Crossover Learning for Fast Online Video Instance Segmentation

Shusheng Yang[1,2*], Yuxin Fang[1*], Xinggang Wang[1†], Yu Li[2],
Chen Fang[3], Ying Shan[2], Bin Feng[1], Wenyu Liu[1]

[1]School of EIC, Huazhong University of Science & Technology
[2]Applied Research Center (ARC), Tencent PCG [3]Tencent

Abstract

Modeling temporal visual context across frames is critical for video instance segmentation (VIS) and other video understanding tasks. In this paper, we propose a fast online VIS model termed CrossVIS. For temporal information modeling in VIS, we present a novel crossover learning scheme that uses the instance feature in the current frame to pixel-wisely localize the same instance in other frames. Different from previous schemes, crossover learning does not require any additional network parameters for feature enhancement. By integrating with the instance segmentation loss, crossover learning enables efficient cross-frame instance-to-pixel relation learning and brings cost-free improvement during inference. Besides, a global balanced instance embedding branch is proposed for better and more stable online instance association. We conduct extensive experiments on three challenging VIS benchmarks, i.e., YouTube-VIS-2019, OVIS, and YouTube-VIS-2021 to evaluate our methods. CrossVIS achieves state-of-the-art online VIS performance and shows a decent trade-off between latency and accuracy. Code is available at https://github.com/hustvl/CrossVIS.

1. Introduction

Video instance segmentation (VIS) [68] is an emerging task in computer vision that aims to perform per-pixel labeling of instances within video sequences. This task provides a natural understanding of the video scenes. Therefore achieving accurate, robust, and fast video instance segmentation in real-world scenarios will greatly stimulate the development of computer vision applications, *e.g.*, autonomous driving, video surveillance, and video editing.

Recently, significant progress has been witnessed in still-image object detection and instance segmentation. However, extending these methods to VIS remains a challenging work. Similar to other video-based recognition tasks, such as video object segmentation (VOS) [45, 46], video object detection (VOD) [49] and multi-object tracking (MOT) [16, 21, 55, 71], continuous video sequences always bring great challenges, *e.g.*, a huge number of frames required to be fast recognized, heavy occlusion, object disappearing and unconventional object-to-camera poses [18].

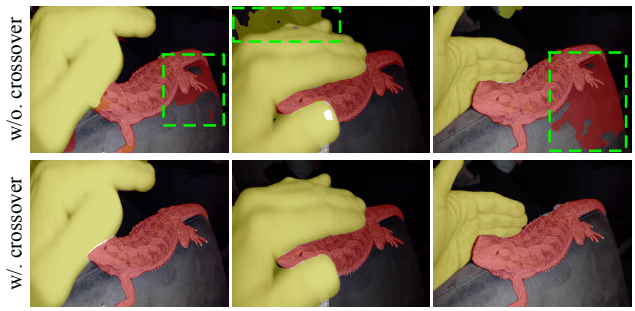

Figure 1. CrossVIS can predict more accurate video instance segmentation results (bottom row) compared with the baseline model without crossover learning (top row).

To conquer these challenges and obtain better performance on these video understanding tasks (VIS, VOS, VOD, and MOT), fully utilizing the temporal information among video frames is critical. Previous deep learning based methods on this topic are in four folds. (1) Pixel-level feature aggregation enhances pixels feature of the current frame using other frames, *e.g.*, STM-VOS [42] and STEm-Seg [1] aggregates pixel-level space-time feature based on Non-local network [57] and 3D convolution, respectively. (2) Instance-level feature aggregation enhances region, proposal or instance features across frames, *e.g.*, MaskProp [2] propagates instance features using deformable convolution [15] for VIS and SELSA [63] fuses instance features using spectral clustering for VOD. (3) Associating instances using metric learning, *e.g.*, MaskTrack R-CNN [68] introduces an association head based on Mask R-CNN [24] and SipMask-VIS [6] adds an adjunctive association head based

*Equal contributions. This work was done while Shusheng Yang was interning at Applied Research Center (ARC), Tencent PCG.

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

on FCOS [53]. (4) Post-processing, *e.g.*, Seq-NMS [23] and ObjLink [44] refine video object detection results based on dynamic programming and learnable object tubelet linking, respectively.

In this paper, we propose a new scheme for temporal information modeling termed crossover learning. The basic idea is to use the instance feature in the current frame to pixel-wisely localize the same instance in other frames. Different from previous pixel/instance-level feature aggregation methods, crossover learning does not require additional network blocks for feature alignment and fusion. It obtains temporal information enhanced features without increasing inference computation cost. Different from metric learning based instance associating methods that require additional metric learning losses, crossover learning is integrated with the instance segmentation loss. Besides, it enables efficient many-to-many relation learning across frames, *i.e.*, the instance pixel features are enforced to be close to the pixels that belong to the same instance and far from pixels that belong to other instances and background. Different from the post-processing methods, crossover learning is end-to-end optimizable with back-propagation.

Since crossover learning is integrated with the instance segmentation loss, it is fully compatible with the other temporal information modeling strategies. In this paper, we further improve the instance association strategy by introducing a global balanced instance embedding learning network branch. Our main contributions are summarized as follows:

- We propose a novel crossover learning scheme that leverages the rich contextual information inherent in videos to strengthen the instance representation across video frames, and weaken the background and instance-irrelevant information in the meantime.
- We introduce a new global balanced instance embedding branch to tackle the association problem in video instance segmentation, which yields better and more stable results than previous pair-wise identity mapping approaches.
- We propose a fully convolutional online video instance segmentation model CrossVIS that achieves strong results on three challenging VIS benchmarks, *i.e.*, YouTube-VIS-2019, OVIS, and YouTube-VIS-2021. To our knowledge, CrossVIS achieves state-of-the-art performance among all online VIS methods and strikes a good speed-accuracy trade-off.

2. Related Work

Still-image Instance Segmentation. Instance segmentation is the task of detecting and segmenting each distinct object of interest in a given image. Many prior works [24, 10, 14, 27, 9, 12, 5, 58, 52, 30, 6, 31] contribute a lot to the rapid developments in this field. Mask R-CNN [24] adapts Faster R-CNN [48] with a parallel mask head to predict instance masks, and leads the two-stage fashion for a long period of time. [27, 9, 13] promote Mask R-CNN and achieve better instance segmentation results. The success of these two-stage models partially is due to the feature alignment operation, *i.e.*, RoIPool [25, 22] and RoIAlign [24]. Recently, instance segmentation methods based on one-stage frameworks without explicit feature alignment operation begin to emerge [5, 4, 8, 65, 58, 59]. As a representative, the fully convolutional CondInst [52] outperforms several state-of-the-art methods on the COCO dataset [36], which dynamically generates filters for mask head conditioned on instances. We build our framework on top of [52] and extend it to the VIS task.

Video Instance Segmentation (VIS). VIS requires classifying, segmenting, and tracking visual instances over all frames in a given video. With the introduction of YouTube-VIS-2019 dataset [68], tremendous progresses [19, 56, 37, 17] have been made in tackling this challenging task. As a representative method, MaskTrack R-CNN [68] extends the two-stage instance segmentation model Mask R-CNN with a pair-wise identity branch to solve the instance association sub-task in VIS. SipMask-VIS [6] follows the similar pipeline based on the one-stage FCOS [53] and YOLACT [5, 4] frameworks. [38] separates all sub-tasks in VIS problem and designs specific networks for each of them, all networks are trained independently and combined during inference to generate the final predictions. MaskProp [2] introduces a novel mask propagation branch on the multi-stage framework [9] that propagates instance masks from one frame to another. As an offline method, MaskProp achieves accurate predictions but suffers from high latency. [32] introduces a modified variational auto-encoder to solve the VIS task. STEm-Seg [1] treats the video clip as 3D spatial-temporal volume and segments objects in a bottom-up fashion. [29] adopts recurrent graph neural networks for VIS task. CompFeat [20] refines features at both frame-level and object-level with temporal and spatial context information. VisTR [60] naturally adopts DETR [7] for VIS task in a query-based end-to-end fashion.

Recently, more challenging benchmarks such as OVIS [47] and YouTube-VIS-2021 [67] are proposed to further promote the advancement of this field. CrossVIS is evaluated on three VIS benchmarks and shows competitive performances. We hope CrossVIS can serve as a strong baseline to facilitate future research.

3. Method

Our goal is to leverage the rich contextual information across different video frames for a more robust instance representation in video instance segmentation (VIS). To this end, we take inspiration from [28, 52, 59] and pro-

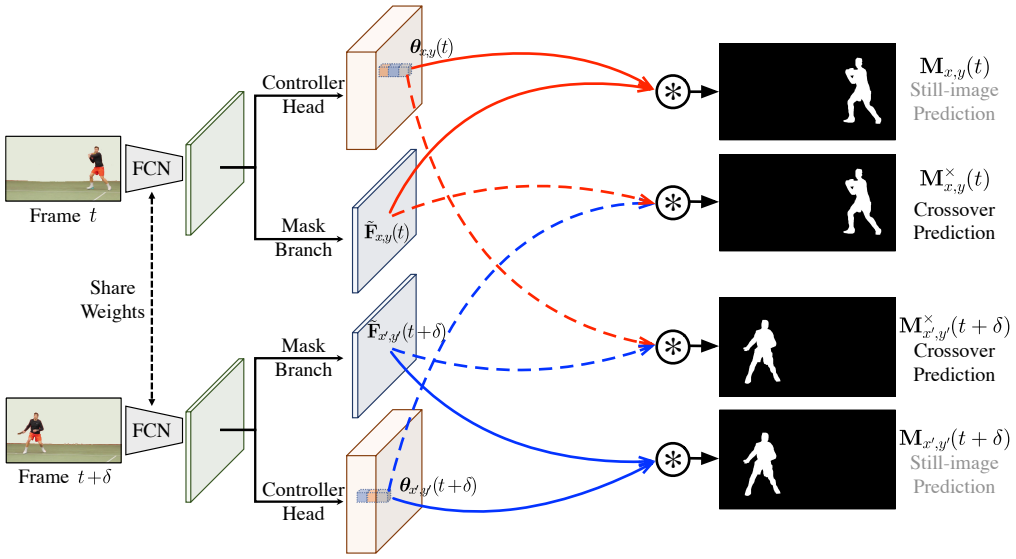

Figure 2. Overview of CrossVIS in the training phase. Two frames at time t and $t+\delta$ are fed into an fully convolutional network (FCN) to generate dynamic filters $\boldsymbol{\theta}_{x,y}(t)$ & $\boldsymbol{\theta}_{x',y'}(t+\delta)$ and mask feature maps $\tilde{\mathbf{F}}_{x,y}(t)$ & $\tilde{\mathbf{F}}_{x',y'}(t+\delta)$. Red lines indicate the dynamic filters and mask feature maps in frame t, blue lines indicate the same in frame $t+\delta$. Solid lines indicate the still-image prediction process, dotted lines indicate the proposed crossover learning scheme. The four "$\circledast$" from top to bottom in the figure correspond to the mask generation process formulated in Eq. (4), Eq. (7), Eq. (6), and Eq. (5), respectively. Classification, localization as well as global balanced instance embedding branches are omitted in the figure for clarification.

pose CrossVIS (see Fig. 2) that consists of two key components tailor-made for VIS task: (1) the crossover learning scheme for more accurate video-based instance representation learning, and (2) the global balanced instance embedding branch for better online instance association.

3.1. Mask Generation for Still-image

For still-image instance segmentation, we leverage the dynamic conditional convolutions [28, 52]. Specifically, our method generates the instance mask $\mathbf{M}_{x,y}$ at location (x, y) by convolving an instance-agnostic feature map $\tilde{\mathbf{F}}_{x,y}$ from the mask branch and a set of instance-specific dynamic filters $\boldsymbol{\theta}_{x,y}$ produced by the controller head. Formally:

$$\tilde{\mathbf{F}}_{x,y} = \text{Concat}\Big(\mathbf{F}_{mask}; \mathbf{O}_{x,y}\Big), \tag{1}$$

$$\mathbf{M}_{x,y} = \text{MaskHead}\Big(\tilde{\mathbf{F}}_{x,y}; \boldsymbol{\theta}_{x,y}\Big), \tag{2}$$

where $\tilde{\mathbf{F}}_{x,y}$ is the combination of mask feature map $\mathbf{F}_{mask}$ and relative coordinates $\mathbf{O}_{x,y}$. $\mathbf{F}_{mask}$ is produced via the mask branch attached on FPN [34] $\{P_3, P_4, P_5\}$ level features. The relative coordinates $\mathbf{O}_{x,y}$ provide a strong localization cue for predicting the instance mask. The MaskHead consists of 3 conv-layers with dynamic filters $\boldsymbol{\theta}_{x,y}$ conditioned on the instance located at (x, y) as convolution kernels. The last layer has 1 output channel and uses sigmoid function for instance mask predictions.

3.2. Crossover Learning

Intuition of Crossover Learning. Still-image instance segmentation needs two types of information [52]: (1) appearance information to categorize objects, which is given by the dynamic filter $\boldsymbol{\theta}_{x,y}$ in our model; and (2) location information to distinguish multiple objects belonging to the same category, which is represented by the relative coordinates $\mathbf{O}_{x,y}$. In the aforementioned still-image instance segmentation model (see Sec. 3.1), for each instance, we have a one-to-one correspondence between the appearance information and location information: given a $\boldsymbol{\theta}_{x,y}$, there exists one and only one $\mathbf{O}_{x,y}$ as the corresponding location information belonging to the same instance. Meanwhile, the connection between different instances is *isolated*.

However, in terms of the VIS task, given a sampled frame-pair from one video, the same instance may appear in different locations of two different sampled frames. Therefore it is possible to use the appearance information from one sampled frame to represent the *same* instance in two *different* sampled frames, guided by *different* location information. We can utilize the appearance information $\boldsymbol{\theta}_{x,y}(t)$ from one sampled frame t to incorporate the location information $\mathbf{O}_{x,y}(t+\delta)$ of the *same instance in another sampled frame* $t+\delta$. By this kind of across frame mapping, we expect the learned instance appearance information can be enhanced and more robust, meanwhile, the background and instance-irrelevant information is weakened.

Formulation of Crossover Learning. Specifically, for a

given video, we denote a detected instance i at time t (or frame t) as:

$$\mathcal{I}_i(t) = (c_i(t), \boldsymbol{\theta}_{x,y}(t), \mathbf{e}_i(t)), \tag{3}$$

where $c_i(t)$ is the instance category, $\boldsymbol{\theta}_{x,y}(t)$ is the dynamic filter for MaskHead, and $\mathbf{e}_i(t)$ is the instance embedding for online association. Without loss of generality, we assume that an instance $\mathcal{I}_i$ exists in both frame t (denoted as $\mathcal{I}_i(t)$) as well as frame $t+\delta$ (denoted as $\mathcal{I}_i(t+\delta)$).

Within each frame, following the setup and notation in Sec. 3.1, at time t, the instance mask of $\mathcal{I}_i(t)$ located at (x, y) can be represented as:

$$\mathbf{M}_{x,y}(t) = \text{MaskHead}\Big(\tilde{\mathbf{F}}_{x,y}(t); \boldsymbol{\theta}_{x,y}(t)\Big). \tag{4}$$

At time $t+\delta$, the instance move from location (x, y) to location (x', y'). So the instance mask of $\mathcal{I}_i(t+\delta)$ can be represented as:

$$\mathbf{M}_{x',y'}(t+\delta) = \text{MaskHead}\Big(\tilde{\mathbf{F}}_{x',y'}(t+\delta); \boldsymbol{\theta}_{x',y'}(t+\delta)\Big), \tag{5}$$

Our crossover learning scheme establishes a connection between the dynamic filter from one frame and the mask feature map from another frame. Specifically, we expect the dynamic filter $\boldsymbol{\theta}_{x,y}(t)$ of $\mathcal{I}_i(t)$ can produce the mask of $\mathcal{I}_i(t+\delta)$ by convolving its mask feature map $\tilde{\mathbf{F}}_{x',y'}(t+\delta)$:

$$\mathbf{M}^{\times}_{x',y'}(t+\delta) = \text{MaskHead}\Big(\tilde{\mathbf{F}}_{x',y'}(t+\delta); \boldsymbol{\theta}_{x,y}(t)\Big), \tag{6}$$

where $\mathbf{M}^{\times}$ with a superscript "$\times$" denotes the instance mask produced by crossover learning. Similarly, we expect the dynamic filter $\boldsymbol{\theta}_{x',y'}(t+\delta)$ of $\mathcal{I}_i(t+\delta)$ can produce the mask of $\mathcal{I}_i(t)$ by convolving its mask feature map $\tilde{\mathbf{F}}_{x,y}(t)$:

$$\mathbf{M}^{\times}_{x,y}(t) = \text{MaskHead}\Big(\tilde{\mathbf{F}}_{x,y}(t); \boldsymbol{\theta}_{x',y'}(t+\delta)\Big). \tag{7}$$

Following [52], during training, the predicted instance masks $\mathbf{M}_{x,y}(t)$, $\mathbf{M}_{x',y'}(t+\delta)$, $\mathbf{M}^{\times}_{x',y'}(t+\delta)$ and $\mathbf{M}^{\times}_{x,y}(t)$ are all optimized by the dice loss [41]:

$$\mathcal{L}_{dice}\left(\mathbf{M}, \mathbf{M}^*\right) = 1 - \frac{2\sum_i^{HW} \mathbf{M}_i \mathbf{M}_i^*}{\sum_i^{HW} \left(\mathbf{M}_i\right)^2 + \sum_i^{HW} \left(\mathbf{M}_i^*\right)^2} \tag{8}$$

where $\mathbf{M}$ is the predicted mask and $\mathbf{M}^*$ is the ground truth mask, i denotes the i^{th} pixel. During inference, the instance mask generation process keeps the same as [52], with no crossover involved.

Advantages of Crossover Learning. For a given instance $\mathcal{I}_i(t)$, its appearance information $\boldsymbol{\theta}_{x,y}(t)$ can learn two kinds of representations: a within-frame one in frame t, and an across-frame one in frame $t+\delta$. At time $t+\delta$, the instance $\mathcal{I}_i(t+\delta)$ may have a different appearance and be in a different context compared with the same instance $\mathcal{I}_i(t)$ at time t. Meanwhile, the background may also changed. The crossover learning enables dynamic filter $\boldsymbol{\theta}_{x,y}(t)$ to identify the same instance representation at both time t and $t+\delta$, regardless of the background and instance-irrelevant information. In this way, we can largely overcome the appearance inconsistency as well as background clutters problems in videos, leveraging the rich contextual information across video frames to get a more accurate and robust instance representation.

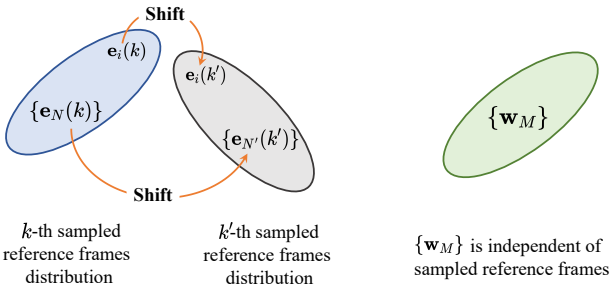

Figure 3. An illustration of pair-wise local embeddings (Fig. 3, left) used in [68, 6], and the proposed instance proxies (Fig. 3, right). For pair-wise local embeddings, $\{\mathbf{e}_N(k)\}$ is the set of all N instance embeddings from k-th sampled reference frames, while in the k'-th sampling, the sampled instance identities will change to $\{\mathbf{e}_{N'}(k')\}$, causing a distribution shift. Even if the same instance $\mathcal{I}_i$ is happened to be sampled in both k-th and k'-th samplings, the corresponding embedding $\{\mathbf{e}_i(k)\}$ may also shift to $\{\mathbf{e}_i(k')\}$ due to occlusion, changing of background and scale variation, *etc*. In contrast, $\{\mathbf{w}_M\}$ is a set of learnable instance-wise weights of the model and independent of sampled reference frames. Therefore $\{\mathbf{w}_M\}$ produces a global, definite convergence status.

3.3. Learning Global Balanced Embeddings for Instance Association

Another crucial sub-task in VIS is the instance association, *i.e*., learning instance embeddings where instances of the same identity are close to each other in the feature space, while instances that belong to different identities are far apart. These embeddings are used for online inference.

In previous tracking-by-detection VIS methods [68, 6], the instance embedding is trained in a *pair-wise* and *local* manner. Specifically, given a key frame at time $t+\delta$ and a reference frame at time t, assuming there is a detected candidate instance $\mathcal{I}_i$ in the key frame as the training sample, and N already identified instances (given by ground truth label during training) in the reference frame as targets. Then, $\mathcal{I}_i$ can only be assigned to one of the N identities if it is one of the already identified instances or a new identity if it is a new instance. The probability of assigning label n

to $\mathcal{I}_i$ is defined as:

$$p_i(n) = \begin{cases} \dfrac{\exp\left(\mathbf{e}_i^\top \mathbf{e}_n\right)}{1 + \sum_{j=1}^{N} \exp\left(\mathbf{e}_i^\top \mathbf{e}_j\right)} & \text{if } n \in [1, N], \\ \dfrac{1}{1 + \sum_{j=1}^{N} \exp\left(\mathbf{e}_i^\top \mathbf{e}_j\right)} & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathbf{e}_i$ and $\mathbf{e}_n$ denote the instance embedding of $\mathcal{I}_i$ from the key frame and $\mathcal{I}_n$ from the reference frame, respectively. $p_i(n)$ is optimized by cross-entropy loss:

$$\mathcal{L}_{CE} = -\log(p_i(n)). \tag{10}$$

However, this approach suffers from the following issues (see Fig 3, left) [50, 43]: the feature space where $\mathbf{e}_i$ and $\{\mathbf{e}_N\} \coloneqq \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N\}$ live in is defined by the sampled frames, and the decision boundary is closely related to the instance embeddings $\{\mathbf{e}_N\}$ from the reference frame. Therefore, the optimization and instance association processes highly depend on stochastic frame sampling, which probably lead to unstable learning and slow convergence. We also observe a relatively large fluctuation in AP when using pair-wise embeddings (see σ_{AP} in Tab. 7).

To remedy these problems and get a globally definite convergence status for instance embeddings, we train our model as a M-class classification problem where M equals to the number of all different identities in the whole training set. We then employ a set of learnable instance-wise weights $\{\mathbf{w}_M\} \coloneqq \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ as proxies of instances (see Fig 3, right) to replace the embeddings of instances $\{\mathbf{e}_N\}$ defined by the sampled frame pair directly [40, 11, 61]. In this way, the probability of assigning label n to $\mathcal{I}_i$ is reformulated as:

$$p_i(n) = \frac{\exp\left(\mathbf{e}_i^\top \mathbf{w}_n\right)}{\sum_{j=1}^{M} \exp\left(\mathbf{e}_i^\top \mathbf{w}_j\right)}. \tag{11}$$

$p_i(n)$ is also optimized by cross-entropy loss:

$$\mathcal{L}_{CE} = -\log(p_i(n)). \tag{12}$$

However, the M-class classification problem is hard to extend to large-scale datasets (*e.g.*, $M = 3,774$ for the YouTube-VIS-2019 training set) as all the negative classes participate in the loss computation, resulting in a large pos-neg samples imbalance issue. Moreover, the large gradient produced by these negative samples from the instance embedding branch dominates the learning process[1], which can negatively affect the optimization of all sub-tasks. To remedy these problems, we adopt focal loss [35] as the objective for our global instance embedding to balance the pos-neg samples as well as the learning of each sub-task:

$$p_i(n) = \begin{cases} \sigma(\mathbf{e}_i^\top \mathbf{w}_n) & \text{if } \mathcal{I}_i = \mathcal{I}_n, \\ 1 - \sigma(\mathbf{e}_i^\top \mathbf{w}_n) & \text{otherwise,} \end{cases} \tag{13}$$

$$\mathcal{L}_{id} = \mathcal{L}_{Focal} = -\alpha_t (1 - p_i(n))^\gamma \log(p_i(n)), \tag{14}$$

where $\sigma(\cdot)$ is the sigmoid function, α_t and γ follow the definition in [35]. $\mathcal{I}_i = \mathcal{I}_n$ means the two instances belong to the same identity. $\mathbf{e}_i$ is generated by the proposed global balanced instance embedding branch which shares a common structure as the classification branches of [52].

3.4. Training and Online Inference

We jointly train detection, segmentation, crossover learning and instance association tasks in an end-to-end manner. The multi-task loss for each sample is:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{seg} + \mathcal{L}_{cross} + \mathcal{L}_{id}. \tag{15}$$

$\mathcal{L}_{det}$ and $\mathcal{L}_{seg}$ denote the object detection loss and still-image instance segmentation loss in [52]. $\mathcal{L}_{cross}$ denotes the crossover learning loss:

$$\begin{aligned} \mathcal{L}_{cross} = {} & \mathcal{L}_{dice}(\mathbf{M}^{\times}_{x,y}(t), \mathbf{M}^{*}_{x,y}(t)) \\ & + \mathcal{L}_{dice}(\mathbf{M}^{\times}_{x',y'}(t+\delta), \mathbf{M}^{*}_{x',y'}(t+\delta)), \end{aligned} \tag{16}$$

where $\mathcal{L}_{dice}$ is formulated in Eq. (8). $\mathcal{L}_{id}$ denotes the instance embedding loss defined in Eq. (13) & Eq. (14).

During inference, the testing video is processed by CrossVIS frame by frame in an online fashion. We follow the inference procedure described in [68, 20].

4. Experiments

4.1. Dataset

We evaluate the proposed CrossVIS on three challenging video instance segmentation benchmarks, *i.e.*, YouTube-VIS-2019 [68], OVIS [47] and YouTube-VIS-2021 [67].

YouTube-VIS-2019 is the first dataset for video instance segmentation, which has a 40-category label set, $4,883$ unique video instances and $131k$ high-quality manual annotations. There are $2,238$ training videos, 302 validation videos, and 343 test videos in it.

OVIS dataset is a recently proposed very challenging VIS dataset with the philosophy of perceiving *object occlusions* in videos, which could reveal the complexity and the diversity of real-world scenes. OVIS consists of $296k$ high-quality instance masks (about $2\times$ of YouTube-VIS-2019) and 5.80 instance per video (about $3.4\times$ of YouTube-VIS-2019) from 25 semantic categories, where object occlusions usually occur. There are 607 training videos, 140 validation videos, and 154 test videos in this dataset.

[1] For the classification sub-task, the large amount of easy negative samples can be handled by focal loss [35]. For regression and segmentation sub-tasks, only positive samples participate in training.

Methods	Backbone	Aug.	Type	FPS	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
IoUTracker+ [68]	ResNet-50		Online	-	23.6	39.2	25.5	26.2	30.9
OSMN [69]	ResNet-50		Online	-	27.5	45.1	29.1	28.6	33.1
DeepSORT [62]	ResNet-50		Online	-	26.1	42.9	26.1	27.8	31.3
FEELVOS [54]	ResNet-50		Offline	-	26.9	42.0	29.7	29.9	33.4
SeqTracker [68]	ResNet-50		Offline	-	27.5	45.7	28.7	29.7	32.5
MaskTrack R-CNN [68]	ResNet-50		Online	32.8	30.3	51.1	32.6	31.0	35.5
MaskProp [2]	ResNet-50	✓✓	Offline	$< 6.2^{\dagger}$	40.0	-	42.9	-	-
SipMask-VIS [6]	ResNet-50		Online	34.1	32.5	53.0	33.3	33.5	38.9
SipMask-VIS [6]	ResNet-50	✓	Online	34.1	33.7	54.1	35.8	35.4	40.1
STEm-Seg [1]	ResNet-50	✓✓	Near Online	4.4	30.6	50.7	33.5	31.6	37.1
Johnander *et al*. [29]	ResNet-50	✓✓	Online	~ 30	35.3	-	-	-	-
CompFeat [20]	ResNet-50	✓✓	Online	< 32.8	35.3	56.0	38.6	33.1	40.3
VisTR [60]	ResNet-50		Offline	30.0	34.4	55.7	36.5	33.5	38.9
CrossVIS	ResNet-50		Online	39.8	34.8	54.6	37.9	34.0	39.0
CrossVIS	ResNet-50	✓	Online	39.8	36.3	56.8	38.9	35.6	40.7
CrossVIS-Lite	DLA-34		Online	48.5	33.0	52.7	35.0	33.9	39.5
CrossVIS-Lite	DLA-34	✓	Online	48.5	36.2	56.7	38.4	35.1	42.0
MaskTrack R-CNN [68]	ResNet-101		Online	28.6	31.9	53.7	32.3	32.5	37.7
MaskProp [2]	ResNet-101	✓✓	Offline	$< 5.6^{\dagger}$	42.5	-	45.6	-	-
STEm-Seg [1]	ResNet-101	✓✓	Near Online	2.1	34.6	55.8	37.9	34.4	41.6
VisTR [60]	ResNet-101		Offline	27.7	35.3	57.0	36.2	34.3	40.4
CrossVIS	ResNet-101		Online	35.6	36.6	57.3	39.7	36.0	42.0

Table 1. Comparisons with some state-of-the-art VIS models on **YouTube-VIS-2019** `val` set. The compared methods are listed roughly in the temporal order. "✓" under "Aug." indicates using multi-scale input frames during training. "✓✓" indicates using stronger data augmentation (*e.g*., random crop, higher resolution input, *etc*.) [2, 29] or additional data [1, 20, 29]. The FPS with superscript "†" is not reported in [2] and is estimated using its utilized components [9, 26, 66, 3]. For the definition of online and offline, we follow [33, 39].

YouTube-VIS-2021 dataset is an improved and augmented version of YouTube-VIS-2019 dataset, which has $8,171$ unique video instances and $232k$ high-quality manual annotations (about $2\times$ of YouTube-VIS-2019). There are $2,985$ training videos, 421 validation videos, and 453 test videos in this dataset.

Unless specified, AP and AR in this paper refer to the average precision and average recall defined in [68]. Following previous works [68, 6, 2, 1], we report our results on the validation set to evaluate the effectiveness of the proposed method.

4.2. Implementation Details

Similar to the setup of [68, 6], we initialize CrossVIS with corresponding CondInst instance segmentation model [52, 26, 70, 34] pre-trained on COCO `train2017` [36] with $1\times$ schedule. Then we train the CrossVIS on VIS datasets with $1\times$ schedule. The pre-train procedure on COCO follows `Detectron2` [64] and `AdelaiDet` [51]. $1\times$ schedule on VIS datasets refers to 12 epoch [68]. The learning rate is set to 0.005 initially following SipMask-VIS [6] and reduced by a factor of 10 at epoch 9 and 11. Most FPS data is measured with a 2080 Ti GPU. For single-scale training, we resize the frame to 360×640. For multi-scale training, we follow the setting in SipMask-VIS. During inference, we resize the frame to 360×640. For our main results, we evaluate the proposed CrossVIS on YouTube-VIS-2019, OVIS and YouTube-VIS-2021 datasets, respectively. Our ablation study is conducted on the YouTube-VIS-2019 dataset using models with ResNet-50-FPN [26, 34] backbone.

4.3. Main Results

Main Results on YouTube-VIS-2019 Dataset. We compare CrossVIS against some state-of-the-art methods in Tab. 1. The comparison is performed in terms of both accuracy and speed. (1) When using the single-scale training strategy, CrossVIS achieves 34.8 AP using ResNet-50 and 36.6 AP using ResNet-101, which is the best among all the online and near online methods in Tab. 1. CrossVIS also outperforms the recently proposed offline method VisTR. (2) When using the multi-scale training strategy, CrossVIS achieves 36.3 AP and 39.8 FPS with ResNet-50, which outperforms SipMask-VIS, STEm-Seg and VisTR with the stronger ResNet-101 backbone. (3) Moreover, CrossVIS achieves the best speed-accuracy trade-off among all VIS approaches in Tab. 1. We also present a more efficient CrossVIS-Lite model with DLA-34 backbone, achieving 36.2 AP and 48.5 FPS, which shows a decent trade-off between latency and accuracy.

MaskProp [2] is a state-of-the-art offline VIS approach that proposes a novel mask propagation mechanism in combination with Spatiotemporal Sampling Network [3], Hy-

Methods	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
SipMask-VIS	10.3	25.4	7.8	7.9	15.8
MaskTrack R-CNN	10.9	26.0	8.1	8.3	15.2
STEm-Seg	13.8	32.1	11.9	9.1	20.0
CrossVIS	14.9	32.7	12.1	10.3	19.8

Table 2. Comparisons with some VIS models on the recently proposed very challenging **OVIS** `val` set. We use ResNet-50 backbone and $1\times$ schedule for all experiments.

Methods	Aug.	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN		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

Table 3. Comparisons with some VIS models on the recently proposed **YouTube-VIS-2021** `val` set. We use ResNet-50 backbone and $1\times$ schedule for all experiments.

brid Task Cascade mask head [9], High-Resolution Mask Refinement post-processing, longer training schedule and stronger data argumentation. MaskProp can achieve very high accuracy but suffer from low inference speed so it is far from real-time applications and online scenarios. Meanwhile, CrossVIS is designed to be an efficient online VIS model and focusing more on the speed-accuracy trade-off. Overall, the experiment results demonstrate the effectiveness of the proposed approach.

Main Results on OVIS Dataset. OVIS is a much more challenging VIS benchmark than YouTube-VIS-2019 and all methods encounter a large performance degradation on this dataset. CrossVIS achieves 14.9 AP, surpassing all methods investigated in [47] under the same experimental conditions. We hope CrossVIS can serve as a strong baseline for this new and challenging benchmark.

Main Results on YouTube-VIS-2021 Dataset. YouTube-VIS-2021 dataset is an improved and augmented version of YouTube-VIS-2019 dataset. We evaluate the recently proposed MaskTrack R-CNN and SipMask-VIS on this dataset using official implementation for comparison. As shown in Tab. 3, CrossVIS surpasses MaskTrack R-CNN and SipMask-VIS by a large margin. We hope CrossVIS can serve as a strong baseline for this new and challenging benchmark.

4.4. Ablation Study

Does Better VIS Results Simply Come from Better Still-image Instance Segmentation Models? The answer is no. We prove this in Tab. 4: (1) Compared with MaskTrack R-CNN using ResNet-101 backbone, CrossVIS is 0.2 $AP^{\texttt{COCO}}_{\texttt{mask}}$ lower, which indicates that our pre-trained model is relatively weaker in terms of still-image instance segmentation on COCO. But for the VIS task, our model is 2.9 $AP^{\texttt{VIS}}$ higher. (2) We implement a VIS baseline called CondInst-VIS which replaces the Mask R-CNN part in MaskTrack R-CNN by CondInst. Therefore the only differences between CrossVIS and CondInst-VIS are the proposed crossover learning scheme and global balanced instance embedding branch. Compared with CondInst-VIS with ResNet-50 and MaskTrack R-CNN with ResNet-101, we conclude that they achieve similar $AP^{\texttt{VIS}}$ under similar $AP^{\texttt{COCO}}_{\texttt{mask}}$. Meanwhile, CrossVIS is 2.7 $AP^{\texttt{VIS}}$ better than CondInst-VIS under the same $AP^{\texttt{COCO}}_{\texttt{mask}}$. The above two observations prove that the improvement in $AP^{\texttt{VIS}}$ mainly comes from the proposed two modules instead of better pre-trained models or baseline.

Method	Backbone	Sched.	$AP^{\texttt{VIS}}$	$AP^{\texttt{COCO}}_{\texttt{mask}}$
MaskTrack R-CNN	ResNet-50	$1\times$	30.3	34.7
MaskTrack R-CNN	ResNet-101		31.9	**35.9**
CondInst-VIS	ResNet-50		32.1	35.7
CrossVIS	ResNet-50		**34.8**	35.7

Table 4. Comparisons between CrossVIS and other baselines in terms of both $AP^{\texttt{VIS}}$ and $AP^{\texttt{COCO}}_{\texttt{mask}}$ on YouTube-VIS-2019 `val` set.

Method	Backbone	FPS (360×640)	$AP^{\texttt{VIS}}$
Mask R-CNN [24]	ResNet-50	41.6	-
CondInst [52]		42.8 (+1.2)	-
MaskTrack R-CNN	ResNet-50	32.8	30.3
CrossVIS		39.8 (+7.0)	34.8

Table 5. Efficiency comparisons on YouTube-VIS-2019 `val` set.

Does the Efficiency of CrossVIS Simply Come from the Efficiency of CondInst? The answer is no. We prove this in Tab. 5. In terms of the inference speed, CondInst is only 1.2 FPS faster than Mask R-CNN in instance segmentation task (similar conclusions are also reported in [52]). Meanwhile, CrossVIS is 7.0 FPS faster than MaskTrack R-CNN in VIS task. This is mainly because: (1) crossover learning adds no extra parameters and can bring cost-free improvement during inference. (2) The global balanced embedding branch adopts a lightweight fully convolutional design compared to the fully connected design in MaskTrack R-CNN. Therefore the efficiency of CrossVIS mainly comes from the efficient design of crossover learning and global balanced embedding.

Crossover Learning. Here we investigate the effectiveness of the proposed crossover learning scheme in Sec. 3.2. During training, we randomly sample frame pairs with a sample time interval $\delta \in [-T, T]$. The results are shown in Tab. 6. We conclude that: (1) When the sample time interval δ is small, *e.g.*, $\delta = [-1, 1]$, the crossover learning brings moderate improvement compared to the baseline. This makes sense because the sampled two frames are quite similar to each other when the δ is small. Under this circumstance, the crossover learning degenerates to naïve

Crossover	$T=1$	$T=3$	$T=5$	$T=10$	$T=15$	$T=20$	$T=\infty$
	33.1	33.4	33.5	33.5	33.6	33.4	33.5
✓	$33.6_{\uparrow(0.5)}$	$34.2_{\uparrow(+0.8)}$	$34.6_{\uparrow(+1.1)}$	$34.6_{\uparrow(+1.1)}$	$34.3_{\uparrow(+0.7)}$	$\mathbf{34.8}_{\uparrow(+1.4)}$	$\mathbf{34.8}_{\uparrow(+1.3)}$

Table 6. Effect of crossover learning and sample time interval on AP. We randomly sample two frames at time t and $t+\delta$ respectively, where the sample time interval $\delta \in [-T, T]$. The "↑" indicates the AP improvement of the model with crossover learning compared to the model without crossover learning under the same T.

Embedding	Loss	AP $\pm$ σ_{AP}	AP_{50}	AP_{75}
Pair-wise	$\mathcal{L}_{CE}$	33.1 ± 0.78	51.9	34.9
Pair-wise	$\mathcal{L}_{Focal}$	33.3 ± 0.72	52.1	35.0
Global	$\mathcal{L}_{CE}$	33.4 ± 0.27	53.9	35.7
Global	$\mathcal{L}_{Focal}$	$\mathbf{34.8 \pm 0.25}$	**54.6**	**37.9**

Table 7. Study of instance association embeddings. To quantitate the fluctuation in results, we conduct 5 independent experiments for each configuration. We report the AP using the *median* of 5 runs. σ_{AP} denotes the *standard deviation* of 5 runs.

Baseline	COL	GBE	AP
✓			32.1
✓	✓		33.1
✓		✓	33.5
✓	✓	✓	34.8

Table 8. Impact of integrating **C**ross**O**ver **L**earning (COL) and **G**lobal **B**alanced **E**mbedding (GBE) into CondInst-VIS baseline.

still-image training. (2) When the sample time interval δ becomes larger, the scene and context become different and diverse between two frames in the sampled frame pair. The baseline without crossover learning cannot explicitly utilize the cross-frame information therefore only has limited improvement. However, crossover learning can benefit significantly from the larger δ and achieves up to 1.4 AP improvement compared to the baseline. (3) The proposed crossover learning scheme is quite insensitive to the variations of T. Overall, models trained with crossover scheme are ~ 1 AP higher than baselines under a wide range of time intervals, *i.e.*, from $T=3$ to $T=\infty$ as shown in Tab. 6.

These results prove the analysis in Sec. 3.2 that the crossover scheme can leverage the rich contextual information across video frames to get a more accurate and robust instance representation.

Instance Association Embeddings. We study the instance association embeddings in Tab. 7. As expected in Sec. 3.3, (1) In terms of AP, the effect from "global" (using learnable $\{\mathbf{w}_M\}$ instead of sampled $\{\mathbf{e}_N\}$) and "balanced" (using $\mathcal{L}_{Focal}$ instead of $\mathcal{L}_{CE}$) are *equally important and interdependent*: Using $\mathcal{L}_{Focal}$ instead of $\mathcal{L}_{CE}$ for pair-wise embedding can only bring 0.2 AP improvement, for large pos-neg imbalance do not exist in the pair-wise scheme. Using global instead of pair-wise embedding optimized by $\mathcal{L}_{CE}$ can only bring 0.3 AP improvement, for there exists a large pos-neg imbalance issue. But together, global and balanced embedding can bring 1.7 AP improvement. So global and balanced are both *indispensable* for good performance. (2) In terms of AP fluctuation, using the global embedding has a much smaller standard deviation σ_{AP} than the pair-wise embedding regardless of the loss function, which indicates that the global embedding can produce a more definite convergence status and more stable results.

Component-wise Analysis. We investigate the effects of crossover learning and global balanced embedding individually and simultaneously in Tab. 8. Using crossover learning and global balanced embedding individually can bring 1.0 AP and 1.4 AP improvement, respectively. In terms of AP, global balanced embedding is slightly higher. Meanwhile, crossover learning adapts CondInst naturally for VIS task during training and brings cost-free improvement during inference. Together, the two components bring 2.7 AP improvement, which is larger than $1.0+1.4$ AP when used solely. Therefore the proposed two components are fully compatible with each other. They show synergy and their improvements are complementary.

5. Conclusion

In this paper, we introduce a novel VIS solution coined as CrossVIS, which performs the best among all online video instance segmentation methods in three challenging VIS benchmarks. Moreover, CrossVIS strikes a decent trade-off between latency and accuracy. We also show that the accuracy and efficiency of CrossVIS are not simply come from the instance segmentation framework but stems from the proposed design. Extensive study proves that crossover learning can bring cost-free improvement during inference, while the lightweight global balanced embedding can help stabilize the model performance. We believe that the proposed approach can serve as a strong baseline for further research on the VIS, and sheds light on other video analysis and video understanding tasks.

Acknowledgement

This work was in part supported by NSFC (No. 61733007, No. 61876212 and No. 61773176) and the Zhejiang Laboratory under Grant 2019NB0AB02.

References

[1] Ali Athar, S. Mahadevan, Aljosa Osep, L. Leal-Taixé, and B. Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020.

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

[3] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *ECCV*, 2018.

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time instance segmentation. *arXiv preprint arXiv:1912.06218*, 2019.

[5] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: real-time instance segmentation. In *ICCV*, 2019.

[6] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *ECCV*, 2020.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[8] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *CVPR*, 2020.

[9] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *CVPR*, 2019.

[10] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, 2018.

[11] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[12] Xinlei Chen, Ross B. Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *ICCV*, 2019.

[13] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *ECCV*, 2020.

[14] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[16] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian D. Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *IJCV*, 2020.

[17] Minghui Dong, Jian Wang, Yuanyuan Huang, Dongdong Yu, Kai Su, Kaihui Zhou, Jie Shao, Shiping Wen, and Changhu Wang. Temporal feature augmented network for video instance segmentation. In *ICCVW*, 2019.

[18] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017.

[19] Qianyu Feng, Zongxin Yang, Peike Li, Yunchao Wei, and Yi Yang. Dual embedding learning for video instance segmentation. In *ICCVW*, 2019.

[20] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. *arXiv preprint arXiv:2012.03400*, 2020.

[21] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

[22] Ross B. Girshick. Fast r-cnn. In *ICCV*, 2015.

[23] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. In *ICCV*, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 2015.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[27] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019.

[28] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NeurIPS*, 2016.

[29] Joakim Johnander, Emil Brissman, Martin Danelljan, and M. Felsberg. Learning video instance segmentation with recurrent graph neural networks. *arXiv preprint arXiv:2012.03911*, 2020.

[30] Youngwan Lee and Jongyoul Park. Centermask : Real-time anchor-free instance segmentation. In *CVPR*, 2020.

[31] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *CVPR*, 2020.

[32] Chung-Ching Lin, Ying Hung, Rogério Feris, and Linglin He. Video instance segmentation tracking with a modified vae architecture. In *CVPR*, 2020.

[33] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

[34] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[35] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[37] Xiaoyu Liu, Haibing Ren, and Tingmeng Ye. Spatio-temporal attention network for video instance segmentation. In *ICCVW*, 2019.

[38] Jonathon Luiten, Philip H. S. Torr, and Bastian Leibe. Video instance segmentation 2019: A winning approach for combined detection, segmentation, classification and tracking. In *ICCVW*, 2019.

[39] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 2020.

[40] Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, 2012.

[41] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

[42] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019.

[43] Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense instance similarity learning. *arXiv preprint arXiv:2006.06664*, 2020.

[44] Tang Peng, Wang Chunyu, Wang Xinggang, Liu Wenyu, Zeng Wenjun, and Wang Jingdong. Object detection in videos by high quality object linking. *TPAMI*, 2020.

[45] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[46] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[47] Jiyang Qi, Yan Gao, Xiaoyu Liu, Yao Hu, Xinggang Wang, Xiang Bai, Philip HS Torr, Serge Belongie, Alan Yuille, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021.

[48] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2015.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. In *IJCV*, 2015.

[50] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020.

[51] Zhi Tian, Hao Chen, Xinlong Wang, Yuliang Liu, and Chunhua Shen. AdelaiDet: A toolbox for instance-level recognition tasks. https://git.io/adelaidet, 2019.

[52] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020.

[53] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019.

[54] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.

[55] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019.

[56] Qiang Wang, Yi He, Xiaoyun Yang, Zhao Yang, and Philip H. S. Torr. An empirical study of detection-based video instance segmentation. In *ICCVW*, 2019.

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[58] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *ECCV*, 2020.

[59] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *NeurIPS*, 2020.

[60] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *arXiv preprint arXiv:2011.14503*, 2020.

[61] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, 2020.

[62] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.

[63] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9217–9225, 2019.

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

[65] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020.

[66] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[67] Ning Xu, Linjie Yang, Jianchao Yang, Dingcheng Yue, Yuchen Fan, Yuchen Liang, and Thomas S. Huang. Youtube-vis dataset 2021 version. https://youtube-vos.org/dataset/vis, 2021.

[68] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

[69] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K. Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, 2018.

[70] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, 2018.

[71] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, 2020.