TCOVIS: Temporally Consistent Online Video Instance Segmentation

Junlong Li Bingyao Yu Yongming Rao Jie Zhou Jiwen Lu*
Department of Automation, Tsinghua University, China
Beijing National Research Center for Information Science and Technology, China

Abstract

In recent years, significant progress has been made in video instance segmentation (VIS), with many offline and online methods achieving state-of-the-art performance. While offline methods have the advantage of producing temporally consistent predictions, they are not suitable for real-time scenarios. Conversely, online methods are more practical, but maintaining temporal consistency remains a challenging task. In this paper, we propose a novel online method for video instance segmentation, called TCOVIS, which fully exploits the temporal information in a video clip. The core of our method consists of a global instance assignment strategy and a spatio-temporal enhancement module, which improve the temporal consistency of the features from two aspects. Specifically, we perform global optimal matching between the predictions and ground truth across the whole video clip, and supervise the model with the global optimal objective. We also capture the spatial feature and aggregate it with the semantic feature between frames, thus realizing the spatio-temporal enhancement. We evaluate our method on four widely adopted VIS benchmarks, namely YouTube-VIS 2019/2021/2022 and OVIS, and achieve state-of-the-art performance on all benchmarks without bells-and-whistles. For instance, on YouTube-VIS 2021, TCOVIS achieves 49.5 AP and 61.3 AP with ResNet-50 and Swin-L backbones, respectively. Code is available at https://github.com/jun-long-li/TCOVIS.

1. Introduction

Video instance segmentation (VIS) is a challenging and representative video understanding task recently introduced in [37]. It aims at detecting, segmenting and tracking instances across a video. VIS is attracting increasing attention for various real-world applications such as video editing, video surveillance, augmented reality and autonomous driving. Recently introduced VIS methods can be roughly categorized into two groups: offline methods and online methods. Offline methods [2, 16, 19, 31, 33, 35] take as input the whole video and perform the segmentation of instance sequence for the whole video at once. Online methods [8, 34, 17, 10, 38], on the contrary, take as input a video frame by frame and generate the pre-frame object instances while associating the frame-wise results across frames. Both offline and online methods have achieved impressing performance on the VIS task.

Offline methods have an inherent advantage in producing temporally consistent predictions, since delicate temporal communication and association mechanisms can be adopted throughout the video [39, 33, 14] to handle the overall temporal information and impose an explicit constraint on the temporal consistency. However, the video-in and video-out offline manner is not suitable for real-time scenarios. Conversely, online methods are more practical and making considerable progress but suffer from temporal inconsistency (as shown in Figure 1), remaining a great challenge.

Online methods rely on specific instance association applied across frames, since only one frame is observed at a time. Existing association techniques can be grouped into two categories, including tracking-by-detection and query propagation-based paradigms. Tracking-by-detection methods [37, 34, 32] generate the per-frame instances independently by existing instance segmentation models [11, 26, 5] and track instances via tracking heads [37] or instance embeddings matching [34, 15]. In this way, the features of different frames are isolated before tracking, which results in temporal inconsistency. Query propagation-based methods [13, 10, 41] are inspired by query-based methods [3, 25] and they propagate the query across frames to decode a unique instance without heuristic matching algorithms. Despite the explicit temporal link of queries, the temporal consistency is impaired by the Local Matching and Propagating (LocPro) scheme, where they first perform local optimal matching between the predictions and ground truth at the beginning of the video, and then propagate the assignment across frames, forcing all features from subsequent frames to follow. The LocPro is not suitable for the holistic optimization across frames and results in temporal inconsis-

*Corresponding author

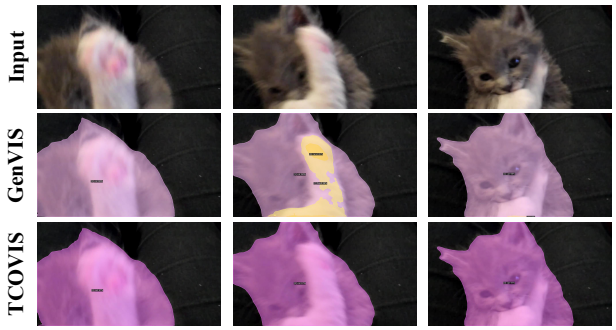

Figure 1. Visualization of predictions from the previous online method (the online GenVIS [13]) and our TCOVIS. The previous method generates temporally inconsistent predictions, while our proposed TCOVIS achieves temporal consistency and outperforms the previous method (Best viewed in color).

tency. Thus, achieving temporal consistency is challenging for online methods and also not comprehensively investigated by previous online VIS methods.

In this paper, we propose a novel online method for video instance segmentation, named TCOVIS, to fully exploit the temporal information within a video. We take as the baseline framework an existing online VIS model (GenVIS [13]) with the query propagation-based instance association. We first introduce the global instance assignment strategy to perform global optimal matching. Different from the previous online methods [13, 41, 10], which obtain per-frame matching cost, assign labels locally on the beginning frame and propagate across frames, we collect the predictions across frames, compute the global matching cost with the video segmentation ground truth and supervise the model with the global instance assignment, encouraging features across the video to be optimized for a global optimal objective. As online methods focus on improving the representative ability of the semantic instance embeddings [38, 34], which is achieved via learning more discriminative semantic embedding in those using heuristic matching [34, 15], or via reviewing the memory of semantic embedding across frames [13] in query propagation-based methods, the spatial features are not comprehensively investigated. We further propose spatio-temporal enhancement module, leveraging the spatial information from the previous frame to enhance the temporal consistency. We perform spatial matting on the pixel embeddings to retrieve the instance-wise spatial features and adopt the cross-attention layer to aggregate the spatial and semantic features across frames, thus realizing the spatio-temporal enhancement. As shown in Figure 1, the previous online method [13] produces temporally inconsistent results, as exemplified by an object abruptly appearing in front of the cat in a mid-frame, while the proposed TCOVIS outperforms the previous one and generates temporally consistent predictions.

To validate the effectiveness of the proposed method, experiments are conducted on four widely adopted VIS benchmarks, i.e., YouTube-VIS 2019 [37], YouTube-VIS 2021, YouTube-VIS 2022 and Occluded VIS (OVIS) [24]. Without bells-and-whistles, our proposed method achieves state-of-the-art performance on all benchmarks, outperforming other online methods, e.g., on YouTube-VIS 2021, TCOVIS achieves 49.5 AP and 61.3 AP with ResNet-50 and Swin-L backbones, respectively.

Our main contributions are summarized as follows:

- TCOVIS performs a novel global instance assignment strategy for online video instance segmentation. The model is optimized for the global optimal objective to generate more temporally consistent predictions.
- The further proposed spatio-temporal enhancement module captures the spatial feature and aggregates it with the semantic feature between frames, which fully utilizes the spatial information and facilitates the temporal consistency enhancement.
- The proposed method achieves state-of-the-art performance on four widely used video instance segmentation benchmarks (YouTube-VIS 2019/2021/2022 and OVIS). Such achievements demonstrate the effectiveness of our proposed method.

2. Related Works

Offline Video Instance Segmentation. Offline methods take as input the whole video and predict instance sequence for all frames at once. Mask propagation and box ensemble techniques are used to improve the predictions and association [1, 2, 19], but they are not end-to-end learnable due to the complex inference process. VisTR [31] extends DETR [42] from the image domain and introduces the transformer [27] to the VIS domain. EfficientVIS [35] and IFC [16] relax the heavy overhead of VisTR via an iterative query-video interaction and memory token communication, respectively. TeViT [39] contains a vision transformer [6] backbone instead of CNN and efficiently builds correspondence between the instance and query. VITA [14] models relationships among instances with the distilled condensed object tokens, without using the dense spatio-temporal backbone features. Offline methods exploit rich temporal knowledge from the whole clip and have the advantage of producing temporally consistent results, however, the offline manner is not suitable for the application in real-time scenarios.

Online Video Instance Segmentation. Instead of processing the entire video before predictions, online methods only leverage the information from the previous frames and segment the video frame-by-frame. The association paradigms of the previous online methods roughly fall into two groups: Tracking-by-detection and Query-propagation.

Most online methods [37, 29, 21, 34] follow the tracking-by-detection paradigm. MaskTrack R-CNN [37] is the baseline method and extends the Mask R-CNN [11] with an extra tracking head for temporal association. CrossVIS [38] proposes a crossover learning scheme to utilize the current contextual information for other frames. VISOLO [9] builds on the image instance segmentation method SOLO [30] and takes advantage of the grid form previous information for memory matching and features aggregation. MinVIS [15] and IDOL [34] make use of the discriminative instance embeddings for matching between frames. With the heuristic matching technique designed for instance association across frames, temporal inconsistency comes from the frame-wise modeling before tracking.

Object association with query propagation has been explored in the multi-object tracking (MOT) task [22, 40]. TrackFormer [22] tracks the seen objects of previous frames with a track query subset and detects the newly appeared objects in current frame with an extra object query subset. MOTR [40] extends the paired-frames training scheme to multiple frames for long-range temporal association. The query propagation-based object association is recently introduced to VIS [41, 10, 13]. ROVIS [41] follows TrackFormer [22] to detect and track instances with two subsets of queries. InsPro [10] and GenVIS [13] propagate the queries without heuristics, i.e., handcrafted rules to combine two types of queries, and achieve association across frames. However, previous query propagation-based methods also propagate the local assignment to supervise the model with the local optimal results, which leads to temporal inconsistency across the entire video. We adopt the query propagation framework but introduce the global instance assignment to enhance the temporal consistency.

3. Method

Given a video clip with consecutive image frames, online video instance segmentation methods generate frame-level object instances upon on instance segmentation models [26, 5], utilizing the instance queries propagated from previous frames [35, 13]. We have already discussed that better performance of the online VIS method relies on more temporally consistent instance features among a video. To this end, we propose a novel end-to-end online VIS method TCOVIS (Figure 2), to improve the temporal consistency of instance features via the global instance assignment strategy and spatio-temporal enhancement module. In this section, we first introduce the online VIS pipeline in Section 3.1. Then the details of the proposed global instance assignment strategy and spatio-temporal enhancement module will be described in Section 3.2 and Section 3.3, respectively. Finally, in Section 3.4, we describe the overall loss for training the model end-to-end.

3.1. Online Video Instance Segmentation

Following state-of-the-art VIS methods [15, 4], we adopt the advanced Masked-attention Mask Transformer (Mask2Former [5]) as the image instance segmentation network (Img. SegNet.) in this paper. Assume that the input video clip with T frames is denoted as $x \in \mathbb{R}^{T\times 3\times H\times W}$. With each frame $H \times W$ as input, frame-wise activation map is extracted by the backbone and Transformer encoder. Then following the query-based mechanism of DETR [3], N_{fq} object queries of C dimensions are used to parse an input frame, which are called frame object queries $f \in \mathbb{R}^{C\times N_{fq}}$. Each object in the frame is decoded by the frame object queries from the spatial features through a multiple-level Transformer decoder, and represented as an object embedding of C dimensions. The object embeddings are used for classification and together with the pixel embeddings from the pixel decoder generating the mask for object instances. Class predictions are produced through a linear layer from the object embeddings. Mask embeddings are generated by an MLP linked to the object embeddings, and the model finally segments objects by pixel-wise dot product between per-pixel embeddings $\mathcal{P} \in \mathbb{R}^{C\times \frac{H}{S}\times \frac{W}{S}}$ and mask embeddings $\mathcal{M} \in \mathbb{R}^{C\times N_{fq}}$, where S is the stride of the spatial feature map.

As for online video segmentation, we follow the online scheme of GenVIS [13] which propagates the instance queries from previous frames. To resolve the computational limitation, the framework adopts VITA [14] that regards the frame object queries f as a concise representation of objects in a frame and then feeds them into the *Object Encoder* $\mathscr{E}$ for intra-frame relationship. The *Object Decoder* $\mathscr{D}$ takes as input N_v video instance queries q and aggregates information from frame object queries. The temporal instance association is implemented through a query-based temporal propagation mechanism, where the output of $\mathscr{D}$, instance prototypes denoted as p, are concise representations of instances [13], are not only utilized for classifying and segmenting in current frame at t, but also serve as the instance queries for the next frame at $t+1$, i.e.,

$$q^{t+1} = p^t = \mathscr{D}(q^t, \mathscr{E}(f^t)). \tag{1}$$

With this propagation mechanism, the model can simply run in an online manner and associate the frame-wise outputs without heuristic matching algorithms.

We leave out the Instance Prototype Memory module in [13]. During the training process, we freeze the image instance segmentation model and only train the following modules, to efficiently make use of memory. More implementation details are described in Section 4.2.

3.2. Global Instance Assignment

With the query propagation mechanism discussed above, the associated video instance queries along the temporal

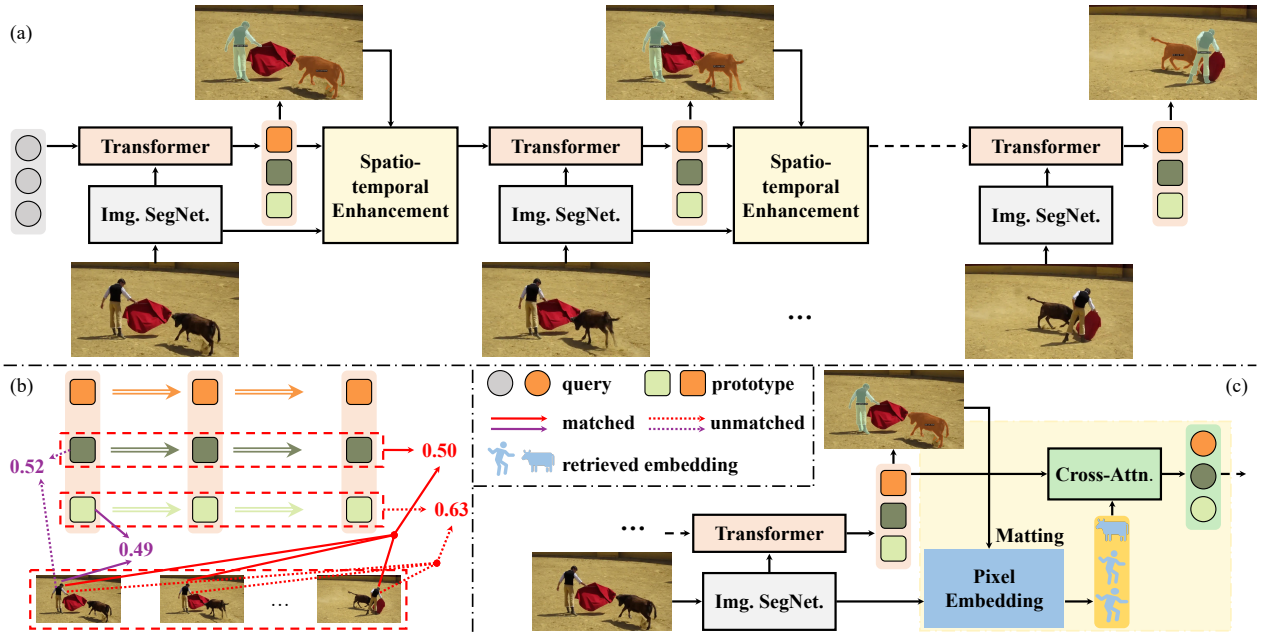

Figure 2. (a) The overall illustration of our TCOVIS. It supervises the model with a global optimal objective during training and utilizes spatial features via the enhancement module between frames. (b) Details of global instance assignment. Different from the local matching and propagating technique, we conduct global optimal matching and assignment. (c) In the spatio-temporal enhancement module, we perform matting on pixel embeddings according to predicted masks and then aggregate the spatial and semantic features between frames to enhance temporal consistency. Numbers in colors denote the local matching loss and global matching loss (Best viewed in color).

dimension $\{q_k^{1:T}\}_{1:N_v}$, are used to extract the features of a unique instance, i.e. the k-th instance, throughout the whole input video clip. Meanwhile, the associated instance prototypes $\{p_k^{1:T}\}_{1:N_v}$, i.e. the corresponding outputs from $\mathscr{D}$, represent this unique instance within the whole video. For brevity, we define the associated instance prototypes/features as a **set** of prototypes/features.

Previously, online VIS methods that associate frame-level results without heuristic matching algorithms [13, 41, 10] adopt a *Local Matching and Propagating (LocPro)* technique. This technique matches the predictions and ground truth on a frame level during training. In other words, it conducts one-to-one bipartite matching between ground truth instances and predictions of a frame (what we regard as *Local Matching*), and propagates the matching from previous frames to subsequent frames. Specifically, given a video clip consisting of T consecutive frames, it first computes the pair-wise matching cost $\mathcal{L}_{match}^{t}$ between predictions and ground truth instances on the initial frame (the first frame or the frame objects appear):

$$\mathcal{L}_{match}^{t=1} = \lambda_{cls}\mathcal{L}_{cls}^{t=1} + \lambda_{bce}\mathcal{L}_{bce}^{t=1} + \lambda_{dice}\mathcal{L}_{dice}^{t=1}. \quad (2)$$

The frame-wise $\mathcal{L}_{match}^{t}$ is composed of categorical loss and mask loss. The categorical loss adopts the cross entropy loss $\mathcal{L}_{cls}^{t}$. The mask loss consists of a binary cross entropy loss $\mathcal{L}_{bce}^{t}$ and a dice loss [23]. From the cost matrix, it follows DETR [3] and uses Hungarian algorithm [18] for optimal frame-level matching.

As discussed in Section 1, *Local Matching* only obtains optimal matching on the initial frame, and cannot achieve the global optimal matching of the whole video. As illustrated in Figure 2 (b), both dark and light green prototypes attempt to represent the person. With local matching, the ground truth will be assigned to the light one with a smaller local matching cost, however, the dark one performs better from a global perspective and is supposed to be assigned. Training the model with *LocPro* forces all the predictions of subsequent frames to conform to the initial frame, resulting in temporal inconsistency of the instance features among the video, because inappropriate previous matching brings accumulative error to the model.

We introduce **Global Instance Assignment** (GIA) strategy and expect two functionalities: (1) all frames in the video clip are considered when conducting global matching, and (2) the global assignment encourages the instance features among frames to be optimized for a global optimal objective, both of which serve the temporal consistency of instance features.

During training, different from previous online VIS methods [8, 34] that conduct frame-level local matching and provide supervision signals to each frame according to the matched pairs, we leave out the halfway matching and supervision. Consecutive input frames of a video clip pass through the VIS model in an online manner and the model generates the predictions of all frames $\{\hat{\mathbf{y}}_k\}_{1:N_v} =$

$\{\hat{y}_k^t\}_{1:N_v}^{1:T}$, each of which consists of a category probability $\hat{c}_k^t$ and a segmentation mask probability $\hat{m}_k^t$. To conduct global assignment, we collect predictions of all frames and as well the ground truth video instance segmentation $\{\mathbf{y}_k\}_{1:N_{gt}} = \{y_k^{1:T}\}_{1:N_{gt}}$, including the category label c_k and its binary segmentation masks $\mathbf{m}_k = m_k^{1:T}$. Since the ground truth of an instance only has one category label, we first compute the average predicted category probability across a video clip: $\bar{c}_k = \sum_{t=1}^{T} \hat{c}_k^t / T$ and also collect the masks $\hat{\mathbf{m}}_k = \hat{m}_k^{1:T}$. The global matching cost is defined as:

$$\begin{aligned}\mathcal{L}_{match}^{global} = \lambda_{cls}\mathcal{L}_{cls}(c_k, \bar{c}_{\sigma(k)}) + \lambda_{bce}\mathcal{L}_{bce}(\mathbf{m_k}, \hat{\mathbf{m}}_{\sigma(k)}) \\ + \lambda_{dice}\mathcal{L}_{dice}(\mathbf{m_k}, \hat{\mathbf{m}}_{\sigma(k)}),\end{aligned} \tag{3}$$

where $\sigma \in \mathfrak{S}_{N_v}$ is a permutation of N_v elements. One-to-one bipartite global matching between $\{\hat{\mathbf{y}}_k\}_{1:N_v}$ and $\{\mathbf{y}_k\}_{1:N_{gt}}$ is performed to find the global optimal assignment and the objective can be formally described as:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_{N_v}}{\arg\max} \sum_{k=1}^{N_{gt}} \mathcal{L}_{match}^{global}(\mathbf{y}_k, \hat{\mathbf{y}}_{\sigma(k)}). \tag{4}$$

Following prior work [28, 7, 42], we use Hungarian algorithm [18] to search for the global optimal assignment. In contrast to the prior methods that compute matching loss only for $t = 1$, our method considers the masks and ground truth of the whole clip, computes the matching loss globally, and conducts the global assignment. Finally, given the global optimal assignment, we use the N_{gt} matched video-level predictions to supervise the model with the globally matched instances across the entire video.

With the proposed assignment strategy, GIA, the temporal consistency of instance features across the video clip can be effectively enhanced, since we consider all frames as a whole to search for the optimal objective. Specifically, when the set of instance features across the video represents the target instance well in the first few frames, but fails to track it later, this strategy helps to find a more appropriate set to be optimized. As the global matching cost is lower, the selected features fit the target more closely and are more temporally consistent.

3.3. Spatio-temporal Enhancement

Previous online video instance segmentation methods focus on improving the representative ability of the semantic instance embeddings [38, 34, 15, 13]. The spatial features are not comprehensively investigated to boost the temporal association for online VIS. Thus, we further introduce **Spatio-temporal Enhancement** module (STE), leveraging the spatial information from the previous frame to enhance the temporal consistency of the online model.

Given the t-th frame in the video clip, with the framework described in Section 3.1, the mask embedding of the k-th instance is generated from the instance prototype through an MLP: $\mathcal{M}_k^t = MLP(p_k^t)$, and then the model segments the mask $\hat{m}_k^t$ by pixel-wise dot product between $\mathcal{P}_k^t$ and $\mathcal{M}_k^t$, which can be formulated as: $\hat{m}_{k,i,j}^t = \langle \mathcal{P}_{k,i,j}^t, \mathcal{M}_k^t \rangle$, where i and j denote the spatial position of the pixel.

As illustrated in Figure 2 (c), the spatial information of the frame is encoded in the pixel embeddings. To extract instance-wise spatial features, they can be exploited together with the predicted mask. Specifically, we perform spatial matting on pixel embeddings $\mathcal{P}_k^t$, similar to image matting, to retrieve the instance-wise pixel embedding according to $\hat{m}_k^t$. For each instance, we conduct pixel-wise multiplication between the original pixel embeddings and the binary mask to obtain the retrieved embeddings:

$$\mathcal{R}_k^t = \mathcal{P}_k^t \odot \hat{m}_k^t, \tag{5}$$

where $\odot$ denotes the element-wise multiplication. Many of the retrieved embeddings $\mathcal{R}_k^t$ are redundant since they describe the same instance, and directly mining the spatial information with them is computationally inefficient. Average pooling is adopted to obtain a concise representation of the spatial features for each instance:

$$\mathcal{S}_k^t = \frac{\sum_{i,j} \mathcal{R}_{k,i,j}^t}{\sum_{i,j} \mathbb{1}(\hat{m}_{k,i,j}^t = 1)}, \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The concise spatial features $\mathcal{S}_k^t$ are sent to the next frame for temporal association.

Having received from the previous frame the propagated semantic instance queries, i.e., $q^{t+1} = p^t$, and the spatial features, we follow the standard multi-head cross-attention layer (MHCA) [27] to incorporate the features from two aspects. The instance prototype p^t is used as the query to decode the spatial features:

$$q_k^{t+1} = \mathrm{MHCA}(p_k^t, \mathcal{S}_{1:N_v}^t), \tag{7}$$

where q_k^{t+1} is the updated instance query now. Notably, the positional embedding is shared by the spatial feature and instance query with regard to the same instance, which helps the model align the instance-wise information between frames. Finally, the updated instance query is fed into the *Object Decoder* $\mathscr{D}$ to produce spatio-temporally enhanced features.

By performing the proposed enhancement module on the spatial features between frames, we effectively boost the spatio-temporal association of the features. In this way, the temporal consistency of the features is enhanced via the information from the spatial dimension, and the online video instance segmentation model manages to predict more temporally consistent results.

3.4. Overall Loss

The overall loss for training with a video clip as input is a linear combination of categorical and mask losses using the one-to-one global optimal assignment $\hat{\sigma}$:

$$\mathcal{L}_{overall} = \lambda_{cls}\mathcal{L}_{cls}(c_k, \bar{c}_{\hat{\sigma}(k)}) + \lambda_{bce}\mathcal{L}_{bce}(\mathbf{m_k}, \hat{\mathbf{m}}_{\hat{\sigma}(k)}) + \lambda_{dice}\mathcal{L}_{dice}(\mathbf{m_k}, \hat{\mathbf{m}}_{\hat{\sigma}(k)}), \tag{8}$$

where $\mathcal{L}_{cls}$ is the cross entropy loss, $\mathcal{L}_{bce}$ is the binary cross entropy loss and $\mathcal{L}_{dice}$ is the dice loss [23].

4. Experiments

In this section, we evaluated the proposed TCOVIS on four benchmark datasets. Furthermore, we provided in-depth ablation studies of the effectiveness of TCOVIS. Finally, we presented several visualizations of the predictions from our model.

4.1. Datasets

We evaluated our approach on four VIS datasets: YouTube-VIS 2019 dataset [37], YouTube-VIS 2021 dataset [37], YouTube-VIS 2022 dataset [37] and OVIS dataset [24]. We present a brief description of them:

YouTube-VIS 2019 & 2021 & 2022: YouTube-VIS 2019 [37] is the first VIS dataset and comprises 40 predefined categories of objects. This dataset included 4,883 unique video instances with 131,000 high-quality manual annotations. We followed the widely utilized training/test set split: 2,238 videos were selected for training, 302 videos were adopted for validation and 343 videos were used for testing. Further, YouTube-VIS 2021 improved the 40-category label set and added 4883 more unique video instances. Then, YouTube-VIS 2022 contained 71 additional long evaluation videos on the top of YouTube-VIS 2021.

OVIS: Occluded video instance segmentation (OVIS) is also a challenging VIS dataset. OVIS included 901 videos in total with 25 semantic categories. This dataset contained 5,223 unique video instances and we followed the widely utilized training/test set split: 607 videos were selected for training, 607 videos were used for validation and 154 videos were adopted for testing.

Following [37], the video-level average precision (AP) and average recall (AR) were adopted as the evaluation metrics on both YouTube-VIS and OVIS.

4.2. Implementation Details

We adopted the framework of GenVIS [13] which is built on VITA [14], but left out the similarity loss in VITA and the memory module in GenVIS. With the global assignment, the total loss as well the hyper-parameters were set the same as the video-level loss in VITA, which is a temporally extended loss function [16]. The model was trained with pseudo-videos from COCO images [20] as data augmentation, and with a batch size of 8 video clips of 6 frames. As we froze the backbone and image segmentation model, all experiments were conducted with 8 RTX 2080 Ti GPUs. The method was implemented on detectron2 [36].

4.3. Main Results

Following the standard evaluation metrics [37], we compared TCOVIS with state-of-the-art approaches on four VIS benchmarks: YouTube-VIS 2019/2021/2022 and OVIS.

YouTube-VIS 2019&2021. From Table 1, we can observe that TCOVIS has achieved very competitive performance using both lightweight backbones (ResNet-50) and powerful ones (Swin-L). Moreover, our method can even show better performance than offline methods, such as VITA [14] and offline version GenVIS [13]. On the more difficult YouTube-VIS 2021 dataset, TCOVIS surpasses GenVIS [13] in AP not only with ResNet-50 by 2.4 but also with Swin-L by 1.7.

YouTube-VIS 2022. As shown in Table 2, TCOVIS performed best in both AP and AR on YouTube-VIS 2022 dataset, which is more challenging than 2019&2021 datasets. Especially with powerful backbone Swin-L, TCOVIS outperformed GenVIS [13] in AP with a huge margin of 4.9, which shows the effectiveness of our method for the complex scenarios.

OVIS. Table 3 presents the comparisons on OVIS dataset, and we can find that TCOVIS also achieved the best 46.7 AP and 19.1 AR with Swin-L backbone. The results demonstrate that TCOVIS can deal with complicated situations where objects are heavily occluded in others. With ResNet-50 backbone, the performance is still the second best compared to the previous state-of-the-art methods.

4.4. Ablation Study

In this section, we provided ablation studies and discuss the effects of different settings in the proposed method. The experiments are conducted with a ResNet-50 [12] backbone on YouTube-VIS 2019/2021 [37] valid set.

Effectiveness of the proposed assignment strategy and enhancement module. The ablation studies on the global instance assignment strategy and the spatio-temporal enhancement module are shown in Table 4. As for the assignment strategy, compared to LocPro as the baseline, the model with the proposed global assignment outperformed the baseline model by more than 0.8 on AP, AP_{50} and AP_{75} on YouTube-VIS 2019, and more than 1.2 on YouTube-VIS 2021. The consistently significant improvement indicates that the strategy with global optimal matching contributes to better overall segmentation, while LocPro only considers the local optimal results. Besides, the baseline forces the posterior features to conform to those at the very beginning leading to temporal inconsistency, since the accumulative

	Method	Type	YouTube-VIS 2019					YouTube-VIS 2021				
			AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	EfficientVIS [35]	Offline	37.9	59.7	43.0	40.3	46.6	34.0	57.5	37.3	33.8	42.5
	IFC [16]	Offline	41.2	65.1	44.6	42.3	49.6	35.2	55.9	37.7	32.6	42.9
	Mask2Former-VIS [4]	Offline	46.4	68.0	50.0	-	-	40.6	60.9	41.8	-	-
	TeViT† [39]	Offline	46.6	71.3	51.6	44.9	54.3	37.9	61.2	42.1	35.1	44.6
	SeqFormer [33]	Offline	47.4	69.8	51.8	45.5	54.8	40.5	62.4	43.7	36.1	48.1
	VITA [14]	Offline	49.8	72.6	54.5	49.4	61.0	45.7	67.4	49.5	40.9	53.6
	$GenVIS_{semi\text{-}online}$ [13]	Offline	51.3	72.0	57.8	49.5	60.0	46.3	67.0	50.2	40.6	53.2
	CrossVIS [38]	Online	36.3	56.8	38.9	35.6	40.7	34.2	54.4	37.9	30.4	38.2
	VISOLO [9]	Online	38.6	56.3	43.7	35.7	42.5	36.9	54.7	40.2	30.6	40.9
	MinVIS [15]	Online	47.4	69.0	52.1	45.7	55.7	44.2	66.0	48.1	39.2	51.7
	IDOL [34]	Online	49.5	**74.0**	52.9	47.7	58.7	43.9	68.0	49.6	38.0	50.9
	$GenVIS_{online}$ [13]	Online	50.0	71.5	54.6	49.5	59.7	47.1	67.5	51.5	**41.6**	54.7
	TCOVIS	Online	**52.3**	73.5	**57.6**	**49.8**	**60.2**	**49.5**	**71.2**	**53.8**	41.3	**55.9**
Swin-L	SeqFormer [33]	Offline	59.3	82.1	66.4	51.7	64.4	51.8	74.6	58.2	42.8	58.1
	Mask2Former-VIS [4]	Offline	60.4	84.4	67.0	-	-	52.6	76.4	57.2	-	-
	VITA [14]	Offline	63.0	86.9	67.9	56.3	68.1	57.5	80.6	61.0	47.7	62.6
	$GenVIS_{semi\text{-}online}$ [13]	Offline	63.8	85.7	68.5	56.3	68.4	60.1	80.9	66.5	49.1	64.7
	MinVIS [15]	Online	61.6	83.3	68.6	54.8	66.6	55.3	76.6	62.0	45.9	60.8
	IDOL [34]	Online	**64.3**	**87.5**	**71.0**	55.6	69.1	56.1	80.8	63.5	45.0	60.1
	$GenVIS_{online}$ [13]	Online	64.0	84.9	68.3	**56.1**	**69.4**	59.6	80.9	65.8	**48.7**	65.0
	TCOVIS	Online	64.1	86.6	69.5	55.8	69.0	**61.3**	**82.9**	**68.0**	48.6	**65.1**

Table 1. Quantitative results on **YouTube-VIS 2019 and 2021 validation** sets. The results are respectively grouped by method types (Offline or Online) and backbone networks (ResNet-50 and Swin-L). We **bold** the best performance and underline the second. † denotes using MsgShifT [39] backbone which has a similar weight scale with ResNet-50.

	Method	Type	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	VITA [14]	Offline	32.6	53.9	39.3	30.3	42.6
	GenVIS [13]	Offline	37.2	58.5	42.9	33.2	40.4
	MinVIS [15]	Online	23.3	47.9	19.3	20.2	28.0
	GenVIS [13]	Online	37.5	**61.6**	41.5	32.6	42.2
	TCOVIS	Online	**38.6**	59.4	**41.6**	**32.8**	**46.7**
Swin-L	VITA⋆ [14]	Offline	41.1	63.0	44.0	39.3	44.3
	GenVIS [13]	Offline	44.3	69.9	44.9	39.9	48.4
	MinVIS⋆ [15]	Online	33.1	54.8	33.7	29.5	36.6
	GenVIS [13]	Online	45.1	69.1	47.3	39.8	48.5
	TCOVIS	Online	**51.0**	**73.0**	**53.5**	**41.7**	**56.5**

Table 2. Quantitative results on **YouTube-VIS 2022 validation** dataset. We **bold** the highest accuracy and underline the second. ⋆: Reproduced by [13].

	Method	Type	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
ResNet-50	TeViT† [39]	Offline	17.4	34.9	15.0	11.2	21.8
	VITA [14]	Offline	19.6	41.2	17.4	11.7	26.0
	GenVIS [13]	Offline	34.5	59.4	35.0	16.6	38.3
	CrossVIS [38]	Online	14.9	32.7	12.1	10.3	19.8
	VISOLO [9]	Online	15.3	31.0	13.8	11.1	21.7
	MinVIS [15]	Online	25.0	45.5	24.0	13.9	29.7
	IDOL [34]	Online	30.2	51.3	30.0	15.0	37.5
	GenVIS [13]	Online	**35.8**	**60.8**	36.2	**16.3**	**39.6**
	TCOVIS	Online	35.3	60.7	**36.6**	15.7	39.5
Swin-L	VITA [14]	Offline	27.7	51.9	24.9	14.9	33.0
	GenVIS [13]	Offline	45.4	69.2	47.8	18.9	49.0
	MinVIS [15]	Online	39.4	61.5	41.3	18.1	43.3
	IDOL [34]	Online	42.6	65.7	45.2	17.9	49.6
	GenVIS [13]	Online	45.2	69.1	48.4	**19.1**	48.6
	TCOVIS	Online	**46.7**	**70.9**	**49.5**	**19.1**	**50.8**

Table 3. Quantitative results on **OVIS validation** set. We **bold** the highest accuracy and underline the second. † denotes using MsgShifT [39] backbone.

error impairs the model during the training process. Our proposed assignment strategy encourages the features of all time to fit the global optimal objective, which effectively enhances the temporal consistency.

As for the effectiveness of the spatio-temporal enhancement module, the comparison is also shown in Table 4. The results show that the enhancement module brings improvements of 0.9 AP on YouTube-VIS 2019&2021, compared to using the proposed assignment strategy individually. In particular, the proposed module improved the performance by 1.2 and 1.7 in AP_{50} on two datasets, respectively. The performance improvements demonstrate that the proposed module effectively captures the spatial information from the previous frame and aggregates the semantic and spatial features across time. As a result, the delicate spatio-temporal design further enhances the temporal consistency of fea-

LocPro	GIA	STE	YouTube-VIS 2019					YouTube-VIS 2021				
			AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
✓	-	-	50.6	71.5	56.1	47.6	58.2	47.4	68.1	52.1	39.8	53.1
-	✓	-	51.4	72.3	57.0	49.5	59.9	48.6	69.5	53.6	40.6	55.0
-	✓	✓	52.3	73.5	57.6	49.8	60.2	49.5	71.2	53.8	41.3	55.9

Table 4. Ablation study of the method for Global Instance Assignment strategy (GIA) and the Spatio-temporal Enhancement module (STE) on the YouTube-VIS 2019 / 2021 validation sets. LocPro denotes the Local Matching and Propagating technique.

Architecture	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Cross-Attn.	49.5	71.2	53.8	41.3	55.9
Concat.	47.9	68.6	52.7	41.4	55.1
Self-Cross	48.2	70.0	51.8	41.0	54.1
Dec. Mod.	48.5	69.7	53.9	40.6	54.7
Resp. Pos.	48.8	70.3	52.5	41.5	55.5
Ins. Attn.	49.2	70.6	53.5	40.9	55.1

Table 5. Ablation study of the module manipulations of the Spatio-temporal Enhancement module on YouTube-VIS 2021.

tures for the online video instance segmentation model.

Incorporating the assignment strategy and the enhancement module, our proposed method gained remarkable performance improvements of 1.6 AP on YouTube-VIS 2019 and 2.1 AP on YouTube-VIS 2021 over the LocPro baseline. In a nutshell, the experimental results indicate that the temporally consistent features learned by our proposed TCOVIS significantly boost the performances on online video instance segmentation.

Different manipulations of spatio-temporal enhancement. In Table 5, we compared our proposed spatio-temporal enhancement module with other optional manipulations. *Cross-Attn.* denotes our proposed manipulation following the standard multi-head cross-attention layer to decode the spatial features with shared positional embedding described in Section 3.3. *Concat.* indicates that we concatenated the corresponding spatial feature and prototype of an instance followed by an MLP to get the updated query. *Self-Cross* stands for adding an extra self-attention layer ahead of the cross-attention layer. *Dec. Mod.* is decoder modulation, reviewing the spatial feature for every Transformer decoder layer. *Resp .Pos.* denotes respective positional embeddings were used for the spatial feature and prototype when we performed cross-attention. *Ins. Attn.* indicates the instance-wise decoding in cross-attention layer.

As shown, compared to all its counterparts, our proposed manipulation achieved the best performance. The first three variants are related to the feature aggregation, where *Concat.* is too naive to model the spatial and semantic features, while *Self-Cross* obscures the spatial feature. We inferred the performance decrease of *Dec. Mod.* comes from the information redundancy when aggregating them in every layer. The last two experiments studied the instance-wise correspondence. In the *Resp. Pos.* setting, explicit correspondence for spatial and semantic features of the same instance between frames is absent. *Ins. Attn.* only focuses on the instance itself across time and can slightly enhance the temporal consistency, however, neglecting the spatial features of other instances, the module fails to capture the spatial relationship among instances. The results confirm that the proposed manipulation effectively exploits the spatial information along the temporal dimension to enhance the

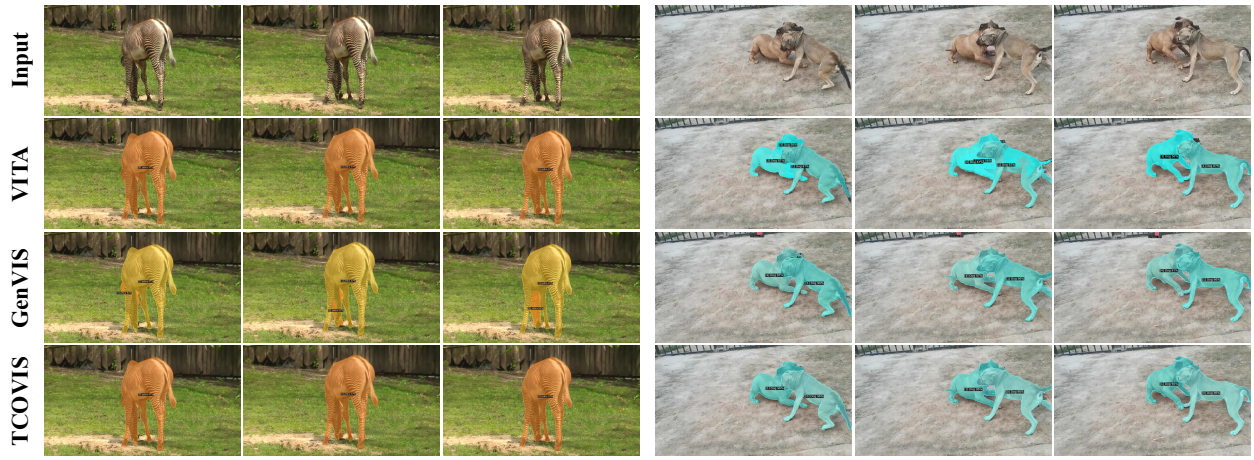

Figure 3. Qualitative results of TCOVIS, compared with VITA [14] and GenVIS [13]. On the left are the predictions on YouTube-VIS 2019 [37] and on the right are on OVIS [24]. Objects displayed in the same color denote the same instance. Our TCOVIS shows more temporally consistent results in these challenging scenes, where there are occlusions of the instance itself or others (Best viewed in color).

Figure 4. Qualitative results of TCOVIS in four challenging cases on YouTube-VIS 2019 [37] and OVIS [24]: (a) fast movement, (b) heavy occlusion, (c) low resolution, and (d) crowded scene. Objects displayed in the same color denote the same instance. The qualitative results demonstrate the effectiveness and robustness of TCOVIS (Best viewed in color).

temporal consistency of features.

4.5. Qualitative results

In Figure 3, we show the qualitative comparisons of the proposed TCOVIS with VITA [14] and GenVIS [13] on YouTube-VIS 2019 and OVIS datasets. In the left scene where there is a zebra with self-occlusion, GenVIS fails to segment its head resulting in fragmented predictions, however, TCOVIS performs temporally consistent segmentation with impressive accuracy. On the right is a difficult case where there are two dogs with similar appearances grappling with each other. VITA incorrectly detects more than two dogs and fails to track the left one, while GenVIS fails to handle the margin of two instances, e.g. the nose and the paw of the left dog. TCOVIS successfully tracks and segments the instances, demonstrating its effectiveness.

In Figure 4, we provide more qualitative results of the proposed method in variously challenging cases, which are all chosen from the mentioned benchmarks [37, 24]. As shown in Figure 4 (a), our method successfully tracks the surfer and the surfboard with fast movement. In the second row, we present a difficult case with severe occlusion, in which our method performs admirably by accurately segmenting the claw (at the bottom) of a parrot despite the presence of a wooden stick that partially obstructs its lower body. Figure 4 (c) illustrates a low-resolution case and our method still achieves good performance. In Figure 4 (d), we depict a crowded scene where TCOVIS is capable of handling multiple instances that share similar appearances and exhibit complex interactions. All the qualitative results in the challenging situations demonstrate the effectiveness and robustness of the proposed method.

5. Conclusion

In this paper, we propose a new online video instance segmentation method, TCOVIS, to fully exploit the temporal information within a video and produce temporally consistent predictions. Based on the query propagation framework, we propose a global instance assignment strategy to perform global optimal matching with the consideration of the entire video and supervise the model with the global optimal objective. We further devise a spatio-temporal enhancement module to capture the spatial feature and aggregate it with the semantic feature between frames. The effectiveness of our method is evaluated with experimental results on YouTube-VIS 2019/2021/2022 and OVIS.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0160102 and in part by the National Natural Science Foundation of China under Grant 62125603.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, pages 158–177, 2020. 2

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, pages 9739–9748, 2020. 1, 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1, 3, 4

[4] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 3, 7

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 1, 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[7] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, pages 6910–6919, 2021. 5

[8] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. In *AAAI*, pages 1361–1369, 2021. 1, 4

[9] Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, and Seon Joo Kim. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *CVPR*, pages 2896–2905, 2022. 3, 7

[10] Fei He, Haoyang Zhang, Naiyu Gao, Jian Jia, Yanhu Shan, Xin Zhao, and Kaiqi Huang. Inspro: Propagating instance query and proposal for online video instance segmentation. In *NeurIPS*, 2022. 1, 2, 3, 4

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1, 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[13] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *CVPR*, pages 14623–14632, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 9

[14] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. VITA: Video instance segmentation via object token association. In *NeurIPS*, 2022. 1, 2, 3, 6, 7, 8, 9

[15] De-An Huang, Zhiding Yu, and Anima Anandkumar. MinVIS: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 1, 2, 3, 5, 7

[16] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *NeurIPS*, 34, 2021. 1, 2, 6, 7

[17] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *NeurIPS*, 34, 2021. 1

[18] Harold W Kuhn. The hungarian method for the assignment problem. *NRL*, 1955. 4, 5

[19] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *ICCV*, pages 1739–1748, 2021. 1, 2

[20] 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*, pages 740–755, 2014. 6

[21] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *CVPR*, pages 9816–9825, 2021. 3

[22] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, pages 8844–8854, 2022. 3

[23] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571, 2016. 4, 6

[24] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *IJCV*, 130(8):2022–2039, 2022. 2, 6, 8, 9

[25] 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*, pages 14454–14463, 2021. 1

[26] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, pages 282–298, 2020. 1, 3

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2, 5

[28] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, pages 5463–5474, 2021. 5

[29] Tao Wang, Ning Xu, Kean Chen, and Weiyao Lin. End-to-end video instance segmentation via spatial-temporal graph neural networks. In *ICCV*, pages 10797–10806, 2021. 3

[30] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, pages 649–665, 2020. 3

[31] 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*, pages 8741–8750, 2021. 1, 2

[32] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *CVPR*, pages 12352–12361, 2021. 1

[33] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, pages 553–569, 2022. 1, 7

[34] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, pages 588–605, 2022. 1, 2, 3, 4, 5, 7

[35] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. In *CVPR*, pages 959–968, 2022. 1, 2, 3, 7

[36] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[37] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, pages 5188–5197, 2019. 1, 2, 3, 6, 8, 9

[38] 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*, pages 8043–8052, 2021. 1, 2, 3, 5, 7

[39] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *CVPR*, pages 2885–2895, 2022. 1, 2, 7

[40] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *ECCV*, pages 659–675. Springer, 2022. 3

[41] Zitong Zhan, Daniel McKee, and Svetlana Lazebnik. Robust online video instance segmentation with track queries. *arXiv preprint arXiv:2211.09108*, 2022. 1, 2, 3, 4

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2, 5