executes fully offline inference on this dataset. Without bells and whistles, our TeViT achieves 45.9 AP when using a single-scale training strategy and outperforms the previous state-of-the-art methods by a large margin. Multi-scale training strategy further boosts the performance to 46.6 AP. Meanwhile, our method also achieves competitive inference speed. With about 10 AP higher, our method is still faster than VisTR.

Main results on YouTube-VIS-2021 dataset. Tab. 2

Methods	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
SipMask [†] [6,63]	10.3	25.4	7.8	7.9	15.8
MaskTrack R-CNN [†] [62,63]	10.9	26.0	8.1	8.3	15.2
STEm-Seg [‡] [2,39]	13.8	32.1	11.9	9.1	20.0
CrossVIS [63]	14.9	32.7	12.1	10.3	19.8
CMaskTrack R-CNN [‡] [39]	15.4	33.9	13.1	9.3	20.0
TeViT	17.4	34.9	15.0	11.2	21.8

Table 3. Comparisons on OVIS dataset. Methods with superscript “†” and “‡” are reported in [63] and [39] respectively.

shows the final results of several VIS methods and ours on YouTube-VIS-2021 dataset. Due to the video length in YouTube-VIS-2021 is longer than our inference clip length ($T = 36$), TeViT performs near online tracking described in Sec. 3.5 on this dataset. TeViT obtains 37.9 AP, outperforming the previous state-of-the-art method by 2.7 AP.

Main results on OVIS dataset. The results on the OVIS dataset are shown in Tab. 3. Our method also performs near online inference on OVIS dataset. TeViT achieves a relatively higher performance of 17.4 AP on the val split, surpassing previous state-of-the-art methods. Compared to CMaskTrack R-CNN [39] which presents an elaborate-designed feature calibration plug-in to alleviate occlusion, our TeViT still gains 2.0 AP improvement, which shows that our temporal context modeling designs are helpful to segment occluded instances.

4.4. Ablation Study

Effect of frame-level & instance-level temporal context modeling. We investigate the effects of messenger shift mechanism and spatiotemporal query interaction individually and simultaneously in Tab. 4. Using messenger shift mechanism and spatiotemporal query interaction individually brings 0.6 and 2.7 AP improvements respectively. The results show that both frame-level and instance-level temporal context modeling can obviously improve VIS performance. In addition, the instance-level one brings more significant performance gain. The two designs together brings 3.4 ($> 0.6 + 2.7$) AP improvements over a high-performance baseline. Besides the remarkable performance improvements, our designs only bring 0.27% computation overhead on our baseline (82.19 GFLOPs *vs.* 81.97 GFLOPs), which demonstrates our design is very efficient.

Variants of spatiotemporal query interaction. In Tab. 5, we investigate the effectiveness of our spatiotemporal query interaction comparing to its variants. A naive query interaction method in [17] without using temporal interaction, denoted as “Spatial Only”, serves as a baseline. “Fused Space-Time” in the Row 2 denotes fusing video instance queries together and performing spatial and temporal interaction within a single MHSA, which is the same as in [57]. As the results show: (1) Our spatiotemporal query interaction

MSM STQI	GFLOPs	AP $\pm \sigma_{AP}$	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
	81.97	42.5 \pm 0.47	67.6	44.0	43.0	52.7
✓	82.19	43.1 _{↑(+0.6)} \pm 0.71	67.2	47.8	43.5	52.4
✓	81.97	45.2 _{↑(+2.7)} \pm 0.85	68.9	50.2	44.0	53.0
✓	82.19	45.9 _{↑(+3.4)} \pm 0.58	69.1	50.4	44.0	53.4

Table 4. Component-wise analysis on TeViT. MSM denotes the messenger shift mechanism and STQI denotes spatiotemporal query interaction. Without applying STQI implies only one MHSA is performed for query interaction within each frame (excluding Eq. 4).

Manip.	AP $\pm \sigma_{AP}$	AP ₅₀	AP ₇₅
None	45.2 \pm 0.85	68.9	50.2
MHSA + FFN	44.5 \pm 1.07	69.2	49.3
Shift	45.9 \pm 0.58	69.1	50.4

Table 6. Study of the manipulations upon messenger tokens. Our method obtains the highest AP and a relatively stable performance (σ_{AP}) among all settings.

Manip.	AP	AP ₅₀	AP ₇₅
None	45.2	68.9	50.2
Conv	41.8	63.7	45.1
MHSA + FFN	43.1	67.2	49.1
Msg Shift	45.9	69.1	50.4

Table 7. Study of frame-level feature aggregation. Compared to other frame-level feature manipulations, our messenger shift (Row 4) obtains the best results.

Interaction	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
Spatial Only [17]	43.1	67.2	47.8	43.5	52.4
Fused Space-Time [57]	43.9 _{↑(+0.8)}	69.5	48.4	42.9	52.0
Ours	45.9 _{↑(+2.7)}	69.1	50.4	44.0	53.4

Table 5. Variants of spatiotemporal query interaction. ‘‘Spatial Only’’ denotes the image-level instance segmentation heads in [17], ‘‘Fused Space-Time’’ denotes applying MHSA to all video instance queries at a single run, which is the same as in [57].

M	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
8	45.3	69.0	48.9	44.5	52.4
16	45.4	70.3	49.9	44.0	51.7
32	45.9	69.1	50.4	44.0	53.1

Table 8. Impact of messenger token numbers. M indicates the number of messenger tokens. We increase M from 8 to 32 and observe the effects on final performance.

achieves the best performance among three variants. (2) Compared to the spatial-only query interaction, joint spatiotemporal query interaction brings only 0.8 AP improvements. However, our method achieves 2.7 AP gains. We argue this is because the one-to-one corresponding between instance queries are misaligned in the joint spatiotemporal attention, while ours is not.

Different messenger token manipulation methods. We compare our messenger shift mechanism with two other optional manipulations in Tab. 6. None indicates there are no extra manipulations conducted on messenger tokens, thus no temporal information is employed. MHSA+FFN stands for the same operation in [22] which performs extra MHSA and FFN on messenger tokens, and Shift denotes our messenger shift mechanism. Different from [22], we do not conduct any extra pre-training process so that both the messenger tokens and MHSA layers with FFN are randomly initialized and trained from scratch. As results have shown: (1) Our method achieves the best AP and outperforms None by 0.7 AP and MHSA + FFN by 1.4 AP. (2) Compared to conducting MHSA + FFN upon messenger tokens, our method obtains a more stable performance. MHSA + FFN shows more fluctuating final results (see σ_{AP} in Tab. 6) while ours is much more stable. The results confirm that the randomized MHSA + FFN is unable to capture temporal context while our parameter-free shift operation works.

Different manipulations on frame-level feature aggregation. We also compare our messenger shift mechanism with other optional frame-level feature aggregation manipulations in Tab. 7. None indicates no manipulation is conducted to aggregate frame-level temporal features. Conv. and MHSA + FFN denotes using newly introduced con-

volution layers or transformer layers to achieve temporal feature aggregation. As results have shown, our messenger shift mechanism achieves the best performance (*i.e.*, 45.9 AP) compared to its all counterparts. Meanwhile, we observe apparent performance decrease by using newly introduced Conv. or MHSA + FFN as aggregation layers. We infer such a performance decrease comes from the enormous newly introduced parameters, while our messenger shift mechanism eliminates this performance decrease by the nearly parameter-free design.

Number of messenger tokens. In Tab. 8, we test our method with number of messenger tokens increases from 8 to 32. Compared to less messenger tokens ($M = 8, 16$), more messenger tokens ($M = 32$) achieves better results. Unless specified, our experiments are conducted with 32 messenger tokens.

Training and inference clip length. We also investigate the effects of clip length in both the training and testing phase. From Tab. 9, we find that: (1) Our method shows great tolerance to short length of training clip. Only trained with 2 or 3 frames, our method can effectively learn temporal context and obtains comparable results to previous methods. (2) The performance improvements by increasing the length of the training clip gradually gets saturated. Increasing training clip length from 2 to 3 and 3 to 5 brings 3.2 AP and 1.6 AP gains respectively while increasing training clip length from 5 to 7 only obtains slight 0.4 AP profit. Besides, a longer training clip requires more training computations and memory budgets. To this end, we set the training clip length of our method to $T = 5$ as a compromise between performance and training costs.

Tab. 10 gives the results of TeViT under different in-

T	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
2	41.1	64.8	44.3	41.2	50.2
3	44.3	69.2	49.3	43.7	52.1
5	45.9	68.9	50.2	44.0	53.0
7	46.3	71.9	51.6	44.0	53.4

Table 9. Effect of training clip length on AP. “T” indicates the number of frames for each video clip during training.

Method	MST	FPS	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
VisTR [57]		51.1	36.2	59.8	36.9	37.2	42.4
IFC [22]	✓	107.1	41.2	65.1	44.6	42.3	49.6
Ours w/o. Eq. 4		78.1	36.8	78.3	38.8	38.3	46.2
Ours		76.8	41.7	67.8	44.8	41.3	49.9
Ours	✓	76.8	42.3	67.6	44.0	43.0	52.7

Table 11. Comparisons with ResNet-50 as backbone.

inference settings. “T” indicates the input clip length during the inference phase, and “S” indicates strides. It shows that our TeViT obtains promising performance under various inference setups. Even with $T = 5$ and $S = 3$, TeViT still achieves 41.7 AP, which indicates TeViT can serve as a strong baseline for both offline and online video understanding scenarios.

Performance under ResNet backbone. We compare our method with previous transformer-based methods using ResNet-50 [20] as backbone network in Tab. 11. As results show, our method achieves 41.7 AP on YouTube-VIS-2019 dataset with single-scale training (Row 4), even outperforming VisTR and IFC with multi-scale training strategy. The performance goes a step further to 42.3 AP when the multi-scale strategy is applied (Row 5). It’s worth noting that with ResNet-50 as backbone network, the messenger tokens, and messenger shift mechanism are unable to proceed, so our method achieves such high performance with only the STQI mechanism. We also investigate the improvements by our STQI head. Simply taking off the spatial MHSA in Eq. 4, the final performance drops from 41.7 AP to 36.8 AP (Row 3), demonstrating the effectiveness of our STQI head directly.

Revisiting messenger tokens. Inspired by MSG-Transformer [15], we re-initialize messenger tokens in inference phase and obvious the influence on performance in Tab. 12. As the results show, when we re-initialize messenger tokens to zero, the performance merely drops 0.4 AP (compare Row 2 to Row 1). Randomly initialize the messenger token in inference phase leads to a similar performance decrease (Row 3). We think this phenomenon implies that the messenger tokens contain only a few or not

T	S	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
5	1	42.1	66.8	46.7	41.5	51.3
5	3	41.7	63.2	46.1	41.8	50.8
10	5	44.1	66.4	48.3	42.9	51.9
15	8	44.7	67.0	48.7	43.4	52.5
20	10	46.0	67.5	50.0	43.7	53.1
36	18	45.9	68.9	50.2	44.0	53.0

Table 10. Study the impact of clip length and stride during inference phase. “T” and “S” indicates the clip length and overlapping stride respectively.

Train	Inference	AP	AP ₅₀	AP ₇₅	AR ₁	AR ₁₀
learnable	learned	45.9	68.9	50.2	44.0	53.0
learnable	zero	45.5 _{↓(-0.4)}	68.5	49.9	43.8	52.5
learnable	random	45.6 _{↓(-0.3)}	68.2	49.7	43.7	52.4

Table 12. Revisiting messenger tokens in inference phase.

specific information in themselves. On the contrary, they play the role of summarizing frame-level contexts, and exchanging them across adjacent frames.

5. Conclusion

In this paper, we provide lightweight and effective solutions to fully exploit temporal context for VIS. Based on existing ViTs and query-based image-level instance segmentation methods, we propose the TeViT VIS method that contains the messenger shift and spatiotemporal query interaction mechanisms. TeViT performs both frame-level and instance-level temporal feature interactions while only bringing a few parameters and marginal extra computational costs. Experiments on YouTube-VIS-2019, YouTube-VIS-2021, and OVIS show that TeViT can obtain remarkably better results than previous SoTA methods, *e.g.*, IFC, VisTR, MaskProp, and STEm-Seg. We believe the proposed temporal context modeling mechanisms have great potential to be extended to other video understanding tasks.

Limitations. Although the extensive experiments have demonstrated the capacity and efficiency of our TeViT on temporal context modeling, it still suffers effects from occlusion, motion deformation and long time-span videos (*i.e.*, results of TeViT in Tab. 2 and Tab. 3 are far from satisfying). We leave these promising directions as future work.

Broader impact. Although our research does not make direct negative impacts in society, it may be misused by illegal video applications, which could be a potential invasion to human privacy.

Acknowledgement. This work was in part supported by NSFC (No. 61876212 and No. 61733007) and CAAI-Huawei MindSpore Open Fund.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 1, 2
- [2] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 2, 5, 6
- [3] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 2, 5
- [4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 1, 2
- [5] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. 2
- [6] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 2020. 5, 6
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 4
- [8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2
- [9] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiao-xiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2
- [10] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5
- [11] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021. 1, 2
- [12] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *arXiv preprint arXiv:2106.02351*, 2021. 1, 2
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2
- [14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 1, 2
- [15] Jiemin Fang, Lingxi Xie, Xinggang Wang, Xiaopeng Zhang, Wenyu Liu, and Qi Tian. Msg-transformer: Exchanging local spatial information by manipulating messenger tokens. *arXiv preprint arXiv:2105.15168*, 2021. 1, 2, 3, 8
- [16] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *arXiv preprint arXiv:2106.00666*, 2021. 1, 2
- [17] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7
- [18] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. *arXiv preprint arXiv:2012.03400*, 2020. 2
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 8
- [21] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021. 1, 2
- [22] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *arXiv preprint arXiv:2106.03299*, 2021. 1, 2, 3, 4, 5, 6, 7, 8
- [23] Joakim Johnander, Emil Brissman, Martin Danelljan, and Michael Felsberg. Learning video instance segmentation with recurrent graph neural networks. *arXiv preprint arXiv:2012.03911*, 2020. 2
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [25] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4
- [26] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Ji-aya Jia. Video instance segmentation with a propose-reduce paradigm. *arXiv preprint arXiv:2103.13746*, 2021. 2, 5
- [27] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 2
- [28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5
- [30] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*, 2021. 5
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 1, 2

- [32] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 1, 2
- [33] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *ICCV*, 2021. 1, 2
- [34] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 4
- [35] MindSpore. <https://github.com/mindspore-ai/mindspore>. 5
- [36] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asseilmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021. 1
- [37] Thuy C Nguyen, Tuan N Tang, Nam LH Phan, Chuong H Nguyen, Masayuki Yamazaki, and Masao Yamanaka. 1st place solution for youtubevos challenge 2021: Video instance segmentation. *arXiv preprint arXiv:2106.06649*, 2021. 1, 2, 4
- [38] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *CVPR*, 2019. 2
- [39] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: Dataset and challenge. In *NeurIPS Datasets and Benchmarks Track*, 2021. 2, 5, 6
- [40] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. In *ICCV*, 2017. 2
- [41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016. 2
- [42] Hamid Rezaatoughi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 4
- [43] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *arXiv preprint arXiv:2105.05633*, 2021. 1, 2
- [44] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 1, 2, 4
- [45] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 2
- [46] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: A simple and strong anchor-free object detector. *IEEE TPAMI*, 2020. 2
- [47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 1, 2
- [48] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2
- [49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2
- [50] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, 2021. 1, 2
- [51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2
- [52] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019. 2
- [53] Tao Wang, Ning Xu, Kean Chen, and Weiyao Lin. End-to-end video instance segmentation via spatial-temporal graph neural networks. In *ICCV*, 2021. 2
- [54] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021. 2, 3
- [55] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 1, 2, 3
- [56] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2
- [57] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8
- [58] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. *arXiv preprint arXiv:2203.01853*, 2022. 5
- [59] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 1, 2
- [60] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 2
- [61] Ning Xu, Linjie Yang, Jianchao Yang, Dingcheng Yue, Yuchen Fan, Yuchen Liang, and Thomas S. Huang. Youtubevis dataset 2021 version. <https://youtube-vos.org/dataset/vis>, 2021. 2, 5
- [62] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 5, 6
- [63] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *ICCV*, 2021. 2, 5, 6
- [64] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Ying Shan, Bin Feng, and Wenyu Liu. Tracking instances as queries. *arXiv preprint arXiv:2106.11963*, 2021. 1, 2, 4

- [65] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *ACM MM*, 2021. 1, 2
- [66] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *arXiv preprint arXiv:2106.14855*, 2021. 1, 2
- [67] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *ICCV*, 2021. 1, 2
- [68] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2, 4