Instance As Identity: A Generic Online Paradigm for Video Instance Segmentation

Feng Zhu[1,2,3⋆], Zongxin Yang[4], Xin Yu[3], Yi Yang[4], and Yunchao Wei[5,6]

[1] Baidu Research
[2] ReLER, Centre for Artificial Intelligence, University of Technology Sydney
[3] Australian Artificial Intelligence Institute, University of Technology Sydney
[4] CCAI, College of Computer Science and Technology, Zhejiang University
[5] Institute of Information Science, Beijing Jiaotong University
[6] Beijing Key Laboratory of Advanced Information Science and Network
Feng.Zhu@student.uts.edu.au

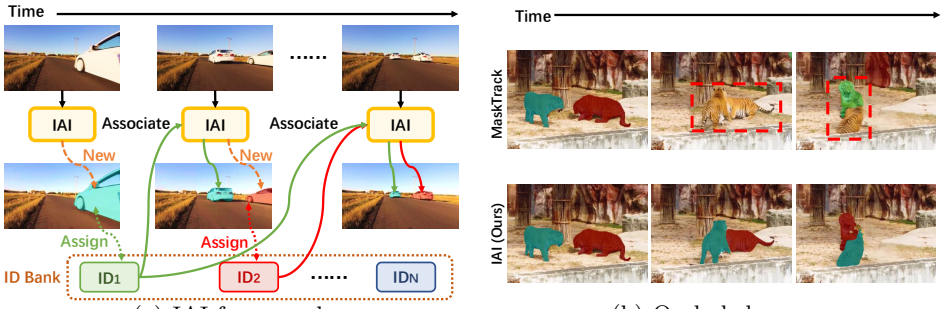
Abstract. Modeling temporal information for both detection and tracking in a unified framework has been proved a promising solution to video instance segmentation (VIS). However, how to effectively incorporate the temporal information into an online model remains an open problem. In this work, we propose a new online VIS paradigm named Instance As Identity (IAI), which models temporal information for both detection and tracking in an efficient way. In detail, IAI employs a novel identification module to predict identification number for tracking instances explicitly. For passing temporal information cross frame, IAI utilizes an association module which combines current features and past embeddings. Notably, IAI can be integrated with different image models. We conduct extensive experiments on three VIS benchmarks. IAI outperforms all the online competitors on YouTube-VIS-2019 (ResNet-101 41.9 mAP) and YouTube-VIS-2021 (ResNet-50 37.7 mAP). Surprisingly, on the more challenging OVIS, IAI achieves SOTA performance (20.3 mAP). Code is available at https://github.com/zfonemore/IAI.

Keywords: Video Instance Segmentation

1 Introduction

Instance Segmentation [27,17,10,13,9,45,6] is an important problem in the computer vision community, which aims to detect and segment objects of specific classes in an image. Benefiting from the rapid growth of deep learning techniques, this problem has achieved great progress in recent years. Most recently, Video Instance Segmentation (VIS) [38], an advanced version of instance segmentation that aims to simultaneously detect, segment, and track different objects of specific categories in videos, is attracting the attention of many researchers due to its wide application prospects, such as augmented reality and video editing.

⋆ Work done during an internship at Baidu.



(a) IAI framework (b) Occluded scene

Fig. 1. (a) The illustration of the IAI paradigm. In the video, the IAI first detects a car in the initial frame and assigns it ID1. In the second frame, IAI associates the first car with ID1, and recognizes the second car as a new instance. IAI assigns ID2 to the new car. In the next frames, IAI associates these two cars using ID1 and ID2. (b) Comparison of MaskTrack and IAI on occluded scenes from OVIS dataset.

The main challenge of VIS lies in how to assemble instance segmentation and tracking into a unified framework. Some latest progresses, e.g., MaskProp [2] and VisTR [36], achieved this target by handling a clip. However, these clip-based approaches fail to perform online inference and thus cannot be applied in real-world applications that requires real-time processing. To address this issue, some online solutions [38,7,22,40] following a tracking-by-detection paradigm are also proposed recently. Although such a paradigm can adopt temporal features for tracking, it does not utilize prior object information (e.g. appearance information and position information) to detect the corresponding ones in following frames. In view of this, we aim to design a novel online pipeline to fully exploit temporal object information and encode it for both detection and tracking processes.

To this end, we propose a new solution named Instance As Identity (IAI), as shown in Fig. 1(a). Within IAI, we first detect instances in the initial frame and assign IDentities (IDs) to them. Then, in the next frame, we directly predict IDs of instances. For those instances that fail to match any previous instances, we assign new unique IDs to them. By conducting this process on each frame, all the instances can be smoothly detected and tracked in an online manner.

To be specific, our IAI is achieved by a novel identification module and an efficient multi-object association module. The identification module consists of a new designed identification head and an identification bank. Particularly, the identification head can dynamically detect new instances and assign IDs, and the identification bank encodes the IDs and masks of objects into ID embeddings for propagating across frames. To construct an efficient multi-object association module, we propose an effective Hybrid Association Block (HAB), which adopts transformer and memory to propagate features for tracking and utilizes a classification projector to encode backbone features. It should be noted that our HAB is very different from the association module proposed in [43], which only

works for the class-agnostic scenarios that the mask of first frame should be correctly provided by human and does not meet the requirements of VIS (i.e., one VIS model should be equipped with the ability of automatically performing instance segmentation and classification by itself). Through these two modules (i.e., identification and multi-object association), our IAI successfully achieves multiple object association at once for both detection and tracking.

To the best of our knowledge, IAI is the first VIS paradigm to use ID to unify detection and tracking in an online way. Besides, our IAI is pretty flexible and could be easily integrated with existing image segmentation models. Surprisingly, as shown in Fig. 1(b), we find that IAI shows a strong capability on handling object occlusion, which is a key problem in VIS. We attribute this robustness to the design of our identification module and association module. First, because our identification module encodes multiple object information into one identification embedding, the enriched surrounding information of each object helps model to separate different instance on occlude scenes. Second, the global memory in our association module helps the model to acquire object information which is absent in long-term occlusion.

We conduct extensive experiments on three challenging VIS benchmarks, i.e., YouTube-VIS-2019, YouTube-VIS-2021 and OVIS, to evaluate the effectiveness and generality of the proposed IAI paradigm. With ResNet-50 [18] as the backbone, the IAI paradigm achieves superior performance on the validation sets of YouTube-VIS-2019 (38.6 mAP) and YouTube-VIS-2021 (37.7 mAP), outperforming all the online model competitors. Particularly, IAI is the first *online* method to achieve an over 40 mAP, i.e., 41.9 mAP, with ResNet-101 as the backbone on YouTube-VIS-2019. Moreover, on the more challenging OVIS dataset, our method outperforms SOTA VIS methods by a large margin (+4.9 mAP), which further proves the robustness of IAI on the occluded scenes.

Overall, we summarize our contributions as follows:

- We propose a generic paradigm for VIS named IAI. IAI achieves superior performance on VIS benchmarks and outperforms all the online methods.
- We propose a novel identification module that can re-identify the previous instance and recognize a new instance, which is the first time in VIS to track instances explicitly using IDs.
- We propose a new hybrid association block as our association module, which combines backbone features with memory ID embeddings.

2 Related Work

VIS is highly related to several tasks, such as image instance segmentation and semi-supervised video object segmentation. In this section, we provide a brief overview of recent studies in VIS and related fields.

Image Instance Segmentation Image instance segmentation algorithms are mainly built on either two-stage frameworks or one-stage frameworks. Though rapid progress has been witnessed in instance segmentation, the classical two-stage architecture Mask R-CNN [17] is still the most popular framework to date.

Many state-of-the-art works are extended on the basis of Mask R-CNN. Mask R-CNN first predicts bounding-box proposals through a regional proposal network and then produces instance masks using the cropped features for proposals. As for one-stage algorithms, CondInst [33] is a good representative, which outperforms many state-of-the-art instance segmentation algorithms. CondInst adopts a dynamic instance-wise mask head to produce instance masks, thus avoiding ROI operations and enabling mask prediction with higher resolution features.

Semi-supervised Video Object Segmentation Semi-supervised video object segmentation (VOS) [30,31] targets at segmenting the given objects with the annotated object masks of the first frame in a video. Many semi-supervised VOS approaches rely on fine-tuning the first frame at test time. Some recent works [12,35,39] propose methods without fine-tuning to achieve a better runtime. STM [29] leverages a memory network to perform long-term propagation. CFBI [42,44] utilizes the feature embedding from the target foreground object and its corresponding background collaboratively. AOT [43] proposes a novel identification mechanism for multi-object association and utilizes a Long Short-Term Transformer to propagate information from memory frames.

Video Instance Segmentation The VIS task consists of classification, segmentation, and tracking of instances in a video. Along with the YouTube-VIS 2019 datasets, [38] proposes a representative algorithm MaskTrack R-CNN. MaskTrack R-CNN employs a tracking branch to the Mask R-CNN framework in order to link the same instance over frames. The VIS task was formally proposed in [38], and most VIS methods follow the tacking-by-detection paradigm of MaskTrack R-CNN. SipMask-VIS [7] adopts a tracking branch similar to the one-stage FCOS [34] and YOLACT [5] framework. CompFeat [16] proposes a temporal attention module and a spatial attention module to extract contextual information in temporal and spatial dimensions. STMask [22] refines spatial features by aligning features between anchors and ground-truth bounding boxes, and designs a temporal fusion module to learn cross-frame information. CrossVIS [40] is built on CondInst [33] and exchanges dynamic filters in two different frames to learn a more robust video-based instance representation. Although these methods are online algorithms, they are not the optimal solution since detection and tracking are conducted in two independent steps.

Different from the tracking-by-detection paradigm, the state-of-the-art algorithm MaskProp [3] designs a mask propagation mechanism to perform detection and tracking simulatenously. MaskProp utilizes deformable convolution [14] and attention to propagate instance features across frames. VisTR [36] and IFC [19] take advantage of the superior sequence modeling of transformers, and extend the transformer-based detection model DETR [8] to solve VIS problem. Despite the promising performance of these methods, they are both offline methods. It remains a challenge to combine detection and tracking in an online paradigm.

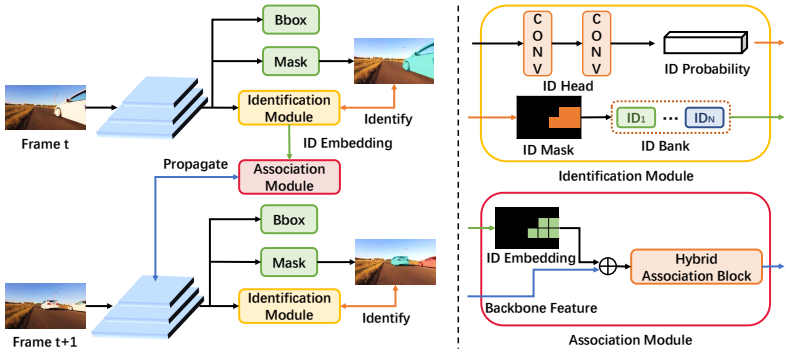


Fig. 2. Overview of the IAI paradigm. In IAI paradigm, we use identification module and association module for tracking and modeling temporal information. The identification consists of an ID head and an ID bank, the former is used to predict ID probability and the latter is used to encode ID mask into ID embedding. The association module is comprised of one hybrid association block and used to propagate information from previous frames to frame t+1.

3 Method

Given an input video $G \in R^{T\times 3\times H\times W}$ comprising of T frames of spatial size $H \times W$, VIS task requires our method to detect, segment and track instances of a predefined category set $\omega = \{1, ..., P\}$ in video G. To be specific, our model predicts an instance mask track $M_G^i \in \{0, 1\}^{T\times H\times W}$ with a class label $c^i \in \{1, ..., P\}$ and a confidence score $s^i \in [0, 1]$ for each detected instance i in G.

In order to solve this challenging problem, we propose a new online paradigm named Instance as Identity (IAI). The detailed framework of IAI is illustrated in Fig. 2. The IAI paradigm designs two modules to extend the original image instance segmentation framework, *i.e.*, identification module and association module. In this section, we first offer a brief introduction to the basic image instance segmentation framework. Then we describe how the ID and association module are designed to combine detection and tracking in an online way.

3.1 Image Instance Segmentation

Commonly, there are two kinds of image instance segmentation frameworks: two-stage framework and one-stage framework. Since our IAI paradigm is not designed on specific image segmentation framework, we take a simplified image segmentation framework for convenience. As it is presented in Fig. 2, the simplified image instance segmentation model contains backbone, bounding box head and mask head. For the image segmentation task, the model firstly uses the backbone to extract features from the image. Then the bounding box head utilizes the object features to classify and regress the bounding box. The mask head utilizes the object features to predict the mask for each instance.

3.2 Identification Module

As shown in Fig. 3, previous tracking-by-detection VIS algorithms always add a tracking branch to the image instance segmentation model to achieve instance association. In this way, temporal information is only utilized for tracking but not for detection. To overcome this disadvantage, MaskProp proposes a mask propagation paradigm to combine detection and tracking. Maskprop processes each instance independently for association across and aggregates all the single-object predictions into a multi-object prediction. However, this post ensemble paradigm is not efficient for multiple object association in video tasks.

Revisit ID Mechanism. As for multi-object learning and understanding, ID Mechanism [43] was recently proposed for associating and re-identifying multiple given objects in video. The ID mechanism consists of an ID embedding and an ID decoding. The ID embedding module utilizes an identity bank and a random permutation matrix to embed the mask of multiple different targets for propagation. The ID decoding module predicts the targets' probabilities using the aggregated feature. Although this ID mechanism provides a good idea for multiple object association in the video, it is impractical to directly apply it to video instance segmentation. There are two main challenges for application in VIS: (1) No targets and ground truth will be given in the first frame, which means nothing to be encoded in the ID embedding; (2) Once a new instance appears in the intermediate frame, the ID decoding module is unable to recognize it and always treats it as background. **Improved ID Mechanism in VIS.** To address these challenges, we propose an improved ID mechanism for VIS. In the improved ID mechanism, we will assign each instance a unique ID for the entire video. First, we will detect new instances and assign them unique IDs. Then we will use a similar ID embedding method to encode the mask of different objects. Finally, we will predict IDs for previously detected instances and recognize new instances in subsequent frames. Through this improved ID mechanism, our IAI paradigm could achieve multiple object association more effectively.

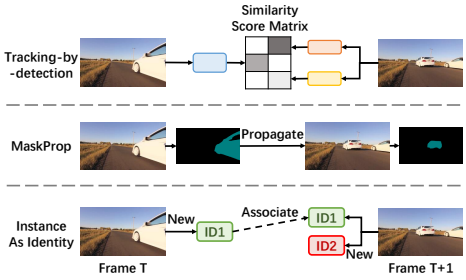


Fig. 3. Tracking patterns of different VIS paradigms.

Here, we take an example to illustrate our improved ID mechanism. In the first frame, we detect a new instances i and assign it a new ID d. Then we encode ID and mask information of instance i into ID embedding and save it to memory. In subsequent frames, the ID probability of detected instances will be predicted. With the predicted IDs, different objects are tracked across frames.

ID Head. In order to recognize new instances and match previous instances through ID, we design a new ID head to predict the ID probability for all object

proposals. As seen in Fig. 2, the ID head is parallel with other heads,*i.e.*, classification head, bounding box head, and mask head, and shares the same features with them, which means no additional cues are required. The ID head predicts ID probability for all the object proposals. As the number of instances could be various in different frames, we set a number N which is large than the maximum amount of objects in a video of the benchmark (e.g. 20 for YouTube-VIS 2019) as the number of IDs in the ID head. Moreover, the ID head predicts a specific N-1-th ID for all the new instances and then assign specific IDs for them. We use IDs from 0 to N-2 to denote the detected instances in previous frames, and the N-th ID means the background class. The ID head does not need elaborate design, and it employs nearly the same structure as the classification head, *e.g.* two convolution layers in Fig. 2.

Post Processing for Inference. As we directly predict the IDs for instances and treat each detection as a unique instance, we use a class-agnostic NMS instead of multiclass NMS. Besides, we use an average of ID score and classification score for NMS. Different from the category prediction, the ID prediction in each frame has to be unique since there could not be two same instances in one frame. The simple ID head is unable to guarantee the uniqueness of ID predictions. Thus we adopt the Hungarian algorithm [20] to assign the unique ID with predicted ID probability as the matching cost. Since there will be various new objects through the video, we set a previously detected object number S to assign ID for new objects. If the object ID is predicted as ID N-1, we will assign it a new ID S+1 and increase S by 1 accordingly. Once S equals $N-1$, we assume there could not be more new instances in the video, and discard the newly detected instances in following frames.

ID Embedding. Assume there are L detected objects in current frame, after the unique ID prediction $U \in \{0, 1, ..., N\}^L$ and mask prediction $M \in \{0, 1\}^{L \times HW}$ are obtained, we produce an ID embedding to propagate these information to following frames. In order to encode ID information and mask information of multiple instances together, we combine U and M to generate the one hot ID masks $Y \in \{0, 1\}^{(N+1) \times HW}$,

$$Y_{U_i} = M_i, \qquad 1 \leq i \leq L. \tag{1}$$

We employ a similar ID embedding method in AOT [43] to encode the ID masks Y. In AOT, an identity bank $D \in R^{(N+1) \times C}$ with C channel dimensions is used to project different instance features into the same feature space. The ID embedding $E \in R^{HW \times C}$ is generated by,

$$E = Y^T D. \tag{2}$$

3.3 Association Module

Revisit Previous VIS Methods. Previous tracking-by-detection methods do not propagate information of one frame to the next frame, and they store features of previous instances for tracking instead. To combine detection and tracking in

a unified model, Maskprop utilizes an attention mechanism to propagate object information. However, this attention mechanism is not efficient since it requires propagating features of every instance in a frame independently, which generates many redundant computations. Moreover, MaskProp separates the video into densely overlapped clips and propagate features from the center frame to all other frames in the clip. Although this propagation manner could avoid information loss during long-term propagation, it takes tremendous computation and memory consumption to perform attention between numerous frames.

New Association Mechanism For VIS. With the compact ID embedding of previous objects, we propose a new association mechanism for VIS. In our new mechanism, we utilize a local memory to store object information of last frame, and maintains a global memory of initial frame to build long-term correlation. Based on these two memory, we use attention operation to get features that contains previous object information and current information. Moreover, we adopt a parallel classification branch to encode backbone features for further classification. This new association mechanism could propagate multiple object information from previous frames to current frame at once, which is much more efficient than MaskProp.

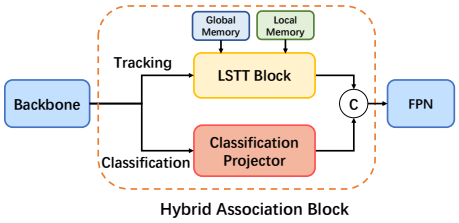


Fig. 4. Illustration of Hybrid Association Block.

Hybrid Association Block (HAB). To implement this new association methcanism, as shown in Fig. 2, we propose an HAB, which is extended on the LSTT block in previous VOS method AOT [43]. As shown in Fig. 4, the new HAB contains a classification projector for additional classification in VIS task. In detail, the LSTT block conducts attention between backbone features and ID embeddings from global and local memory, which learns correlation between frames for object tracking. As for the classification projector, it is 1×1 convolution to encode backbone features for classification. For the output of HAB, we concat outputs from two branches to form the final output.

3.4 Training Details

As for training, we follow the sequential training strategy in [42], in which 5 frames in a video are randomly sampled to form a sequence. For each sequence, we first assign IDs for instances in the sequence since there are no ground truth IDs in the original YouTube-VIS dataset. We assign IDs for instances per frame from 0 to N-1, *e.g.* the first instance assigned 0, the second assigned 1 and so on. One important case should be mentioned is in the frame one instance first appears, the ground truth ID of it should be assigned N.

We train detection and tracking in an end-to-end way, and the loss is

$$L = L_{cls} + L_{bbox} + L_{mask} + L_{id}, \tag{3}$$

where L_{cls}, L_{bbox} and L_{mask} represent the classification loss, bounding box loss and mask loss in image instance segmentation model [17,33]. L_{id} denotes the ID loss, which is implemented with a similar function like L_{cls}. For example, we use focal loss [26] for ID loss when combining with CondInst,

$$L_{id} = -\alpha_t(1 - p_i(d))^{\lambda} log(p_i(d)), \tag{4}$$

where $p_i(d)$ is the probability of assigning ID d to instance i, α_t and λ follow the definition in [26].

4 Experiment

In this section, we conduct extensive experiments to evaluate IAI on three VIS benchmarks, YouTube-VIS-2019 [38], YouTube-VIS-2021 [37] and OVIS [32].

- **YouTube-VIS-2019** is the first large-scale benchmark to video instance segmentation, which consists of 2,883 high-resolution YouTube videos. The dataset is annotated with 4883 unique objects from 40 common categories and contains about 131k instance masks.
- **YouTube-VIS-2021** is extended on the basis of the YouTube-VIS-2019, with more videos and a modified category label set. The YouTube-VIS-2021 contains 3,859 high-resolution YouTube videos, 8,171 unique video instances and approximately 232k high-quality manual annotated masks.
- **OVIS** is a large-scale benchmark for video instance segmentation, which aims to perceive object occlusion in videos. The OVIS dataset consists of 901 videos with severe object occlusions. The dataset is annotated with 5,223 unique instances from 25 commonly seen categories.

We use the average precision (mAP) and average recall (AR) defined in [38] as the evaluation metric. Following previous works, we evaluate all results on validation set through official evaluation servers.

4.1 Implementation Details

Settings. The model is implemented with MMDetection-2.11 [11]. For training, we initialize our model with weights of corresponding instance segmentation model pre-trained on COCO train2017. The instance segmentation models are pretrained with 12 epochs. The VIS network is optimized with AdamW optimizer setting the initial learning rate to 10^{-4}, weight decay to 10^{-4}. The learning rate is reduced by a factor of 10 at 9 and 12 epochs. Specifically, the backbone's learning rate is set to 0.1 of the network's, and weight decay is set to 0.9 of the network's to avoid overfitting. The input size of each frame is resized to 360 $\times$ 640. The number of IDs in ID head is set to 20.

Table 1. Comparisons with SOTA VIS methods on **YouTube-VIS-2019** val set. "✓" under "Aug." means multi-scale augmentation for training, "✓✓" indicates stronger augmentation or additional data. "Temp." means modeling temporal information for detection. We follow [24] to define whether the VIS algorithm is online or offline.

Methods	Backbone	Aug.	Temp.	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}	FPS
offline methods									
STEm-Seg [1]	ResNet-50	✓✓	✓	30.6	50.7	33.5	31.6	37.1	4.4
	ResNet-101			34.6	55.8	37.9	34.4	41.6	2.1
MaskProp [2]	ResNet-50	✓✓	✓	40.0	-	42.9	-	-	<6.2
	ResNet-101			42.5	-	-	45.6	-	<5.6
VisTR [36]	ResNet-50	✓	✓	36.2	59.8	36.9	37.2	42.4	30.0
	ResNet-101			40.1	64.0	45.0	38.3	44.9	27.7
Seq Mask R-CNN [23]	ResNet-50	✓✓	✓	40.4	63.0	43.8	41.1	49.7	-
	ResNet-101			**43.8**	**65.5**	**47.4**	**43.0**	**53.2**	-
IFC [19]	ResNet-50	✓	✓	**41.2**	**65.1**	**44.6**	**42.3**	**49.6**	**107.1**
	ResNet-101			42.6	66.6	46.3	43.5	51.4	89.4
online methods									
MaskTrack R-CNN [38]	ResNet-50			30.3	51.1	32.6	31.0	35.5	32.8
	ResNet-101			31.9	53.7	32.3	32.5	37.7	28.6
SipMask-VIS [7]	ResNet-50	✓		33.7	54.1	35.8	35.4	40.1	34.1
STMask [22]	ResNet-50			33.5	52.1	36.9	31.1	39.2	28.6
	ResNet-101	✓		36.8	56.8	38.0	34.8	41.8	23.4
QueryInst-VIS [15,41]	ResNet-50	✓		36.2	56.7	39.7	36.1	42.9	32.3
CompFeat [16]	ResNet-50	✓✓	✓	35.3	56.0	38.6	33.1	40.3	-
SG-Net [28]	ResNet-50	✓		34.8	56.1	36.8	35.8	40.8	23.0
	ResNet-101			36.3	57.1	39.6	35.9	43.0	19.8
CrossVIS [40]	ResNet-50	✓		36.3	56.8	38.9	35.6	40.7	39.8
	ResNet-101			36.6	57.3	39.7	36.0	42.0	35.6
IAI+CondInst	ResNet-50		✓	37.9	58.8	42.1	38.7	46.8	27.4
	ResNet-50	✓		**38.6**	**60.1**	**41.9**	**38.4**	**45.6**	27.4
	ResNet-101		✓	41.0	61.3	45.3	40.8	47.5	23.7
	ResNet-101	✓		**41.9**	**63.7**	**47.5**	**41.1**	**49.6**	23.7

We randomly sample 5 frames from a video to form a sequence for training. We train our model on VIS datasets with 1× schedule, *i.e.*, 12 epochs. The models are trained on 4 Tesla V100 GPUs with batch size of 16. During inference, the video is processed frame by frame without any test time augmentation. The FPS data of inference is measured on Tesla V100.

4.2 Main Results

YouTube-VIS-2019 Dataset. We apply our IAI paradigm on one-stage segmentation model CondInst, and compare it with state-of-the-art methods in Tab. 1. With simple multi-scale training augmentation, our method achieves 38.6 mAP with ResNet-50 backbone, which outperforms all the online methods in Tab. 1. Moreover, our method even outperforms STEm-seg, STMask

Table 2. Comparisons with SOTA VIS methods on **YouTube-VIS-2021** val set.

Methods	Backbone	Aug.	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN	ResNet-50		28.6	48.9	29.6	26.5	33.8
SipMask-VIS		✓	31.7	52.5	34.0	30.8	37.8
CrossVIS			33.3	53.8	37.0	30.1	37.6
CrossVIS		✓	34.2	54.4	37.9	30.4	38.2
IFC		✓	35.2	57.2	37.5	-	-
IAI+CondInst			**37.7**	**58.0**	**42.3**	**34.6**	**45.6**

Table 3. Comparisons with SOTA VIS methods on **OVIS** val set.

Methods	Backbone	Aug.	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN	ResNet-50		10.8	25.3	8.5	7.9	14.9
SipMask-VIS		✓	10.2	24.7	7.8	7.9	15.8
STEm-Seg		✓	13.8	32.1	11.9	9.1	20.0
QueryInst-VIS		✓	14.7	34.7	11.6	9.0	21.2
STMask			15.4	33.8	12.5	8.9	21.3
CrossVIS		✓	14.9	32.7	12.1	10.3	19.8
CMaskTrack R-CNN		✓	15.4	33.9	13.1	9.3	20.0
IAI+CondInst			18.5	36.8	18.0	11.7	24.0
IAI+CondInst∗			**20.3**	**39.5**	**19.0**	**11.9**	**26.2**

∗ means the model is pretrained on YouTube-VIS 2019 dataset.

and CrossVIS with a stronger Resnet-101 backbone. With the ResNet-101 backbone, our method surpasses the recently proposed online method STMask and CrossVIS by about 5 points in mAP with simple multi-scale augmentation. As for speed, our method achieves a nearly real-time speed at 27.4 FPS. Compared with other online algorithms, we argue that utilizing prior information for detection during inference partly slows down IAI paradigm.

As for the state-of-the-art offline method MaskProp, we argue the high performance of MaskProp partly comes from its combination with multiple strong networks, *e.g.* Spatiotemporal Sampling Network [4], Hybrid Task Cascade mask head [10], High-Resolution Mask Refinement post-process, and complex training augmentations, *e.g.* extra OpenImages [21] datasets and longer training schedule. Meanwhile, MaskProp requires huge computation and memory cost to achieve high performance, which impedes it from online scenarios. We aim to design an efficient online paradigm for VIS, and our method can be integrated with different image segmentation models to solve VIS in an online fashion. Overall, the experimental results prove the effectiveness of the new paradigm.

YouTube-VIS-2021 Dataset. YouTube-VIS-2021 dataset is an upgraded version of YouTube-VIS-2019 dataset, with more videos and an improved class set. We evaluate our method on this new dataset and compare it with some state-of-the-art approaches. As shown in Tab. 2, our algorithm outperforms SipMask-VIS and CrossVIS by a large margin without any training augmentations. The experiment results further demonstrate IAI's advantage over other paradigms.

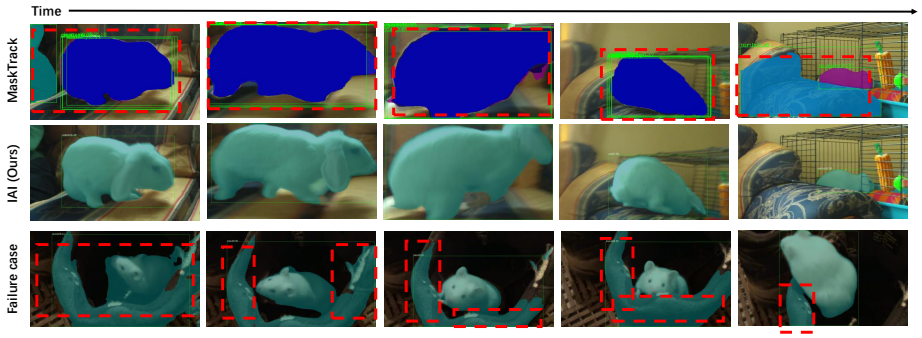


Fig. 5. Qualitative results. (top) Compared with MaskTrack R-CNN, IAI could make better use of temporal information, and performs well even in fast-moving scene. (bottom) Some errors are propagated once mistake happens in previous frames.

OVIS Dataset. To further prove the effectiveness and robustness of our method, we evaluate our method on the OVIS dataset. The OVIS dataset is much harder than YouTube-VIS-2019 and YouTube-VIS-2021 dataset, which contains more instances and more occluded cases per video. As shown in Tab. 3, our methods outperforms SOTA VIS methods by a large margin (+4.9 mAP), which indicates strong ability of our methods on dealing with object occlusion.

Qualitative Results. Fig. 5 visualizes some qualitative results of IAI in comparison with MaskTrack R-CNN. As shown in Fig. 5, IAI segments and tracks object more accurately than MaskTrack R-CNN, especially in fast-moving scenes. IAI makes a better fuse of temporal information, which enables it to handle motion blur and fast-moving target tracking. However, IAI relies on the segmentation quality of the first frame, once a mistake happens in the first frame, IAI might propagate it to the next frames.

4.3 Ablation Study

In this section, we conduct extensive ablation study experiments to prove the general effectiveness of our method. All the experiments are conducted on the YouTube-VIS 2019 dataset. All models are with ResNet-50 FPN [25] as backbone, and trained in 1 $\times$ schedule without any augmentation.

Identification module and association module. We conduct ablation study to prove the effectiveness of two key modules of our method. As shown in Tab. 4, the basic model without ID and association module performs poor, and with both two modules, our IAI achieves superior accuracy. Another important observation is that both two modules are necessary for our IAI paradigm. Without association module, the model could not model previous information and predict accurate IDs. Even worse, the identification module will lead to terrible performance because the ID supervision adds extra noise to model training. Without

Table 4. Experiments of the identification module and association module.

Identification	Association	mAP	AP_{50}	AP_{75}
		24.0	40.4	23.0
✓		12.6	19.5	13.3
	✓	24.5	40.9	25.0
✓	✓	37.9	58.8	42.1

Table 5. Experiments of three key components of the HAB block.

Local	Global	Class	mAP	AP_{50}	AP_{75}
			12.6	19.5	13.3
✓		✓	32.9	50.0	36.7
	✓	✓	34.9	53.1	38.2
✓	✓		36.9	56.7	40.1
✓	✓	✓	37.9	58.8	42.1

identification module, the model could not track the instances and gets similar performance to image model.

HAB block. As the HAB block is the basic component of our association module, we conduct some experiments to verify the effectiveness. We study three key components of the HAB block in Tab. 5: global memory, local memory and classification projector. From the results, we can find that both three components are effective in our IAI paradigm. The global, local memory and classification projector could bring an improvement of 5.0, 3.0 and 1.0 mAP separately.

Image segmentation model. To prove the generality of IAI paradigm, we experiment with both one-stage and two-stage models. In the experiment, we choose CondInst and Mask R-CNN for the representative of one-stage and two-stage models separately. In Tab. 6, we compare our paradigm with the tracking-by-detection paradigm on two image segmentation models. From the results, we could see that IAI paradigm outperforms the tracking-by-detection paradigm on both one-stage and two-stage segmentation models. As for why the IAI on CondInst bring a larger improvement than IAI on MaskTrack, we argue that IAI benefits more from a better image model because more accurate segmentation of first frame (no previous information) can lead to better propagation for next frames.

Table 6. Comparisons with other paradigms on different image instance segmentation frameworks. "Track" means tracking-by-detection paradigm.

Image Model	Paradigm	mAP	AP_{50}	AP_{75}
Mask R-CNN	Track	30.3	51.1	32.6
	IAI	31.7	49.9	34.6
CondInst	Track	32.1	-	-
	IAI	37.9	58.8	42.1

ID loss function. As we introduce a new ID loss in IAI paradigm, we study the effect of the different ID loss functions in Tab. 7. From the results, we could find that focal loss [26] brings a 1.5 mAP improvement over cross-entropy (CE) loss. As the classification loss function is focal loss, this comparison proves that keeping ID loss function consistent with the classification loss function is enough for good performance, which indicates that no additional design is required for the ID loss function.

Table 7. Experiments of different ID loss functions.

L_{id}	mAP	AP_{50}	AP_{75}
CE loss	36.4	53.8	40.9
Focal loss	37.9	58.8	42.1

Table 8. Experiments of different ID head convolution layer numbers.

ID Head	mAP	AP_{50}	AP_{75}
2Conv	37.9	58.8	42.1
3Conv	38.1	61.3	40.9
4Conv	35.3	53.6	37.9

ID head convolution layer number. As the ID head plays an important role in the identification module, and we evaluate the effect of different ID head convolution layer numbers on performance. As shown in Tab. 8, more convolution layers do not bring obvious improvement, and using 4 convolution layers even gets worse accuracy. A possible reason is that ID information is relatively simple compared with appearance information. The appearance information might contain color, shape and other characteristics, while ID information only focuses on similarity between instances. Since ID information is easy to capture, increasing parameters is unable to boost the performance and might cause overfitting.

5 Conclusion

In this paper, we introduce IAI, a novel generic online paradigm for video instance segmentation. The new IAI paradigm successfully utilizes prior object information for both detection and tracking in an online way, and perform multiple object association at once. These advantages make IAI outperform all the online video instance segmentation methods in the challenging YouTube-VIS benchmarks. Notably, the IAI paradigm shows obvious advantages over previous tracking-by-detection paradigm on occluded scenes, outperforming these methods by a large margin on OVIS benchmark. We hope our IAI paradigm could perform as a strong baseline in the VIS and OVIS task, and contribute to future research on video understanding tasks.

Acknowledgment. This work was supported in part by the National NSF of China (No.62120106009), the Fundamental Research Funds for the Central Universities (No. K22RC00010).

References

1. Athar, A., Mahadevan, S., Ošep, A., Leal-Taixé, L., Leibe, B.: Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In: ECCV (2020)
2. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: CVPR (2020)
3. Bertasius, G., Torresani, L.: Classifying, segmenting, and tracking object instances in video with mask propagation. In: CVPR (2020)
4. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: ECCV (2018)
5. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: ICCV (2019)
6. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact++: Better real-time instance segmentation. TPAMI (2020)
7. Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: ECCV (2020)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: Conditional convolutions for instance segmentation. In: ECCV (2020)
9. Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y.: BlendMask: Top-down meets bottom-up for instance segmentation. In: CVPR (2020)
10. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: Hybrid task cascade for instance segmentation. In: CVPR (2019)
11. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
12. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: CVPR (2018)
13. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask r-cnn. In: ECCV (2020)
14. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV (2017)
15. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: ICCV (2021)
16. Fu, Y., Yang, L., Liu, D., Huang, T.S., Shi, H.: Compfeat: Comprehensive feature aggregation for video instance segmentation. In: AAAI (2021)
17. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
19. Hwang, S., Heo, M., Oh, S.W., Kim, S.J.: Video instance segmentation using inter-frame communication transformers. arXiv preprint arXiv:2106.03299 (2021)
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
21. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)

22. Li, M., Li, S., Li, L., Zhang, L.: Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In: CVPR (2021)
23. Lin, H., Wu, R., Liu, S., Lu, J., Jia, J.: Video instance segmentation with a propose-reduce paradigm (2021)
24. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV (2019)
25. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
26. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: ICCV (2017)
27. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
28. Liu, D., Cui, Y., Tan, W., Chen, Y.: Sg-net: Spatial granularity network for one-stage video instance segmentation. In: CVPR (2021)
29. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019)
30. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
31. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
32. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P., Bai, S.: Occluded video instance segmentation. arXiv preprint arXiv:2102.01558 (2021)
33. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: ECCV (2020)
34. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: ICCV (2019)
35. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.C.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: CVPR (2019)
36. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: CVPR (2021)
37. Xu, N., Yang, L., Yang, J., Yue, D., Fan, Y., Liang, Y., Huang., T.S.: Youtube-vis dataset 2021 version. https://youtube-vos.org/dataset/vis (2021)
38. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019)
39. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: CVPR (2018)
40. Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Crossover learning for fast online video instance segmentation. In: ICCV (2021)
41. Yang, S., Fang, Y., Wang, X., Li, Y., Shan, Y., Feng, B., Liu, W.: Tracking instances as queries. arXiv preprint arXiv:2106.11963 (2021)
42. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: ECCV (2020)
43. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. In: NeurIPS (2021)
44. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by multi-scale foreground-background integration. TPAMI (2021)

45. Zhang, R., Tian, Z., Shen, C., You, M., Yan, Y.: Mask encoding for single shot instance segmentation. In: CVPR (2020)