What is Point Supervision Worth in Video Instance Segmentation?

Shuaiyi Huang[1*], De-An Huang[2], Zhiding Yu[2], Shiyi Lan[2], Subhashree Radhakrishnan[2],
Jose M. Alvarez[2], Abhinav Shrivastava[1], Anima Anandkumar[3*]
[1]University of Maryland, College Park [2]NVIDIA [3]Caltech

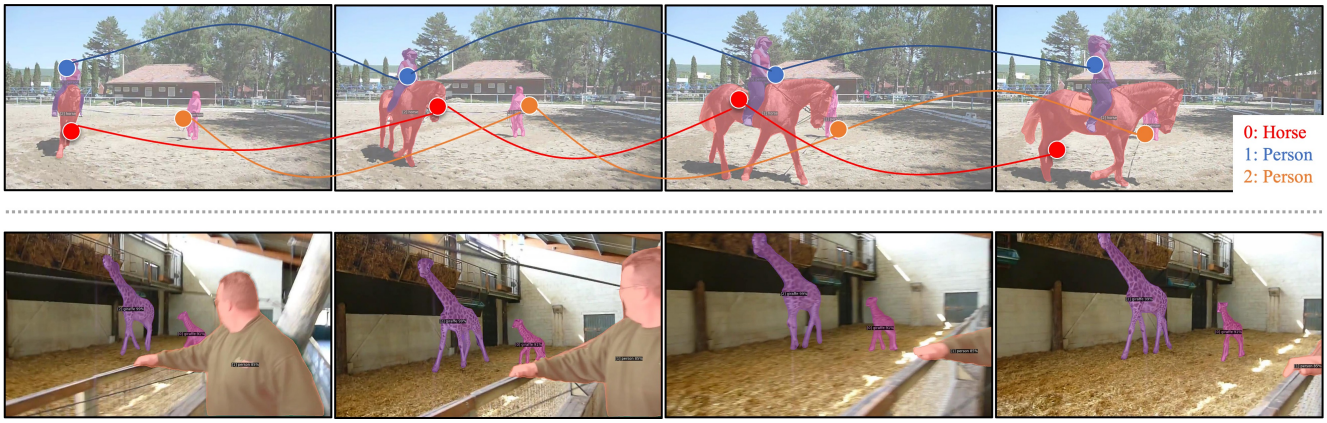

Figure 1. **Point-supervised video instance segmentation in this work (YoutubeVIS-2021)**. Top: point-level annotations in the training set (pseudo masks generated from our method overlaid); Bottom: mask predictions in the validation set.

Abstract

Video instance segmentation (VIS) is a challenging vision task that aims to detect, segment, and track objects in videos. Conventional VIS methods rely on densely-annotated object masks which are expensive. We reduce the human annotations to only one point for each object in a video frame during training, and obtain high-quality mask predictions close to fully supervised models. Our proposed training method consists of a class-agnostic proposal generation module to provide rich negative samples and a spatio-temporal point-based matcher to match the object queries with the provided point annotations. Comprehensive experiments on three VIS benchmarks demonstrate competitive performance of the proposed framework, nearly matching fully supervised methods.

1. Introduction

Video instance segmentation (VIS) is emerging as a challenging vision task which aims to detect, segment, track objects in continuous videos [61]. It has achieved increasing attention recently [9, 18, 21, 35, 39, 45, 67] due to its wide real-world applications such as video editing, 3D reconstruction [20, 31, 69], and view point estimation [30, 52].

Annotating per-frame object masks in videos is time-consuming and even more challenging than annotating image instance segmentation masks. Due to the limited video annotations, a common strategy to train a video instance segmentation model is to first train on image instance segmentation datasets with ground truth mask and category annotations (*e.g.* COCO [40]), and then finetune on video instance segmentation datasets with ground truth masks, category and tracking annotations [13, 29]. However, the categories in image datasets do not necessarily fully overlap with the video datasets, and hence adapting models from the image domain to the video domain has challenging generalization issues due to the emergence of new categories.

There are some recent efforts to reduce video annotation cost for video instance segmentation. They propose to learn with sub-sampled video frames [26], category annotations in videos [42], or without any video annotations [19]. However, these approaches either still require dense masks in the sub-sampled frames [26] or are barely competitive compared with supervised approaches [42] or can only handle the overlapped category between video and image dataset [19]. These limitations of existing approaches show that it is still unclear what is the optimal way to reduce annotation cost for video instance segmentation.

In this paper, we ask the question: *To what level can we*

*Work done during internship / affiliation with NVIDIA.

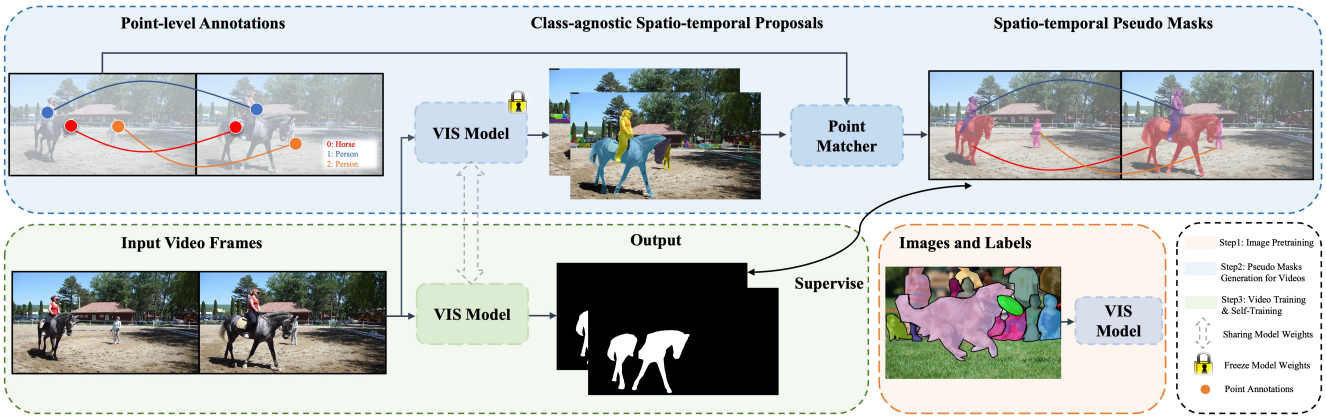

Figure 2. **Method Overview.** Our method consists of class-agnostic spatial-temporal proposal generation, a spatio-temporal point-based matcher to match object queires with point annotations for high-quality pseudo-label generation, and self-training to mitigate the domain gap between images and videos. See text for details.

reduce human annotations in videos and still train an accurate model for video instance segmentation? We believe that point supervision presents a "sweet spot" for annotating objects in videos. Point annotations are significantly cheaper than other alternatives, such as bounding boxes, as one simple click only costs around 1 second [4, 14]. In the most extreme case we considered, every object instance in a frame only contains one labeled point, as shown in Figure 1. Despite the many benefits of point supervision, using it to supervise dense predictions is challenging and raises many ill-conditioned issues, such as the sparsity of ground truth and the lack of informative negative samples.

In this work, we introduce a point-supervised VIS framework (PointVIS) to address these challenges. PointVIS leverages the knowledge from image-based (*e.g.* COCO) pre-raining and guides the VIS task in an open set manner.

Our main contributions are:

- PointVIS is the first attempt to comprehensively investigate video instance segmentation with point-level supervision. Our work significantly reduces the amount of required annotations in VIS and opens up the possibility to address the task with minimal supervision.
- PointVIS overcomes the challenges in point-supervised video instance segmentation, using the proposed class-agnostic proposal generation and a point-based matcher.
- PointVIS is simple to implement, achieving competitive results compared with fully-supervised methods on three major VIS benchmarks.
- We further conduct comprehensive studies in different settings of points with important observations, providing a deeper understanding on what kind of point supervision matters in the VIS task.

The key challenge of using point supervision is the sparsity of ground truth, which further leads to the lack of informative negative samples that are along the boundaries of dense mask annotations. We address these challenges by proposing to (1) generate class-agnostic spatio-temporal instance proposals without video mask annotation and (2) match these proposals with our point annotations to obtain dense pseudo-label for training the VIS model. We leverage COCO pre-trained image instance segmentation model to generate per-frame instance proposals and use bipartite matching on query embeddings to convert them to spatio-temporal video instance proposals. We further design a loss function to match these proposals with our sparse point annotations as there could be multiple proposals that overlap with a single point annotation. We show that both of these designs are crucial to generating high-quality pseudo mask annotations for learning with only point supervision. An overview of our approach is shown in Figure 2.

We further conduct comprehensive studies on how point annotations affect VIS and make the following important observations: (1) even one positive point annotated per video object already achieves good performance, retaining 87% of the performance of fully-supervised methods on Youtube-VIS 2019 [61]; (2) given positive points, increasing negative points improves performance, while adding positive points alone could provide little gain; (3) the positions of positive points have limited effect on performance while the positions of negative points matter more. These observations shed lights on what kind of point supervision matters for video instance segmentation, making it a step closer towards more realistic open-world applications.

2. Related Work

2.1. Image instance segmentation

Supervised Instance Segmentation. Instance segmentation requires bounding-box regression, classification, and pixel-level segmentation of all objects present in images.

After the success of two-stage instance segmentation methods [22, 25, 36, 46, 47], one-stage instance segmentation methods [7, 15, 34, 50, 60, 66] not only significantly improve the accuracy but also reduce the computation cost. Recent video instance segmentation approaches [13, 26, 35, 41, 57, 62] are built on those one-stage methods.

Weakly supervised instance segmentation. Due to the heavy segmentation annotation costs, weakly supervised instance segmentation is a potential way to reduce this cost. Zhou et al. [68] and Ahn [1] propose learning instance segmentation with image-level annotations. Another collection of instance segmentation approaches leverage box-level supervision [2, 24, 32, 51]. Recently, image instance segmentation with both box and point-level supervision shows competitive results [14]. Note that our point-supervised video instance segmentation differs from [14] as we do not use any additional bounding box annotations.

2.2. Video Instance Segmentation

Supervised video instance segmentation. Video Instance Segmentation is a joint task of detection, instance segmentation, and tracking, which was first proposed by Yang et al. [61]. Most previous approaches [18, 35, 45, 62, 67] follow the tracking-by-detection paradigm, which segments and classifies objects and then associates objects across frames. Another trend of video instance segmentation [3, 6, 39] follows the clip-match paradigm, where the video is separated into multiple overlapped clips, and objects are detected and segmented in each clip and then associated across different clips. Tracking-by-regression approaches [41, 57] have also been proposed to generate detections and associate object bounding boxes of the same objects across contiguous frames. Recently, transformer-based approaches for VIS have attracted much attention [13, 29, 56, 63]. Our work is built on top of MinVIS [26] for its excellent performance in VIS using image-based training. We verify that this design also benefits our task and provides better inductive bias in our weakly-supervised learning setting.

Weakly/Semi-supervised video segmentation. As annotations of VIS are expensive, there are emerging methods that aim to learn VIS with reduced annotations [19, 26, 42]. Liu et al. [42] propose to learn VIS by using image-level annotations with correspondences [1, 27, 28]. Fu *et al.* [19] propose to leverage instance segmentation annotations of COCO dataset and learn VIS without video annotations. However, these methods either still require dense masks in the sub-sampled frames [26] or are barely competitive compared with supervised approaches [42] or can only handle the overlapped category between video and image dataset [19]. In contrast, our PointVIS can cover and handle all categories with largely reduced annotation cost. In addition, we for the first time show that video instance segmentation with one point per object can achieve decent performance compared with the fully-supervised counterparts.

2.3. Point-supervised methods.

Recently point-level supervision has attracted growing attention in computer vision, including object localization [65], object detection [11, 53], image instance segmentation [14, 33], and image panoptic segmentation [17, 37]. Point-level interactions are also popular in the field of interactive image segmentation [12, 16, 43] and interactive video segmentation [5, 8, 23, 48], with a focus on label propagation or reducing interaction time. The closest work to ours is PointPanoptic [17] where they use a single point to train image panoptic segmentation. Note that this fundamentally differs from our work, as we aim for reducing video annotations by utilizing pretrained image representations while PointPanoptic [17] aims for reducing image annotations by training from scratch. To our best knowledge, there is no prior work that utilizes point-level supervision for VIS.

3. Method

Our PointVIS design is motivated by three major challenges in point-supervised VIS:

Sparsity in ground truth. Unlike masks and boxes, points do not provide detailed localization and shape information about objects, especially their boundaries, sizes and extreme points. Points are also sparse, which causes difficulties during model training. Our task is further complicated by the lack of negative supervision when only positive points are annotated, which is important to learn a correct decision boundary. Since many VIS methods involve COCO pretraining, we similarly leverage this pipeline to enrich our supervision using object shape priors.

Matching sparse annotations. An important step in recent Transformer-based instance segmentation models is to match their mask proposals with ground truth masks using some costs, usually defined as the intersection-over-union (IoU) between two masks. This matching step is the key to enable end-to-end training without non-maximum suppression (NMS). However, such matching process becomes problematic in our case, since points do not provide informative measurement on how accurate a mask is. Therefore, special designs need to be taken into consideration when defining the cost of matching step in PointVIS.

Novel categories. In real world applications, there is no guarantee that the categories of downstream VIS tasks overlap with those in images. Therefore, a point-supervised framework has to handle arbitrary new categories and learn efficiently with sparse supervision.

Our solution. We therefore propose several designs in PointVIS to address the above challenges. We first pretrain instance segmentation models on COCO and use the pretrained models to generate spatiao-temporal mask proposals in training videoes. Despite the issue mentioned

above, this class-agnostic proposal generation mechanism works well in open set scenarios, providing good coverage on categories not seen in COCO. We then propose a point-based matcher that incorporates annotation-free negative cues from other instances in the same video frame. Finally, we address the generalization issue for new categories in videos by conducting self-training to mitigate the domain gap and refine our results. We iterate the training with pseudo masks from prior round. These solutions together allow us to learn VIS with points effectively.

3.1. Problem Setup

Due to the high annotation cost in videos, in the standard supervised setup, a prevalent strategy to train a video instance segmentation model is to first train on image instance segmentation datasets, and then finetune on video instance segmentation datasets. Formally, there is a full-labeled image instance segmentation dataset $\mathcal{D}_I$ with category space size $\mathbf{O}_I$ and a full-labeled video instance segmentation dataset $\mathcal{D}_V$ with category space size $\mathbf{O}_V$, both of which are annotated with object masks.

In our proposed point-supervised setup, we adopt a video dataset with point annotations (denoted as $\mathcal{D}_{V_p}$) instead of with masks for each video object as in $\mathcal{D}_V$. Specifically, given a video $\mathbf{V} \in \mathbb{R}^{H\times W\times T}$ of T RGB frames with width W and height H, the point annotations for the j^{th} video object in the video $\mathbf{V}$ is denoted as $\mathbf{G}_j = \{\{\mathbf{P}_j^t, \mathbf{L}_j^t\}_{t=1}^T, \mathbf{b}_j\}$, where $\mathbf{P}_j^t \in \mathbb{R}^{N_j^t\times 2}$ are the x-y coordinates of the annotated points at the t^{th} frame, $\mathbf{L}_j^t \in \mathbb{R}^{N_j^t}$ are the corresponding binary labels for the annotated points indicating foreground or background, N_j^t is the number of annotated points, $\mathbf{b}_j \in \mathbb{R}^{\mathbf{O}_V}$ is the one-hot category label and t is the time index. In this setup, a video instance segmentation model $\mathbf{F}$ is first trained on image dataset $\mathcal{D}_I$ with full masks, resulting an image model $\mathbf{F}(;\theta_I)$. Then $\mathbf{F}(;\theta_I)$ is finetuned on video dataset $\mathcal{D}_{V_p}$ with point-level annotations, resulting the final video instance segmentation model $\mathbf{F}(;\theta_{V_p})$.

Supervising a model solely based on annotated points with a loss function like cross-entropy can lead to the model collapsing, especially if only positive points are annotated. Therefore, densifying point annotations and including negative supervision signals become crucial for learning a structured dense prediction task via points.

3.2. Learning VIS with Sparse Points

Class-agnostic proposal generation. To obtain meaningful dense supervision with abundant negative signals, we shift one step back to the image model instead of dwelling on a degenerated video model finetuned with points. Recall that we have image instance segmentation datasets with mask annotations ready at hands. A pretrained image model on such datasets should already know rough object shape, even if never trained on videos and the categories do not fully overlap. Given this simple yet non-trivial discovery, we propose to generate dense class-agnostic spatial-temporal proposals for each video by utilizing a pretrained image model that encodes rich shape prior.

It is challenging to obtain spatio-temporal proposals when video supervision is unavailable. We address this challenge by leveraging COCO pre-trained image instance segmentation model [15] to generate per-frame instance proposals and use bipartite matching on query embeddings to convert them to spatio-temporal video instance proposals. While previous work has used a similar technique for VIS with full video supervision [26], we here show that this enables cross-domain video instance proposals generation trained with image dataset only.

Concretely, given model $\mathbf{F}(;\theta_I)$ trained on image dataset $\mathcal{D}_I$ and a video sequence $\mathbf{V}$ from $\mathcal{D}_{V_p}$, we obtain the initial proposals $\hat{\mathbf{R}}$ for $\mathbf{V}$ by conducting inference as below:

$$\hat{\mathbf{R}} = \mathbf{F}(\mathbf{V};\theta_I) = \{\hat{\mathbf{M}}_r, \hat{c}_r\}_{r=1}^R \tag{1}$$

, where $\hat{\mathbf{M}} \in \mathbb{R}^{H\times W\times T}$ is a spatial-temporal proposal with continous logits after sigmoid but before binarization, $\hat{c}_r$ is the confidence score, R is the maximum number of proposals for a video (e.g. 100).

Given the above dense proposals, there is still one open-set problem worth resolving. As there could be new categories in $\mathcal{D}_{V_p}$ and $\mathbf{F}(;\theta_I)$ has never been finetuned on $\mathcal{D}_{V_p}$, the confidence score $\hat{c}_r$ is not meaningful for every video. To represent the confidence of class-agnostic proposals without the reliance on categories, we propose to use maskness score following [55] to obtain the confidence of a mask as $c_r = \frac{1}{H\times W\times T}\sum_{x,y,z}\hat{\mathbf{M}}_r(x,y,z)$ where x,y are the x-axis and y-axis spatial coordinates, t indexes the time.

Therefore, our final class-agnostic dense spatio-temporal proposals $\mathbf{R}$ for a video $\mathbf{V}$ is denoted as $\mathbf{R} = \{\mathbf{M}_r, c_r\}_{r=1}^R$, where $\mathbf{M}_r \in \mathbb{R}^{H\times W\times T}$ is the binarized mask of $\hat{\mathbf{M}}_r$, c_r is the maskness score.

Point-based matcher. Given the dense class-agnostic proposals, the next challenge is to assign the best proposal to a video object and produce the final pseudo mask when only point annotations are available in videos. This can be challenging since there may be multiple proposals overlapping with a single point, making it difficult to determine which proposal provides the best boundary information.

To address this issue, we develop a matching cost function that combines cues from both point annotations and spatio-temporal proposals to effectively match proposals and video objects with points. With our proposed point-based matcher, we can formulate the pseudo-label filtering problem as a bipartite matching problem between the proposals and the objects. We provide details on the matching process and the design of the matching cost below.

Figure 3. Visualization of point annotations and pseudo masks obtained by our method on Youtube-VIS 2019 [61] (row1), Youtube-VIS 2021 [61] (row2), and OVIS [49] (row3) training set.

Specifically, we search for a permutation $\hat{\sigma}$ between the set of dense proposals and the set of video ground truth with points given a video. Assuming R is larger than the number of objects in the video, we consider $\mathbf{G}$ as a set of video ground truth with size R padded with $\varnothing$ (no object). To find a bipartite matching between $\mathbf{R}$ and $\mathbf{G}$, we search for a permutation $\sigma \in \Omega_R$ of R elements with the lowest cost:

$$\hat{\sigma} = \arg\min_{\sigma \in \Omega_R} \sum_{j=1}^{R} \mathcal{L}_{\text{match}}(\mathbf{G}_j, \mathbf{R}_{\sigma(j)}) \tag{2}$$

where $\mathcal{L}_{\text{match}}(\mathbf{G}_j, \mathbf{R}_{\sigma(j)})$ is a pair-wise matching cost between ground truth $\mathbf{G}_j$ and a proposal with index $\sigma(j)$. This optimal assignment is computed efficiently with the Hungarian algorithm following prior work [10]. Next, we present the specific matching cost $\mathcal{L}_{\text{match}}$ given point annotations.

To penalize proposals that do not overlap with the annotated points consistently, we first define an annotated cost $\mathcal{L}_{\text{ann}}(\mathbf{G}_j, \mathbf{R}_{\sigma(j)})$ calculated over T frames as below:

$$\mathcal{L}_{\text{ann}}(\mathbf{G}_j, \mathbf{R}_{\sigma(j)}) = \sum_{t=1}^{T} \sum_{k=1}^{N_j^t} \mathbb{1}[\mathbf{M}_{\sigma(j)}(\mathbf{P}_j^t(k), t) \neq \mathbf{L}_j^t(k)] \tag{3}$$

where $\mathbb{1}[\cdot]$ is the indicator function, k is the point index, t is the time index, $\mathbf{P}_j^t(k) \in \mathbb{R}^2$ is the x-y coordinates for the k^{th} point of the j^{th} video object at the t^{th} frame.

To combat a server lack of negative points during matching, especially when only positive points are annotated, we further develop a cross instance negative cost $\mathcal{L}_{\text{cineg}}(\mathbf{G}_j, \mathbf{R}_{\sigma(j)})$ to filter out multiple overlapping proposals. The key idea is that the positive points for one video object can serve as accurate negative points for the other video objects in the same video frame. By aggregating the positive point annotations from other video objects in the same frame, we obtain additional accurate negative point annotations for each video object. Therefore, this annotation-free $\mathcal{L}_{\text{cineg}}$ can penalize inaccurate proposals that overlap with the positively annotated points in other video instances.

In addition to the annotated cost and cross instance negative cost, we also define a maskness cost $\mathcal{L}_{\text{maskness}}$ as the negative of the maskness score to favor confident proposals. As a result, our proposed matching cost $\mathcal{L}_{\text{match}}$ is a weighted combination of the annotated cost, cross instance negative cost, and maskness cost as below:

$$\mathcal{L}_{\text{match}} = \lambda_1 \mathcal{L}_{\text{ann}} + \lambda_2 \mathcal{L}_{\text{cineg}} + \lambda_3 \mathcal{L}_{\text{maskness}} \tag{4}$$

where λ_1, λ_2 and λ_3 are the weight balancing parameters.

To handle the birth and death of objects, we compute the matching cost only over video frames where objects show up. After matching, we remove the pseudo-labels for frames where objects die. With our carefully designed matching cost, we can obtain high quality dense pseudo-mask for objects with point annotations via the optimal permutation.

Self-training for generalization. With the above high-quality pseudo masks, we can train our video instance segmentation model on videos with standard loss for mask prediction (cross-entropy and dice loss) and cross-entropy loss for classification following the existing work [13, 26].

Method	Dataset	Sup.	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
TeViT [64]	YouTube-VIS-2019	$\mathcal{M}$	56.8	80.6	63.1	52.0	63.3
IDOL [59]	YouTube-VIS-2019	$\mathcal{M}$	64.3	87.5	71.0	55.6	69.1
MinVIS [26]	YouTube-VIS-2019	$\mathcal{M}$	61.6	83.3	68.6	54.8	66.6
PointVIS (P1)	YouTube-VIS-2019	$\mathcal{P}_1$	53.9 (87.5%)	75.7 (90.9%)	61.8 (90.1%)	47.5 (86.7%)	61.4 (92.2%)
PointVIS (P1N1)	YouTube-VIS-2019	$\mathcal{P}_2$	59.6 (96.7%)	83.3 (100%)	67.1 (97.8%)	52.7 (96.2%)	63.8 (95.8%)
SeqFormer [58]	YouTube-VIS-2021	$\mathcal{M}$	51.8	74.6	58.2	42.8	58.1
IDOL [59]	YouTube-VIS-2021	$\mathcal{M}$	56.1	80.8	63.5	45.0	60.1
MinVIS [26]	YouTube-VIS-2021	$\mathcal{M}$	55.3	76.6	62.0	45.9	60.8
PointVIS (P1)	YouTube-VIS-2021	$\mathcal{P}_1$	46.3 (83.7%)	70.5 (92.0%)	51.1 (82.4%)	37.7 (82.1%)	52.9 (87.0%)
PointVIS (P1N1)	YouTube-VIS-2021	$\mathcal{P}_2$	48.5 (87.7%)	73.0 (95.3%)	54.4 (87.7%)	41.7 (90.8%)	54.1 (89.0%)
MaskTrack [38]	Occluded VIS	$\mathcal{M}$	28.9	56.3	26.8	13.5	34.0
IDOL [59]	Occluded VIS	$\mathcal{M}$	42.6	65.7	45.2	17.9	49.6
MinVIS [26]	Occluded VIS	$\mathcal{M}$	39.4	61.5	41.3	18.1	43.3
PointVIS (P1)	Occluded VIS	$\mathcal{P}_1$	28.6 (72.6%)	49.6 (80.7%)	27.5 (66.6%)	15.0 (82.9%)	32.8 (75.8%)
PointVIS (P1N1)	Occluded VIS	$\mathcal{P}_2$	28.6 (72.6%)	51.2 (83.3%)	27.2 (65.9%)	14.7 (81.2%)	32.2 (74.4%)

Table 1. **Full mask ($\mathcal{M}$) vs. our point supervision ($\mathcal{P}$) on validation set of YouTube-VIS 2019 [61], YouTube-VIS 2021 [61], and OVIS [49].** All results below are based on Swin-L backbone. Our PointVIS results are with self-training.

To generalize from images to videos for new categories, we conduct self-training by regenerating pseudo-labels again from our finetuned video model. The reason is that our pseudo-labels are initially generated from an image model that has never been trained on videos, and there is obviously a domain gap. During self-training, we use confidence score instead of maskness score for pseudo-label matching as the model has been finetuned on videos.

4. Experiments

We evaluate our method on three VIS datasets: YouTube-VIS 2019 [61], YouTube-VIS 2021 [61], and Occluded VIS [49]. We describe our experimental setup (Sec.4.1), compare PointVIS with SOTA fully-supervised methods (Sec.4.2), and provide an ablation study (Sec. 4.3). For more details, please refer to the supplementary material.

4.1. Experimental Setup

Datasets. YouTube-VIS 2019 [61] is a popular dataset for VIS with 2,883 labeled videos, 131K instance masks, and 40 classes. **YouTube-VIS 2021** [61] is an improved version with 8,171 unique video instances and 232k instance masks. **OVIS** [49] is a recently proposed challenging VIS dataset with heavy occlusion and long sequences, containing 296k instance masks and 5.8 instance per video from 25 classes. We synthesize point annotations by randomly sampling points given the ground truth mask in each frame.

Architecture and optimization. We build our PointVIS on top of MinVIS [26] by strictly following its model architecture, training hyper parameters and losses. The only two modifications during video training are: 1) we use pseudo masks obtained from our method given point-level annotations while MinVIS uses annotated masks; 2) we use larger iterations as pseudo-labels require longer time for convergence. All models are pre-trained with COCO instance segmentation [40] then finetuned on videos with Swin backbone [44] unless otherwise stated. For our point-based matcher, we set $\lambda_1 = 5.0$, $\lambda_2 = 5.0$, and $\lambda_3 = 2.0$ for all three datasets. We conduct one iteration of self-training.

Baselines. We propose new baselines for comparison as no prior work is directly applicable to our new point-supervised setting for VIS. 1) **VISP**: naive training MinVIS with points, where only locations with annotated points are supervised during video training. 2) **VISP+CINeg**: adding annotation-free negative point loss to VISP by enforcing cross-instance negative cues on top of it, as described in Sec. 3.2. This strategy is denoted as "CINeg". 3) **VISC**: to ablate the benefits of points and decouple the impact of pre-trained image model, we obtain pseudo-labels by selecting the top k most confident proposals from our proposal set $\hat{\mathbf{R}}$ sorted by the confidence score $\hat{c}_r$.

We also include the existing unsupervised/semi-/weakly-supervised VIS work to compare with our point-supervised setting as follows. 1) **VIS-Unsup** [19]: pretrained COCO image model built on top of SOLO [54] without finetuning on videos. 2) **VIS-Semi** [19]: finetuning on videos on top of VIS-Unsup but without video annotations. Note that VIS-Unsup and VIS-Semi can only handle the overlapped categories, therefore we report AP/AR (seen) in line with [19]. 3) **VIS-Weak** [42]: the first weakly-supervised VIS model that uses category annotations only in each video frame.

Metrics. We use AP and AR for evaluation and train on the training split and evaluate on the validation split using public evaluation servers of the three datasets. Baselines VISC, VIS-Unsup [19] , and VIS-Semi [19] can only handle categories overlapped with the image model, so we report AP/AR (seen) computed by averaging only over the

Model	$\mathcal{L}_{\text{Ann}}$	$\mathcal{L}_{\text{CINeg}}$	$\mathcal{L}_{\text{maskness}}$	Self-Train	mAP
MinVIS [26] (Upperbound)	-	-	-	-	55.3
VISP (P1)	-	-	-	-	0.4
VISP+CINeg (P1)	-	-	-	-	9.7
PointVIS (P1)	✓	-	-	-	40.4
PointVIS (P1)	✓	✓	-	-	45.7
PointVIS (P1)	✓	✓	✓	-	46.0
PointVIS (P1)	✓	✓	✓	✓	47.3

Table 2. **Effects of each component on YouTube-VIS 2019 [61] val-dev.**

Model ID	Sampling method for Pos	Sampling method for Neg	mAP
PointVIS (P1)	Random	-	46.0
PointVIS (P1)	Distance Transform	-	47.1
PointVIS (P1N1)	Random	Random (In-box)	48.6
PointVIS (P1N1)	Random	Random (Out-box-but-in-200%-box)	48.0

Table 3. **Analysis of point selection bias on YouTube-VIS 2019 [61] val-dev.**

Model ID	CINeg	DPPointMatcher	P1	P10	P1N1	P5N5
VISP	-	-	0.4	0.5	27.6	33.0
VISP+CINeg	✓	-	9.7	10.0	45.6	48.9
PointVIS (Ours)	✓	✓	46.0	45.9	48.6	49.4

Table 4. **Effects of additional points on YouTube-VIS 2019 [61] val-dev.** "DPPointMatcher" means Dense Pseudo via Point Matcher. "CINeg" means enforcing additional negative point loss.

overlapped categories for these baselines following [19].

4.2. Comparison with Fully-Supervised SOTA

We compare our PointVIS with recent fully-supervised methods including IDOL [59], MinVIS [26], TeViT [64], SeqFormer [58], and MaskTrack [38], as shown in Table 1. Note that MinVIS [26] serves as the fully-supervised counterpart of PointVIS, therefore we compute the retention rate as the performance of PointVIS divided by the performance of MinVIS. We use $\mathcal{M}$ to indicate full supervision from dense masks, and $\mathcal{P}_n$ to indicate sparse supervision from n points. P1 means only one positive point is labeled. P1N1 means one positive and one negative opint are labeled, and similarly for larger number of points.

YouTube-VIS 2019. With one single point per object (P1), PointVIS achieves 53.9 mAP. With two points per object (P1N1), PointVIS achieves 59.6 mAP, which is 96.7% of the supervised counterpart and even surpassed recent fully-supervised method TeViT [64] by nearly 3 AP. These competitive results demonstrate the potential of our point-supervised video instance segmentation framework.

YouTube-VIS 2021. On this more challenging dataset, we can still achieve 87.7% performance of the supervised counterpart, reaching 48.5 AP, which is only 3 AP away from recent fully-supervised approach SeqFormer.

Model	Backbone	Sup.	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}
VIS-Unsup [19]	R50	$\mathcal{V}$	23.9	43.3	21.5	26.7	37.3
VIS-Semi [19]	R50	$\mathcal{V}$	38.3	61.1	39.8	36.9	44.5
VISC	R50	$\mathcal{V}$	42.0	62.5	47.3	48.7	56.7
PointVIS (P1-Ours)	R50	$\mathcal{P}_1$	47.0	67.4	53.4	44.4	50.9
PointVIS (P1-Ours)	Swin-L	$\mathcal{P}_1$	58.6	80.3	66.2	54.1	64.2

(a) Evaluation on seen categories

Model	Backbone	Sup.	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}
VIS-Weak [42]	R50	$\mathcal{C}$	10.5	27.2	6.2	12.3	13.6
PointVIS (P1)	R50	$\mathcal{P}_1$	38.5	58.9	41.2	36.4	46.4
PointVIS (P1-Ours)	Swin-L	$\mathcal{P}_1$	52.5	74.5	59.2	47.2	61.5

(b) Evaluation on all categories

Table 5. **Comparison with baselines on YouTube-VIS 2019 [61] validation set.**

OVIS. PointVIS achieves 28.6 AP on this challenging dataset, 72.6% of the supervised counterpart, and matches previous fully-supervised methods [38]. The retention rate is relatively lower compared to the other two datasets, likely due to the quality of spatio-temporal proposals, which degrades due to extremely long sequences in OVIS.

4.3. Ablation Study

In this section, we conduct ablation studies to 1) verify the effectiveness of each individual component of our proposed framework; 2) analyze training with subsampled frames; 3) analyze the effect of number of points; 4) analyze point selection bias and 5) compare with baselines on YouTube-VIS 2019 [61]. We split the training set of YouTube-VIS 2019 [61] into train-dev for training and val-dev for testing with Swin-B backbone unless otherwise stated. We do not use self-training by default for simplicity unless otherwise stated. Each ablation is conducted under the same experimental setting for a fair comparison.

4.3.1 Effects of model components.

Table 2 presents the ablation results for each component of our model. The VISP baseline, which trains the VIS model with only point supervision, yields a near-zero AP. Incorporating cross instance negatives (VISP+CINeg) improves it to 9.7 mAP. Our method achieves a decent mAP of 40.4 by incorporating pseudo masks generated through our point-based matcher with only $\mathcal{L}_{\text{Ann}}$ in the matching process. Further addition of $\mathcal{L}_{\text{CINeg}}$ and $\mathcal{L}_{\text{maskness}}$ results in a significant boost of 5.7 mAP, highlighting the importance of incorporating cross instance negatives. Finally, with self-training reducing the domain gap, our PointVIS achieves 47.3 mAP, which is 85.6% of the fully-supervised upperbound.

4.3.2 Analysis of subsampling frames.

PointVIS can be extended to the setting of using a subset of frames following MinVIS [26] (Table 6). With point labels

in only 1% frames, PointVIS (55.0 AP) is only 6.6 AP behind the one uses full-masks in 100% frames, and achieves a retention rate of 93.2% of its counterpart with full-masks in 1% frames. This indicates the practical utility of our work.

PointVIS could be further easily extended to a per-frame setting where frames are annotated with points in parallel w/o tracking annotations. The main change here is that we cannot compute spatio-temporal matching cost. Instead, we change matching to per-frame by computing the matching cost for each annotated frame independently. We refer the resulting per-frame model as PointVIS (PF). PointVIS (PF) achieves competitive results while PointVIS has its own advantages, especially with less frames (Table 6).

4.3.3 Analysis of point selection bias.

To analyze how point selection bias affects performance, we synthesized point annotations using different sampling methods (Table 3). Under the P1 setting, we compared random sampling with human distance transform and found that different methods achieved similar results, but distance transform performed around 1AP better. This suggests that our method is generally robust to annotated point locations, but human annotation can potentially yield higher gains. Under the P1N1 setting, we compared sampling negatives inside versus outside the bounding box and found that the latter was only 0.6 AP behind, indicating that negatives do not need to be constrained within boxes.

4.3.4 Analysis with additional points.

Table 4 summarizes our results on adding more points. Increasing the number of positive points alone (P1 vs P10) did not improve performance because the model lacks additional cues about the background region. Adding both negative and positive points improved the performance of VISP and VISP+CINeg, which rely heavily on point annotations for knowing foreground and background. Our proposed method's performance improved significantly with just one more negative point and remained stable across all point settings, indicating its robustness.

4.3.5 Comparison with additional baselines.

Table 5 shows the comparison with additional baselines. Compared with VIS-Unsup [19] without using video annotations, our PointVIS outperforms VIS-Unsup by more than 10 AP with little annotation overhead. Compared with the VISC baseline that does not use video annotations but generating pseudo-labels by confidence, PointVIS outperforms it by 5 AP showing the benefit of point annotations.

5. Conclusion

In this work, we address the challenging point-supervised video instance segmentation problem and introduce a point-supervised VIS framework. The key ingredients are utilizing object shape prior from a pretrianed COCO image instance segmentation model, and our proposed spatio-temporal point-based matcher for generating high-quality dense pseudo-labels for videos. Our method allows us to reduce annotations to only one point for each object in a video frame, yet retaining high quality mask predictions close to full supervision. We conduct comprehensive experiments on three datasets and achieve competitive performance compared with the fully-supervised methods.

Model	Sup.	Matching	1%	5%	10%	100%
MinVIS [19]	$\mathcal{M}$	-	59.0	59.3	61.0	61.6
PointVIS (PF)	$\mathcal{P}_2$	Per-frame	52.3 (88.6%)	54.5 (91.9%)	54.7 (89.7%)	57.3 (93.0%)
PointVIS	$\mathcal{P}_2$	Spatio-temporal	55.0 (93.2%)	55.5 (93.6%)	56.1 (92.0%)	57.4 (93.2%)

Table 6. **PointVIS (P1N1) with subsampled video frames on YouTube-VIS 2019 validation set (w/o self-training).**

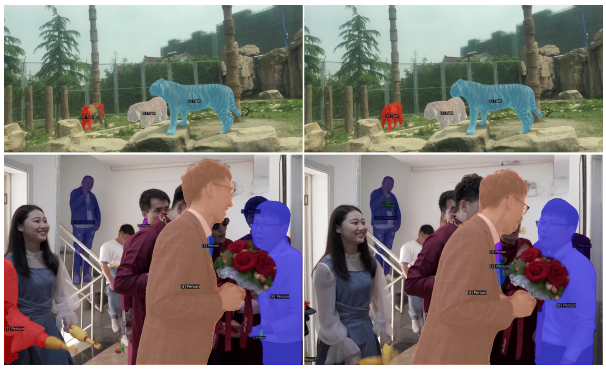

Figure 4. **Failure cases on OVIS [49].** We observe temporal inconsistency (*e.g.* tiger in the top left) or missing instances (*e.g.* person in white) in our pseudo masks.

Limitations. While PointVIS shows promising results in the direction of weakly-supervised video instance segmentation, it has some limitations. Typical failure cases are shown in Figure 4. We observe missing instances and temporal inconsistency. One hypothesis is that our performance is bounded by the quality of proposals, and we believe stronger video instance architecture can mitigate this gap. Another potential direction is to utilize video correspondences for label propagation and other denoising techniques for higher-quality proposals. We leave these as future work.

Broader Impacts. We open up the possibility of learning video instance segmentation with point-level supervision. We hope that our framework can be used to largely reduce annotation cost for a wide range of video recognition tasks in computer vision, including video object detection, video object segmentation, and video panoptic segmentation.

Acknowledgements. This work was partially funded by NSF CAREER Award (2238769) to AS.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 3

[2] Aditya Arun, CV Jawahar, and M Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. In *European Conference on Computer Vision*, pages 254–270. Springer, 2020. 3

[3] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 3

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[5] Arnaud Benard and Michael Gygli. Interactive video object segmentation in the wild. *arXiv preprint arXiv:1801.00269*, 2017. 3

[6] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 3

[7] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019. 3

[8] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 3

[9] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 1–18. Springer, 2020. 1

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5

[11] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8832, 2021. 3

[12] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 3

[13] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 1, 3, 5

[14] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-supervised instance segmentation. *arXiv*, 2021. 2, 3

[15] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 4

[16] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021. 3

[17] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Pointly-supervised panoptic segmentation. In *European Conference on Computer Vision*, pages 319–336. Springer, 2022. 3

[18] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. Compfeat: Comprehensive feature aggregation for video instance segmentation. *arXiv preprint arXiv:2012.03400*, 2020. 1, 3

[19] Yang Fu, Sifei Liu, Umar Iqbal, Shalini De Mello, Humphrey Shi, and Jan Kautz. Learning to track instances without video annotations. In *CVPR*, 2021. 1, 3, 6, 7, 8

[20] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *European Conference on Computer Vision*, pages 88–104. Springer, 2020. 1

[21] Bo He, Xitong Yang, Hanyu Wang, Zuxuan Wu, Hao Chen, Shuaiyi Huang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. Towards scalable neural representation for diverse videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6132–6142, 2023. 1

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICVV*, 2017. 3

[23] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7322–7330, 2021. 3

[24] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32:6586–6597, 2019. 3

[25] Hexiang Hu, Shiyi Lan, Yuning Jiang, Zhimin Cao, and Fei Sha. Fastmask: Segment multi-scale object candidates in one shot. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–999, 2017. 3

[26] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *arXiv preprint arXiv:2208.02245*, 2022. 1, 3, 4, 5, 6, 7

[27] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2010–2019, 2019. 3

[28] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *European Conference on Computer Vision*, pages 267–284. Springer, 2022. 3

[29] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *NeurIPS*, 2021. 1, 3

[30] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 1

[31] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Regognition (CVPR)*, 2019. 1

[32] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *ICCV*, 2021. 3

[33] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Proposal-based instance segmentation with point supervision. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2126–2130. IEEE, 2020. 3

[34] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020. 3

[35] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11215–11224, 2021. 1, 3

[36] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2359–2367, 2017. 3

[37] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[38] Zhuang Li, Leilei Cao, and Hongbin Wang. Limited sampling reference frame for masktrack r-cnn. In *ICCVW*, 2021. 6, 7

[39] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. *arXiv preprint arXiv:2103.13746*, 2021. 1, 3

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 6

[41] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9816–9825, 2021. 3

[42] Qing Liu, Vignesh Ramanathan, Dhruv Mahajan, Alan Yuille, and Zhenheng Yang. Weakly supervised instance segmentation for videos with temporal mask consistency. In *CVPR*, 2021. 1, 3, 6, 7

[43] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. Pseudoclick: Interactive image segmentation with click imitation. *arXiv preprint arXiv:2207.05282*, 2022. 3

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6

[45] Kira Maag, Matthias Rottmann, Serin Varghese, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Improving video instance segmentation by light-weight temporal uncertainty estimates. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 1, 3

[46] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *arXiv preprint arXiv:1506.06204*, 2015. 3

[47] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European conference on computer vision*, pages 75–91. Springer, 2016. 3

[48] Brian L Price, Bryan S Morse, and Scott Cohen. Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. In *2009 IEEE 12th International Conference on Computer Vision*, pages 779–786. IEEE, 2009. 3

[49] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *arXiv preprint arXiv:2102.01558*, 2021. 5, 6, 8

[50] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020. 3

[51] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *CVPR*, 2021. 3

[52] Shubham Tulsiani and Jitendra Malik. Viewpoints and keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1519, 2015. 1

[53] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. *arXiv preprint arXiv:2203.16089*, 2022. 3

[54] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020. 6

[55] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *CVPR*, 2022. 4

[56] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 3

[57] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An

online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2021. 3

[58] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021. 6, 7

[59] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 6, 7

[60] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12193–12202, 2020. 3

[61] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 3, 5, 6, 7

[62] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *ICCV*, 2021. 3

[63] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Ying Shan, Bin Feng, and Wenyu Liu. Tracking instances as queries. *arXiv preprint arXiv:2106.11963*, 2021. 3

[64] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *CVPR*, 2022. 6, 7

[65] Xuehui Yu, Pengfei Chen, Di Wu, Najmul Hassan, Guorong Li, Junchi Yan, Humphrey Shi, Qixiang Ye, and Zhenjun Han. Object localization under single coarse point supervision. *arXiv preprint arXiv:2203.09338*, 2022. 3

[66] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10226–10235, 2020. 3

[67] Tianfei Zhou, Jianwu Li, Xueyi Li, and Ling Shao. Target-aware object discovery and association for unsupervised video multi-object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2021. 1, 3

[68] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018. 3

[69] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5368, 2019. 1